BrainStation Capstone

Drug Review Analysis

Ali Mohammed

Problem Statement

Pharmaceutical companies play a crucial role in providing medications that improve the health and well-being of individuals. Understanding how certain drugs are perceived and the sentiment associated with them can provide valuable insights for these companies, enabling them to make informed decisions and drive business success and patient well-being. By conducting sentiment analysis on drug reviews, pharmaceutical companies can gain a deeper understanding of how their products are perceived by users. Sentiment analysis allows them to extract valuable insights from the textual reviews, uncovering the overall sentiment expressed by users towards specific drugs. This information can help pharmaceutical companies identify areas for improvement, optimize their marketing strategies, and enhance customer satisfaction. For example, if a drug receives overwhelmingly positive sentiment, it indicates high user satisfaction, which can be leveraged in marketing campaigns to promote the drug's benefits and build trust among potential customers. Conversely, if negative sentiment is detected, it can serve as an early warning sign of potential issues or side effects, prompting pharmaceutical companies to investigate and address them promptly. Additionally, being able to predict the conditions mentioned in drug reviews can provide pharmaceutical companies with valuable insights into the efficacy and suitability of specific medications for different medical conditions and other pharmaceutical applications. By analysing the reviews and accurately predicting the associated conditions, pharmaceutical companies can gain a better understanding of which conditions their drugs are commonly prescribed for. This knowledge can inform their decision-making processes, such as optimizing product development strategies, tailoring marketing efforts to specific conditions, and identifying potential areas for future drug development. For example, if a particular drug consistently appears in reviews related to a specific condition, it suggests that the drug is effective in treating that condition, allowing pharmaceutical companies to position it as a suitable treatment option and allocate resources accordingly.

Background

The application of data science techniques to drug reviews and pharmaceutical data is highly valuable for several reasons. Firstly, the volume of available data in the healthcare industry is massive and continues to grow exponentially. By leveraging data science, we can effectively analyse and extract meaningful insights from this vast amount of data, providing valuable information for decision-making processes. Secondly, the subjective nature of drug reviews and patient experiences requires sophisticated analysis to uncover patterns and sentiments hidden within the data. Data science techniques, such as sentiment analysis, allow us to quantify and understand the sentiments expressed in patient reviews, providing valuable insights into drug effectiveness, side effects, and overall patient experiences. Furthermore, pharmaceutical companies can utilize data science techniques to predict and understand patient conditions based on drug reviews. This enables targeted treatment approaches, personalized medicine, and improved patient outcomes.

How has it been addressed in the past?
In the past, addressing this problem typically involved manual analysis of drug reviews, which was time-consuming, labour-intensive, and subject to human bias. Pharmaceutical companies relied on manual reading and categorization of reviews to gain insights, which limited the scalability and accuracy of their analyses.

However, with the advancements in data science and machine learning, sentiment analysis has become a classification problem. Sentiment analysis involves classifying reviews into positive, negative, or neutral sentiments. Machine learning algorithms are trained on labelled data, where the sentiment of reviews is known, to develop models capable of classifying new, unlabelled reviews.

Dataset

To address these objectives, we have chosen the **UCI ML drug review dataset** sourced from **Kaggle.com**. The dataset includes textual reviews and numeric ratings of drugs collected from Drugs.com. It is important to note that text reviews may not always align with the numeric ratings, as extreme positive or negative ratings can sometimes be misleading. Hence, our project focuses on classifying user ratings of drugs based on their textual reviews using traditional machine learning algorithms. The data came in two .csv files, a train.csv and a test.csv.

Here is the **Data Dictionary** for the dataset:

uniqueID	Unique ID assigned to each record
drugName	Unique ID assigned to each record
condition	Name of the condition
review	Patient review
rating	Patient rating on a scale of 1 to 10
date	Date when the review was entered
usefulCount	Number of users who found the review useful

- The training data contains 161297 rows and 7 columns.
- The test data contains 53766 rows and 7 columns.
- The dataset begins on 1-Apr-08 and ends on 9-Sep-17.

