# Sentiment Analysis on Drug Product Reviews

Data Science Capstone Project

Present by Yan Zhang

24Feb2024

# Problem Statement

Objective: Develop a sentiment analsyis model to gauge customer perception in drug reviews.

Challenge:  distilling meaningful insights from the massive unstructured text data.

Reason:

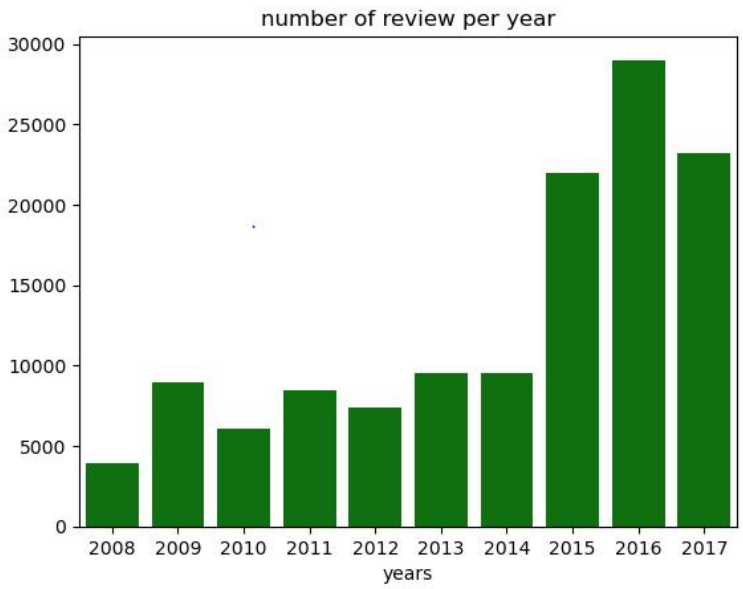
1. Improve and potential modifications in drug products.
2. Provide insights in product development and marketing strategies.

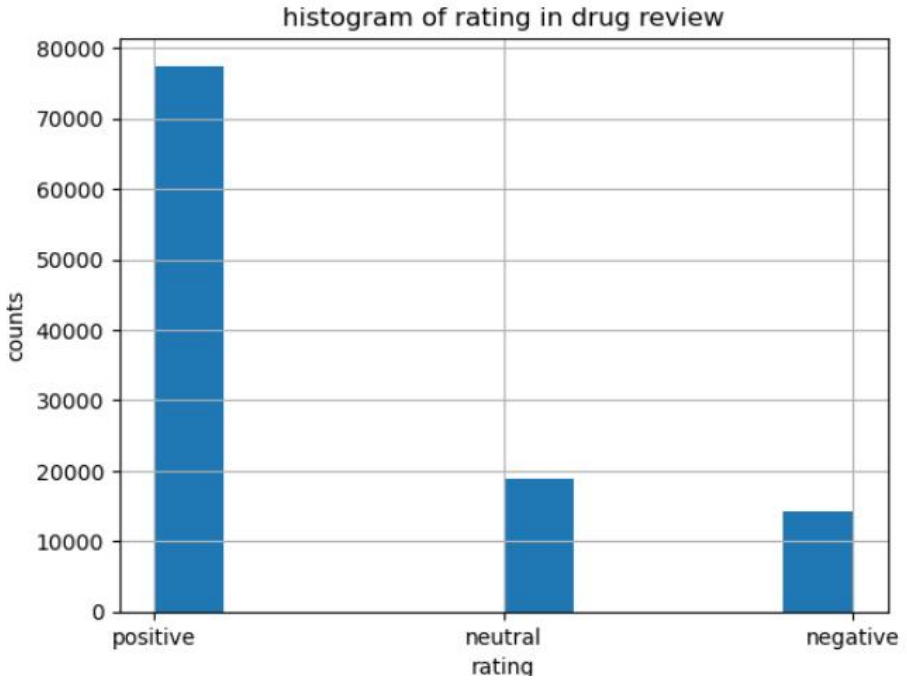
# Data Wrangling

Quality of data:

