Drug Review Sentiment Analysis

Team 3: Haotian Wu, Miao Fang, Jiawei Du
MScA 32009

Our Project Github: https://github.com/whtwht97/Health-Analytics/

01

Background

- Healthcare Problem
- Existing & Future Solutions
- Data Overview

02

Data Processing

- EDA
- Feature Engineering

03

Model Building

- Model Overview
- Model Selection

04

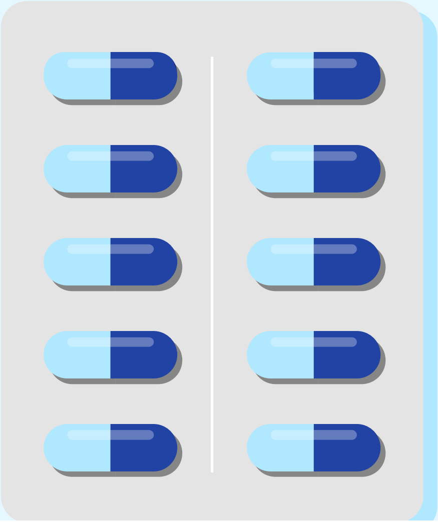
Conclusion

- Healthcare Impact
- Future Work

01 Background

- Healthcare Problem
- Existing & Future Solutions
- Data Overview

Healthcare Problem



Evaluation on drugs and their adverse drug reactions (ADRs) during clinical trials are limited to **standardized conditions** in a **limited test subjects** within a **limited time span**.

- Discrepancies in patient selection and treatment conditions can have significant impact on the evaluation of effectiveness and potential risks of ADRs.

  - **Pharmacovigilance**, post-marketing drug surveillance, plays a major role concerning drug effectiveness and safety

Existing & Future Solutions

Passive Surveillance Programs

Voluntary ADR report to regulatory agencies by patients, HC providers and drug manufacturers. Upon analysis, agencies inform drug hazard warnings.
Example: US FDA - MedWatch

Drawbacks:
underreporting of ADR, latency of notification

Lexicon-based Sentiment Analysis

Recognize sentiment expressions in HC customers' natural language texts by matching textual units with opinion words in lexicons annotated for sentiment polarity.

Drawbacks: polarity of single term differ by context; not suitable for informal and user-expressed texts

ML - based Sentiment Analysis

Train classifiers to detect sentiment-polarity at sentences/document level.

Example: LR, SVM, ELMo+LR, ... **+ Our Model**

Goal: Use ML and leverage quantity and expediency of drug review data to identify customer sentiment and supplement the current Pharmacovigilance system

Data Overview

Data Source

UCL Machine Learning Repository

Datasets

Training Set: 161K
Testing Set: 53K
Drug.com 2008 - 2017

Attributes

1. drugName: name of drug
2. condition: name of condition
3. review: patient review
4. rating: 1-10 patient rating on drug
5. date: date of review entry
6. usefulCount: number of users who found review useful

02

Data Processing

- EDA
- Feature Engineering

Exploratory Data Analysis

Missing Values

Condition:

- 1194 missing values
- **<1%** of total set
- Dropped given immateriality

Other Parameters:

- None missing values

Repetitive Reviews

uniqueID	drugName	condition	review
2817	Cefixime	Sinusitis	"This drug got me well when NOTHING else would...
3855	Suprax	Sinusitis	"This drug got me well when NOTHING else would...

- Same drug with more than one alias
- Removed rows with duplicate reviews
- Resulted in a total 128K rows of dataset