The **Data Types** are:

drugName	object
condition	object
review	object
rating	int64
date	object
usefulCount	int64

Summary of Preprocessing

In the cleaning and preprocessing phase of our project, we focused on filtering and transforming the data to ensure its quality and suitability for analysis. Here is an overview of the main steps we performed after combining the DataFrames:

- **Filtering and Data Reduction:**

We started by filtering the dataset to reduce the number of conditions from 916 to 72. By selecting the top conditions and considering those with both high and low useful counts, we aimed to cover a significant portion (83%) of the total reviews while avoiding any skewness in the data representation.

- **Text Preprocessing:**

We addressed errors in the collected text data, specifically in columns such as conditions, review, and drugName. Text preprocessing techniques were applied to handle issues related to web scraping and normalizing the text. These techniques helped ensure consistency and accuracy in the textual data.

- **Handling Unique IDs:**

The uniqueID column was transformed to serve as the index of the dataset since it was anonymized and only served as an identifier. This adjustment simplified the data structure without compromising its integrity.

- **Date Formatting:**

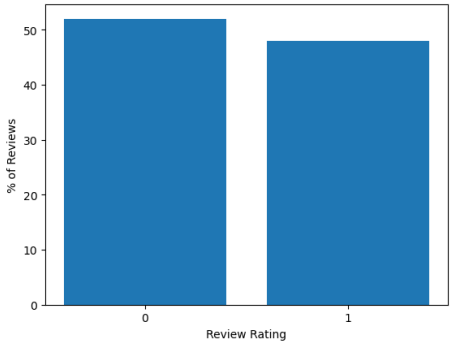
The date column was converted to a datetime format to facilitate temporal analysis and enable time-based insights.

- **Encoding Categorical Variables:**

We applied label encoding to the drugName and condition columns, converting their values into numerical representations. This encoding allowed us to incorporate these categorical variables into our analysis effectively.

- **Rating Transformation:**

To align with our sentiment analysis objective, we transformed the rating column from a scale of 1 to 10 into a binary problem. We defined positive and negative sentiments by equally representing 50% of the ratings. This transformation facilitated the sentiment analysis process and allowed us to focus on classifying reviews as either positive or negative. The decision to start with a binary problem was driven by the nature of the data and the need for efficient sentiment classification.