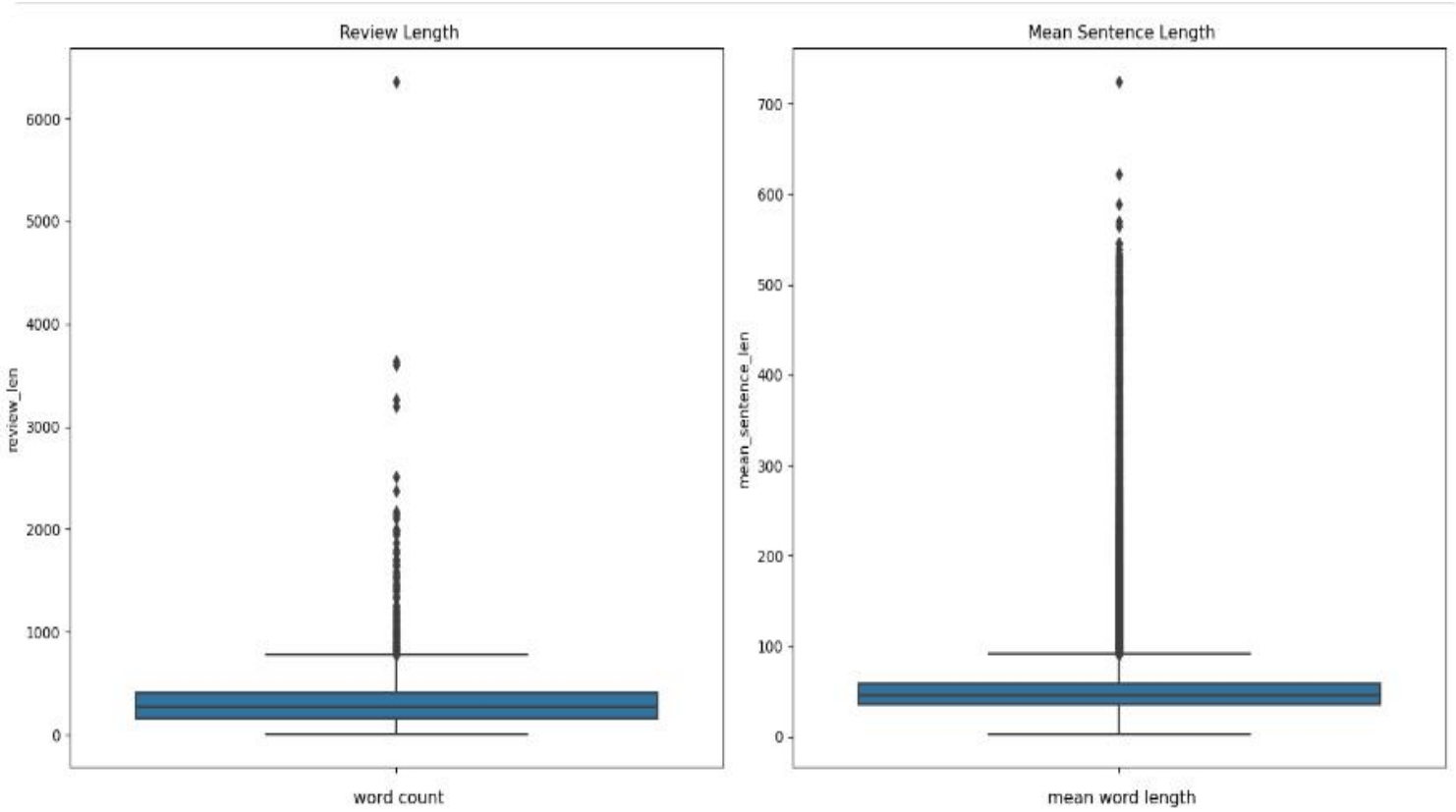
- Missing values, duplication (drop over 85000k)
- Corrected input error in “drugName” and “condition”
- “revew” column text data cleaning: special characters, contraction, whitespace, stopwords