Exploratory Data Analysis

Average Rating

Overall Average Drug Rating: 7.12

Drug Name	Rating
Privine	10
Zutripro	10
Drixoral Cold And Allergy	9.96

Word Cloud

Overall

Negative Review Only

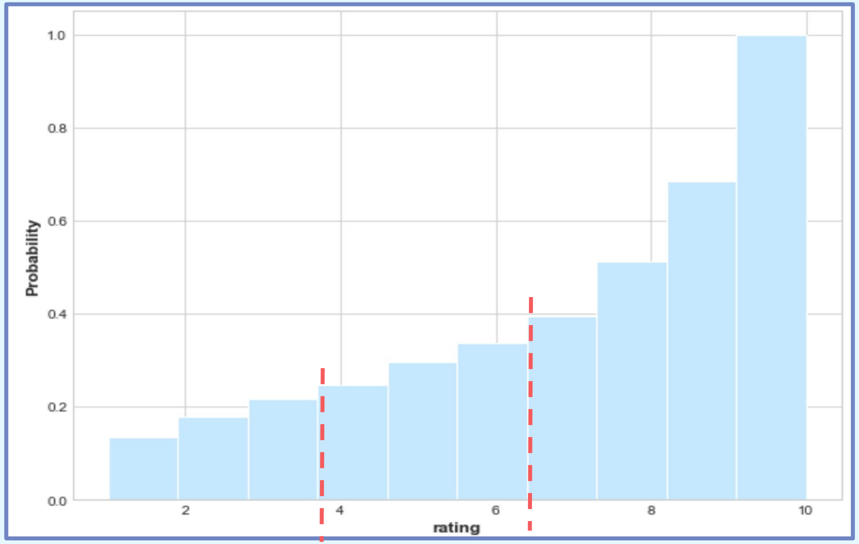
Exploratory Data Analysis

N-gram

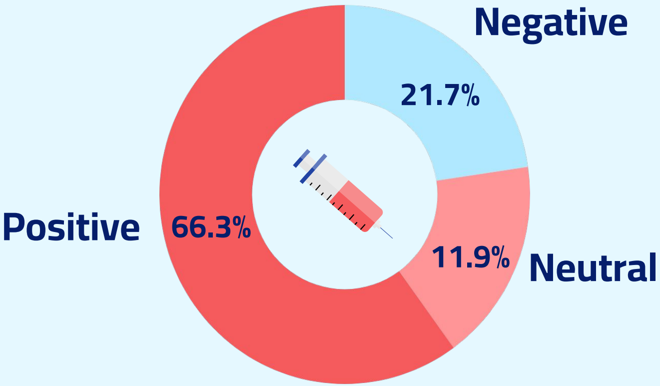
- Contiguous sequence of n items from a given sample of text
- Try out starting with one item
- Until **4-Gram**, start to see interpretability of emotions, thus used for modeling

Feature Engineering

- Ratings -> Sentiment Tag



- ❑ "Negative": 1~3
- ❑ "Neutral": 4~6
- ❑ "Positive": 7~10



Feature Engineering

- **Review Cleaning**

  - Shrink multiple spaces into 1 space.
    Only 2   pills   make me   feels better. –> Only 2 pills make me feels better.
  - Convert all of the characters to lowercase.
    Only 2 pills make me feels better. –> only 2 pills make me feels better.
  - Replace digits with special identifier.
    only 2 pills make me feels better. –> only DG pills make me feels better. √
  - Drop reviews which length is more than 150. (<1%)