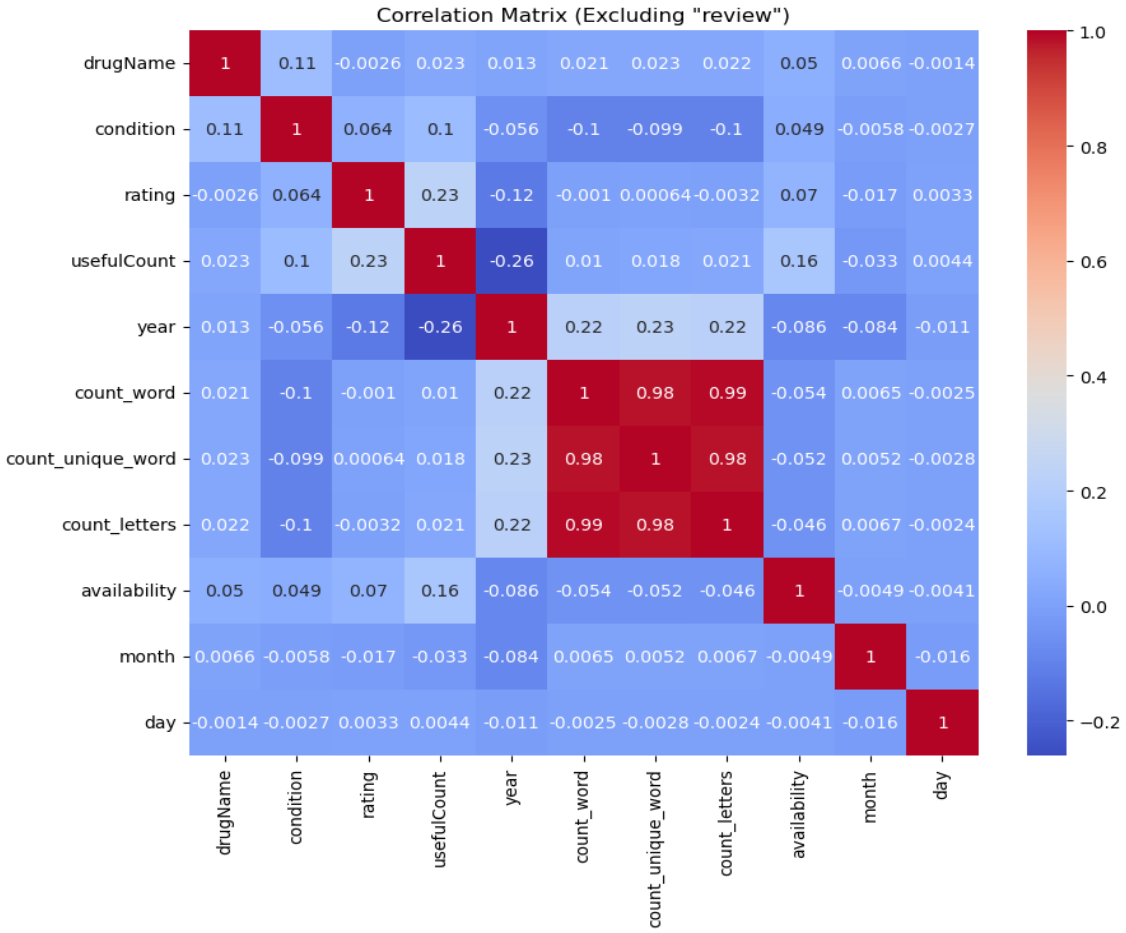
- **Review Column:**

The review column, which contains the patient reviews, underwent its own preprocessing steps, such as removing special characters, punctuation, and stop words. These steps aimed to clean the text data and reduce noise, enabling better analysis and interpretation of the reviews.

- **Feature Engineering:**

Created features relevant to review text. Features like count of words, count of unique words, count of letters. Also web scraped Drugs.com for condition availability; if its prescription or a OTC drug.

Correlation matrix of all our features:

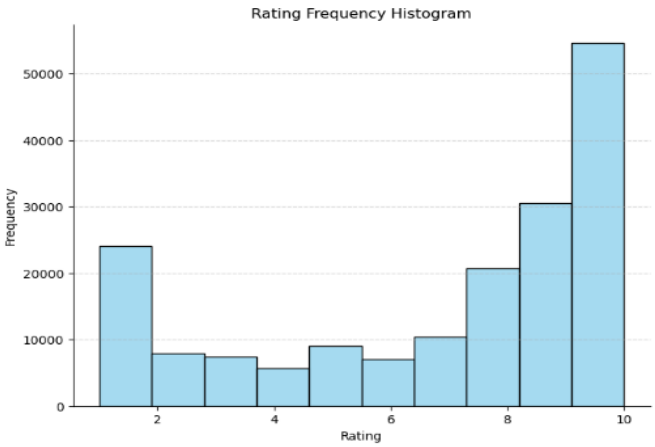


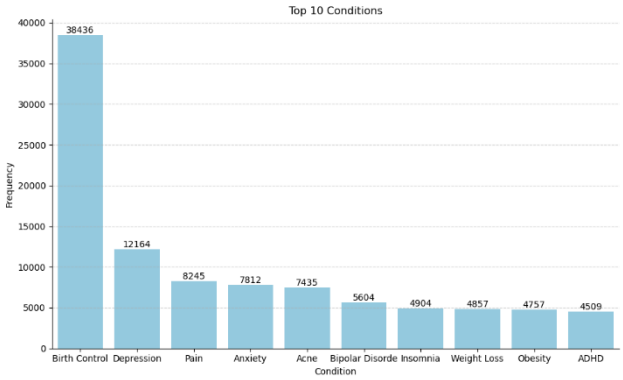
By implementing these cleaning and preprocessing techniques, we ensured the quality, consistency, and suitability of the data for further analysis. These steps prepared the data for exploratory data analysis (EDA) and modelling, allowing us to derive meaningful insights from the drug reviews and facilitate accurate sentiment analysis.

