Problem statement
Can we forecast the sentiment based on the patient reviews on drug products and their medical conditions?

Data Wrangling
The raw drug review datasets, named "train" and "test" in .tsv format, were merged into one dataframe, resulting in a comprehensive dataset with dimensions 215063 rows and 7 columns. 1194 records lacked vital information in the "condition" category, constituting a missing ratio of 0.56%. To enhance the dataset's integrity, these records were removed.
To further refine the dataset, adjustments were made to the data types of the "rating" and "date" columns, ensuring they adhered to the appropriate formats. Additionally, a stringent filtering process was applied, classifying more than 1000 conditions with comments like "users found it useful" as noise, leading to their exclusion from the dataset.
To address issues in the "condition" column, a meticulous cleansing procedure was executed. Typos and inaccurate spellings were rectified. To illustrate, three conditions - "mist (", "me", "mis" - were consolidated into a single condition, "Mist," to ensure accurate counting.
Within the "review" subset, a substantial number of duplications exceeding 85,000 were identified. These duplications exhibited variations solely in drug names for the same condition. Consequently, only the original records were retained to ensure data integrity.
Furthermore, the "review" column underwent an exhaustive cleaning process, employing various functions such as removing special characters, eliminating whitespace, eradicating stopwords, expanding contractions, and implementing stemming and lemmatization, refer to Figure 1 and 2 for a visual representation of these detailed steps.

	review
0	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"
1	"My son is halfway through his fourth week of Intuniv. We became concerned when he began this last week, when he started taking the highest dose he will be on. For two days, he could hardly get out of bed, was very cranky, and slept for nearly 8 hours on a drive home from school vacation (very unusual for him.) I called his doctor on Monday morning and she said to stick it out a few days. See how he did at school, and with getting up in the morning. The last two days have been problem free. He is MUCH more agreeable than ever. He is less emotional (a good thing), less cranky. He is remembering all the things he should. Overall his behavior is better. \r\nWe have tried many different medications and so far this is the most effective."
2	"I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it's the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn't have any other side effects. The idea of being period free was so tempting... Alas."

Figure 1 : ‘review’ column before cleaning.

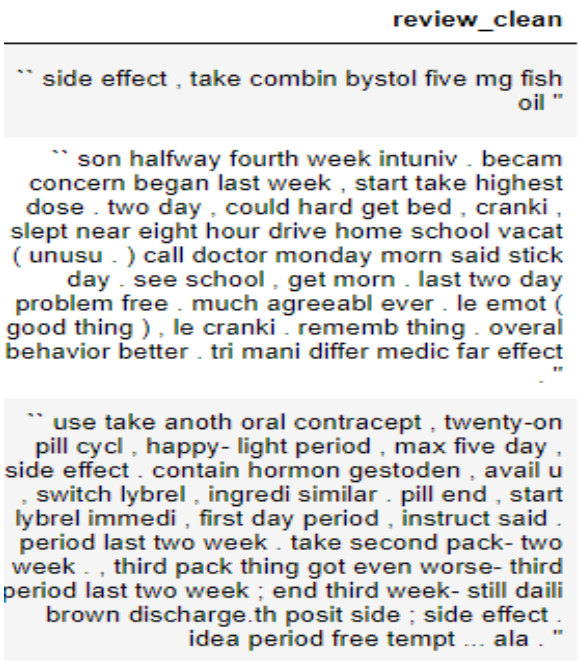

Figure 2 : 'review' column after cleaning.

Visualization during data wrangling revealed top-rated conditions like birth control, depression, and pain. The rating scale ranged from 1 to 10, with drugs like Levonorgestrel appearing in both the top 10 with a rating of 10 and the top 10 with a rating of 1. Analysis of the rating distribution, depicted in Figure 4, uncovered a skewed pattern, with about 70% positive ratings and only 17% and 13% neutral and negative ratings, respectively. This imbalance emphasizes the need for careful interpretation of overall sentiment due to data distribution skewness.