03

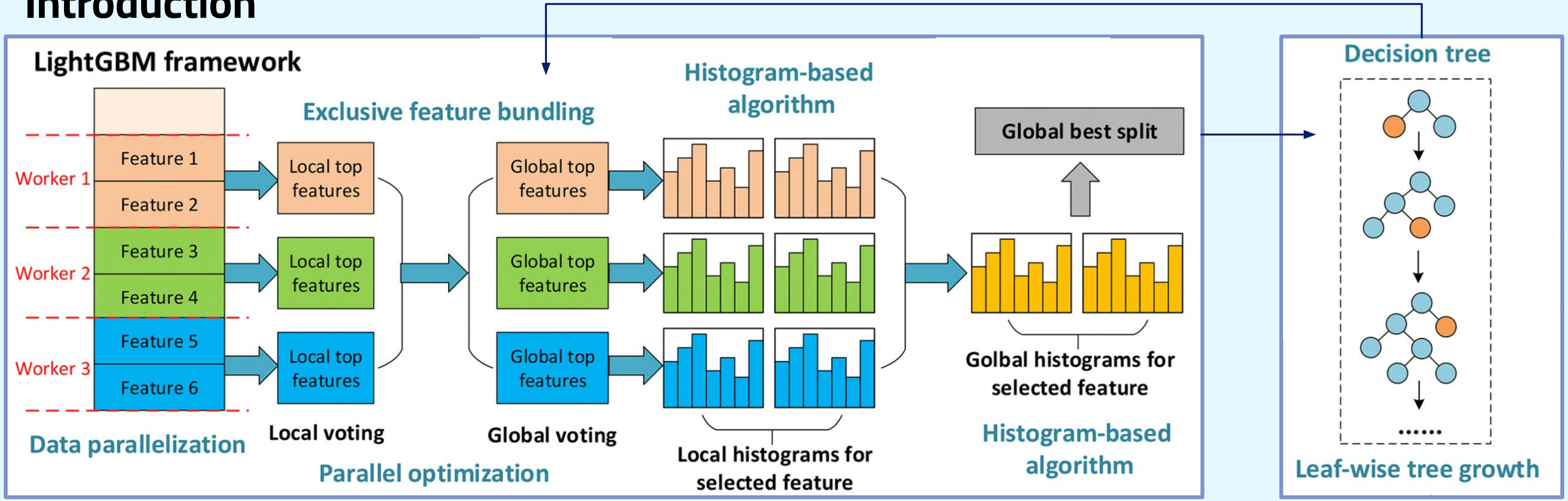
Modeling

- Model Overview
- Model Selection

Model Overview

LightGBM

- **Introduction**



- **Pros & Cons**

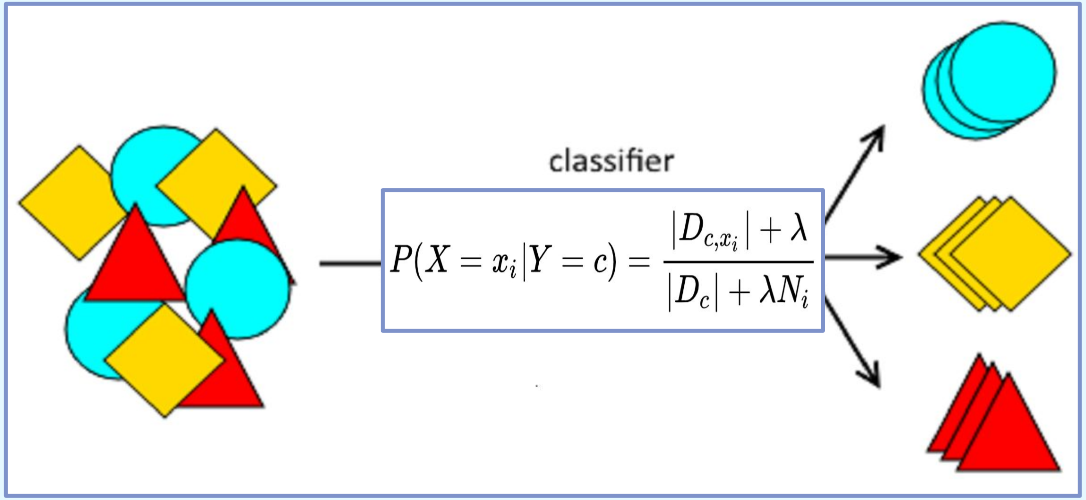
❑ Faster
❑ Less Memory

❑ Overfitting
❑ Sensitive to outliers

Model Overview

Naive Bayes

- **Introduction**



- **Pros & Cons**

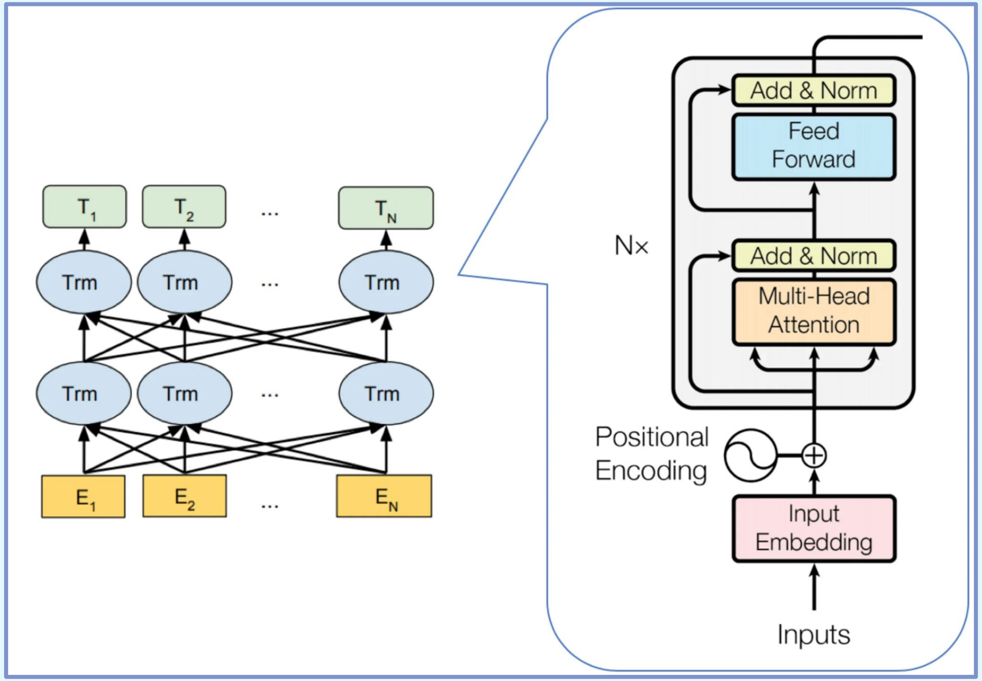
❑ Designed for text
❑ Lot faster than the plan version