Data distribution



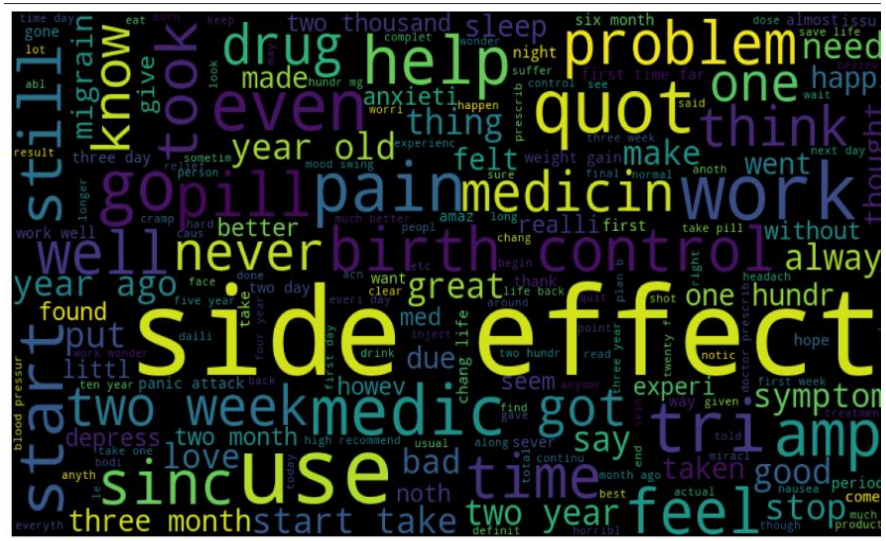



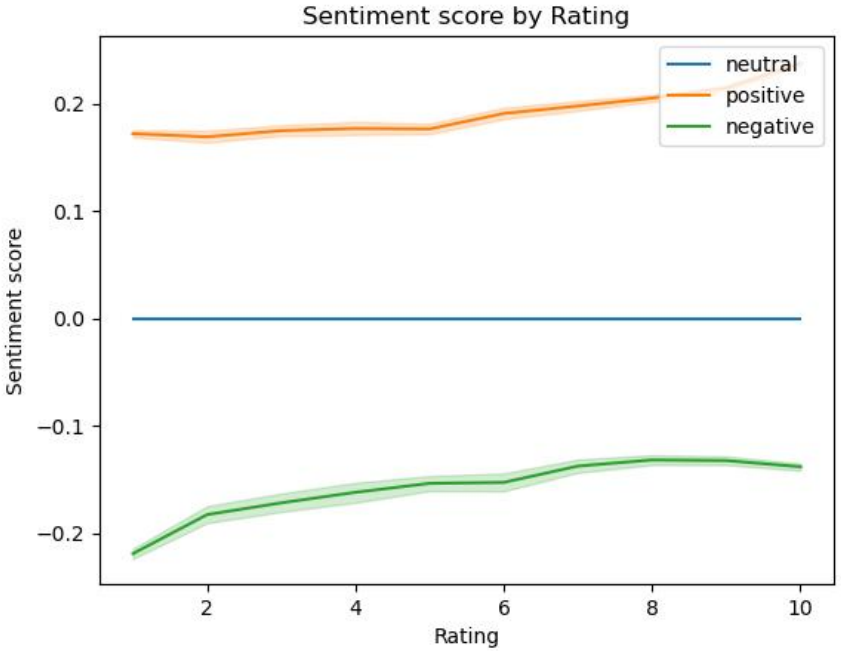
# Data Wrangling




Add new features into dataset:  'word_count', 'mean_word_len', 'unique_word_count'

# Data Wrangling




wordcloud with rating = 10

- wordcloud shows the top words in rating =10, reveal the high rating related to the mecial effective: such as side effect, help, better....
- Target feature sentiment_label was added.