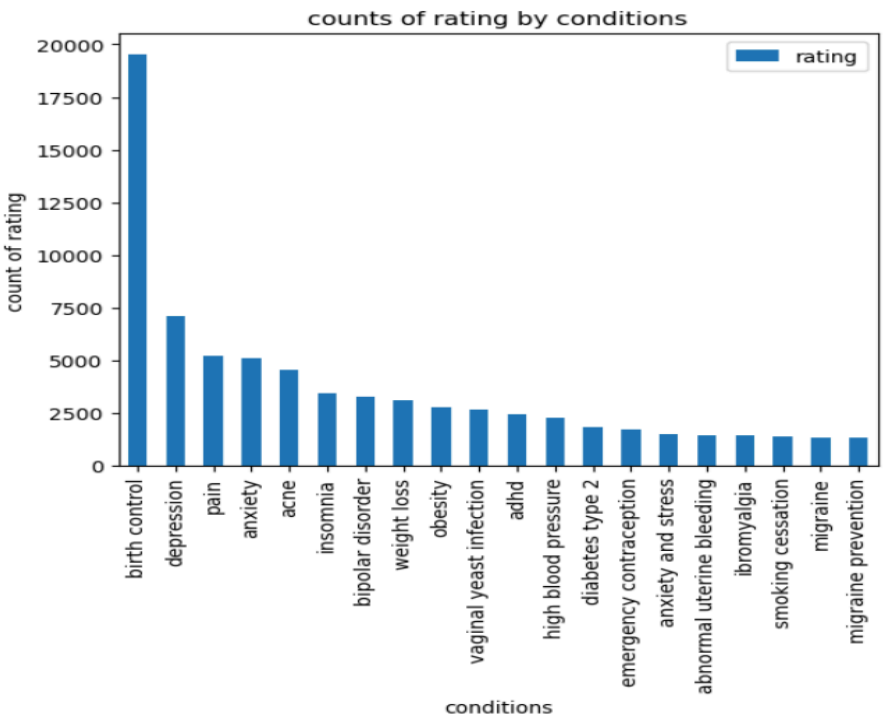

Figure 3 : Top 20 Conditions of Rating

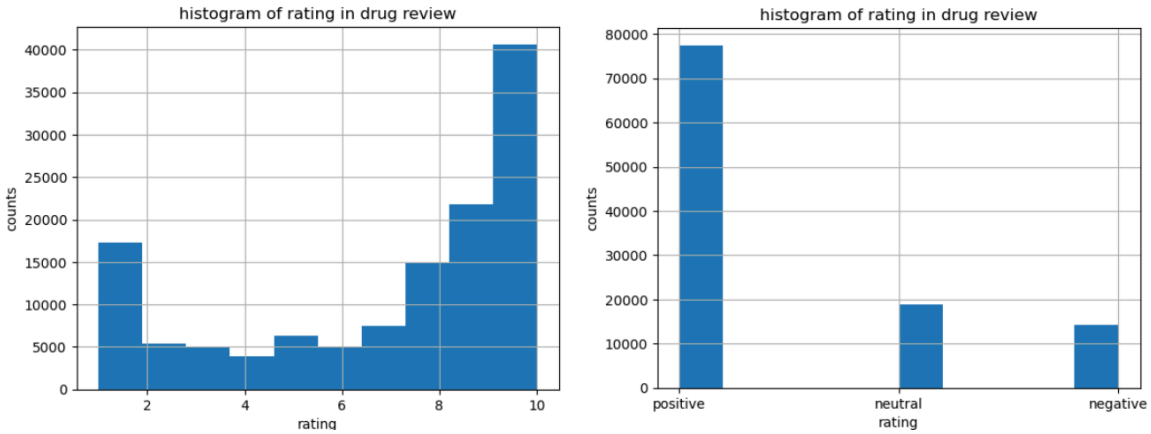

Figure 4 : Histogram of Rating

Exploratory data Analysis

New columns 'word_count,' 'mean_word_len,' 'unique_word_count,' 'Review_len,' 'mean_sentence_len' were introduced. These metrics give insights into the linguistic characteristics and structural attributes of the reviews. These columns revealed the presence of outliers. Specifically, the 'Review_len' exhibited a median around 200, with values below 1000 (1.5 IQR), indicating potential outliers. Similarly, 'mean_sentence_len,' 'word_count,' and 'unique_word_count' displayed medians below 200, 100, and 4, respectively, suggesting the presence of outliers in these dimensions.