❑ It is difficult to get the set of independent predictors

Model Overview

Bert

- **Introduction**

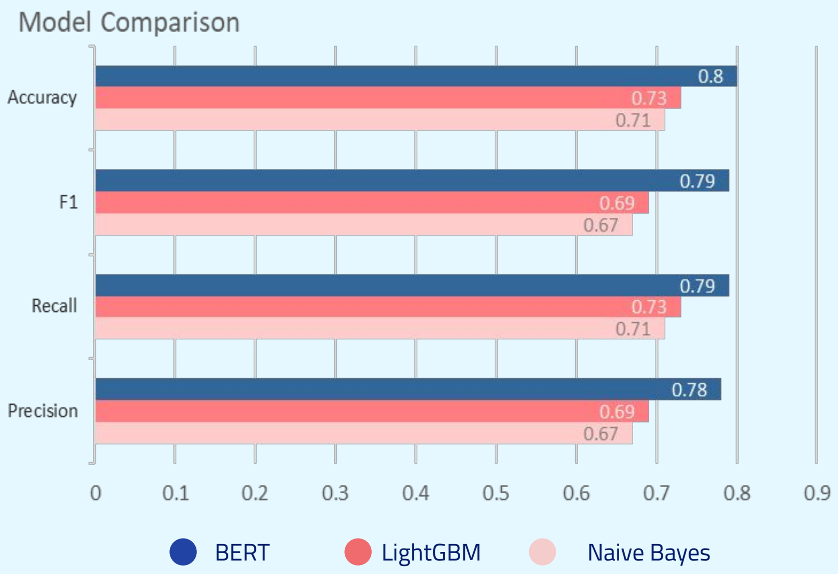


- **Pros & Cons**

❑ Use bi-directional learning to gain context of words from both left to right context and right to left context

❑ Maximum token length is 512, unable to to document-level task
❑ Difficult to do generative tasks
❑ Assumption of independence

Model Selection



Best Model:
BERT

Accuracy	0.80	Recall	0.79
F1	0.79	Precision	0.78

Fine Tuning

Strategy: Transformer

Optimizer	Adams	lr	1e-5
Epochs	4	Batchsize	32

prescription

04

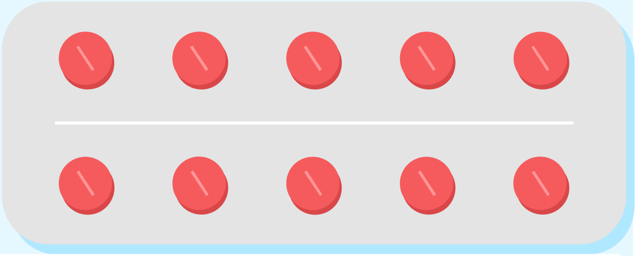
Conclusion

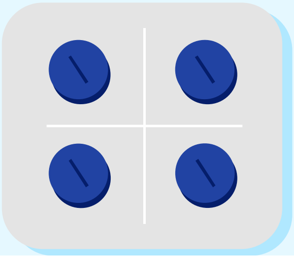
- Healthcare Impact
- Improvement Areas

Healthcare impact

Ensuring Medication Safety

- ❑ Reducing medication errors for physicians
- ❑ Post-Market Drug Surveillance for pharmco
- ❑ Obtain valuable summaries of public opinion for FDA





Effectiveness Evaluation

- ❑ Facilitating patients in making better informed purchase decision
- ❑ Product marketing insights for pharmaco
- ❑ Potential Drug Recommendation at prescription

Weakness

Future Work

Model Selection Consideration

Need more advanced and fit model/technique to improve the classification accuracy

Attempt Bio-Bert and Clinical-Bert

Patients' reviews may not use the same language as scientist & healthcare providers, but we need try to see how it actually performs

Little Improvement in Fine-Tuning

Fine-tuning doesn't help improve the model performance a lot

Learn & Practice

Will try larger number of epochs like 5,10, and different learning rate etc.

Thanks!