# Preprocess

- Tokenize  the'review_clean'
- Encode the categorical features and target 'sentiment_label'
- Extract the 'date' to several new features 'year','month','day'.
- Scale the numerical features using MinMaxScaler.
- Train test split

# Model

- Multinomial Naive Bayes (MNB):baseline model
- Long Short-Term Memory (LSTM):
- Random Forest (RF):

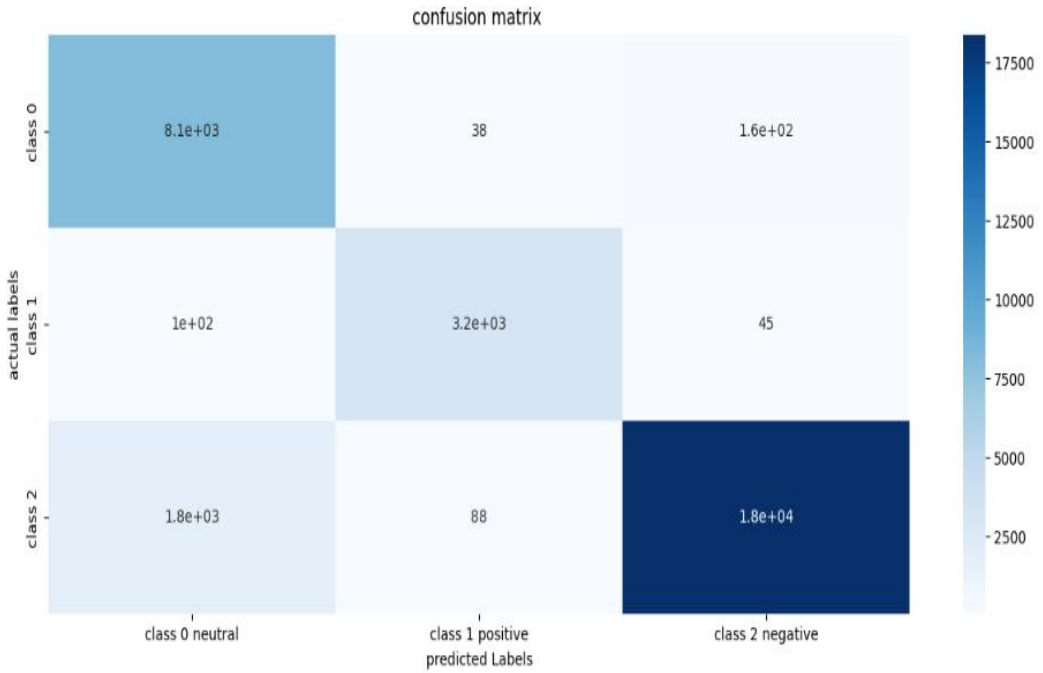
Evaluation metrics: F1_score

# LSTM

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 214, 32)           160000

 lstm (LSTM)                 (None, 100)               53200

 dropout (Dropout)           (None, 100)               0

 dense (Dense)               (None, 3)                 303

=================================================================
Total params: 213503 (834.00 KB)
Trainable params: 213503 (834.00 KB)
```

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.98 | 0.88 | 8321 |
| 1 | 0.96 | 0.96 | 0.96 | 3384 |
| 2 | 0.99 | 0.91 | 0.95 | 20292 |
| accuracy | | | 0.93 | 31997 |
| macro avg | 0.92 | 0.95 | 0.93 | 31997 |
| eighted avg | 0.94 | 0.93 | 0.93 | 31997 |




The F1 score and other scores are high, the dracback is the longer training time, >~1 hours

# Model Metrics

**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 |

LSTM classifier shows the optimal performance and can be selected as the final model.

# Future Work

- Selecting additional data for final modeling, utilizing the remaining dataset for training and modeling purposes will be explored .
- Reconfiguring the training and modeling process to focus on extracting feature importances.