Concurrently, the sentiment analysis was conducted using 'sentiment_subjectivity,' 'sentiment_score,' and 'sentiment_label,' implemented through TextBlob. Additionally, a word cloud enhances the understanding of common word elements across ratings of 10 and 1, refer to Figure 5.

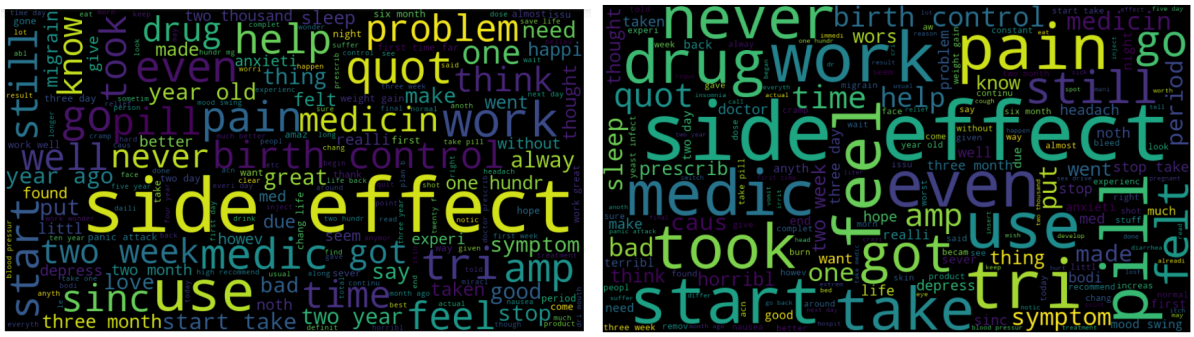

Figure 5 wordCloud of Rate = 10 (left), and Rate =1(right)

Preprocessing

Serve key steps were performed:

- The 'review' text column underwent tokenization, breaking down sentences into individual words for analysis. To prevent overfitting and reduce training time, only the first 200 words in each review column were utilized.
- Categorical features such as 'drugName' and 'condition' were encoded using OrdinalEncoder, converting them into a suitable format for modeling.
- From the 'date' column, additional features like 'year' and 'month' were extracted, providing temporal context for the analysis.
- Target feature 'sentiment' was encoded, creating three sentiment classes to facilitate classification.
- Numerical features underwent scaling using MinMaxScaler to standardize the range of values, ensuring uniformity in the model training process.
- The dataset was split into training (75%) and testing (25%) sets, facilitating robust model evaluation.

Model and Evaluation Metrics

Three models with their unique applications were considered for sentiment analysis

- Multinomial Naive Bayes (MNB) is a probabilistic classification algorithm rooted in Bayes theorem, particularly well-suited for text data processing. Its assumption of feature independence and its simplicity and efficiency make it an excellent baseline model for tasks like text-based sentiment analysis.
- Long Short-Term Memory (LSTM): In sentiment analysis, LSTM excels in capturing dependencies and nuances within text over extended sequences.
- Random Forest (RF), a powerful ensemble learning method,is effective in classification tasks.

Evaluation metrics: the purpose is to accurately predict sentiment while addressing imbalances in the dataset. If accuracy as main metrics may yield misleading insights.

Therefore, the F1 score will be the primary metrics. Other metrics such as training time, precision, recall, and accuracy will be tracked for each model. The ideal model is to achieve a harmonious balance between predictive accuracy, computational efficiency, and the capability to discern sentiment in drug reviews. The results reveals RF shows (Table 1) faster training and highest score. Therefore RF was chose as the winning model.

Table 1: Three Model Performance

Model	F1_score	Precision	Accuracy	Recall	Training Time
MNB	0.35	0.40	0.37	0.50	6 second
LSTM	0.93	0.92	0.93	0.95	~1 hour
RF	1.00	1.00	1.00	1.00	~ 2 min

Conclusion

The sentiment analysis on drug products reveals a random forest classifier is the optimal model.

Future scope of work

The absence of supplementary data limited the final application of the model to real-world data. In the future, if time permits, selecting additional data for final modeling, utilizing the remaining dataset for training and modeling purposes will be explored . Additionally, since the current three models lack feature importances, a potential approach involves reconfiguring the training and modeling process to focus on extracting feature importances. This can provide valuable insights into the underlying data.