Insights, Modelling, and Results

EDA Insights

Sentiment analysis directly rates to ratings and review text analysis. We focused on showing visualizations that allow to support or generate insights. One of the key visualizations is the rating frequency histogram, which illustrates the distribution of ratings given by users. This histogram allows us to understand the overall sentiment expressed by users and identify any trends or patterns in the ratings. We dive in deeper to show more insights of rating in the notebook.





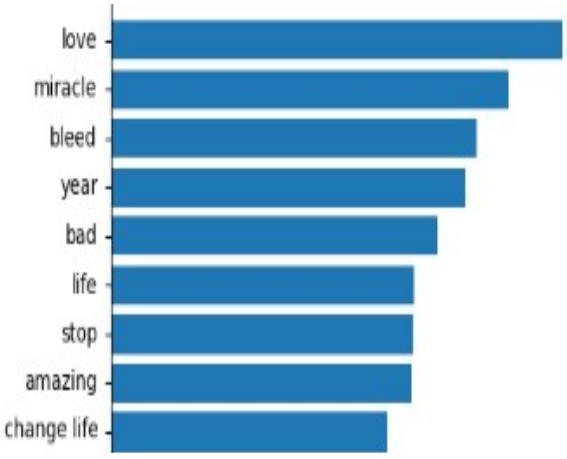
Also by understanding the most commonly discussed conditions, pharmaceutical companies can focus their efforts on developing targeted treatments and improving patient outcomes in these specific areas.

Moreover, we created word clouds that show the words that are high in frequency in the reviews. This serves as a starting point for finding sentiment of words that carry importance in the review text.

Many More insights are generated that are not shown but available in the notebooks of the project.

Modelling

Sentiment analysis plays a crucial role in understanding the emotions and opinions expressed in text data. In our project, we focus on applying sentiment analysis to user reviews of drugs, aiming to gain valuable insights into how users perceive these medications. Our modelling phase involves experimenting with various algorithms, including linear regression, decision trees, random forest, and XGBoost. These models offer different advantages and insights into the sentiment analysis process. Throughout the modelling process, we rigorously evaluate the performance of each model. Through each **Iteration** we add **different preprocessing techniques** such as Count vectorizer, Ngrams,and Mindf to the review text to assess how the models are



Word importance

performing. We also assess metrics such as accuracy and precision, examine feature importance's to identify the key words contributing to sentiment prediction, and thoroughly analyse the strengths and limitations of each approach.By adopting an iterative approach and exploring multiple models, our aim is to develop a robust sentiment analysis solution that accurately captures the sentiment expressed in user reviews of drugs and the words related to the prediction task. In addition, a classification model that's able to classify drug review text from 71 different conditions.

Findings and Conclusions

Sentiment Analysis Modelling Results

	Logistic Regression	Decision Trees	Random Forest	XGBOOST
Train Accuracy	**99%**	**88%**	**87%**	**99%**
Test Accuracy	**88%**	**77&**	**79%**	**97%**

Classification Modelling Results

	Logistic Regression	Random Forest	SciBERT
Train Accuracy	**99%**	**87%**	**XXX**
Test Accuracy	**86%**	**79%**	**XXX**

Note that while the high train accuracy of our models, particularly logistic regression and XGBoost, indicates their strong predictive capabilities, it is essential to consider the possibility of overfitting. Overfitting occurs when a model learns the training data too well and struggles to generalize to unseen data. To address this concern, further research and time are required to fine-tune the models and explore different word embeddings techniques. Word embeddings play a crucial role in capturing the semantic meaning of words and can significantly impact the performance of sentiment analysis models. By experimenting with alternative word embeddings approaches, we can enhance the models' ability to capture nuanced sentiment and improve their generalization on unseen data.

We were able to generate insights based on the drug reviews, moreover we were able to predict the sentiment and the condition of the review. Using this information allowed for insights on the feature or text importance in the reviews which is great for pharmaceutical marketing and human wellbeing.
It is important to acknowledge that our current findings and conclusions are preliminary, and more extensive research and experimentation are necessary. Overcoming the limitations of overfitting and exploring alternative word embeddings techniques will be crucial for achieving robust and reliable sentiment analysis models in the future. Additionally, incorporating more advanced natural language processing techniques and considering deep learning approaches could further enhance the models' performance.