Drug Sentiment Analysis Using Review Text and User Ratings

Group 7

Lecturer: PhD. Le Thi Khuyen

Le Thu Hang
Mai Tra Giang
Nguyen Thi Tuyet May
Vo Thi Minh Phuong

May 10, 2025

Table Contents

Dataset Overview

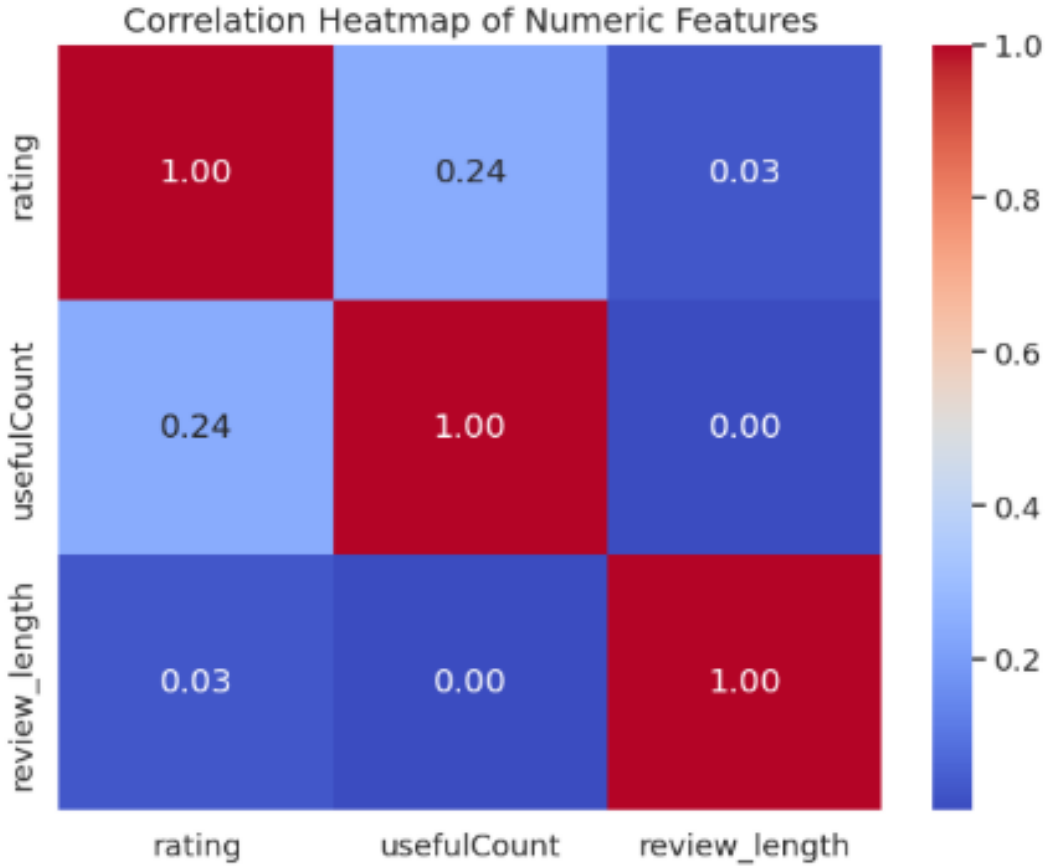
- The dataset contains 110,811 drug reviews from patients, including 4 object features, 4 numerical features.

	patient_id	drugName	condition	review	rating	date	usefulCount	review_length
0	89879	Cyclosporine	keratoconjunctivitis sicca	"i have used restasis for about a year now and...	2.0	April 20, 2013	69	147
1	143975	Etonogestrel	birth control	"my experience has been somewhat mixed. i have...	7.0	August 7, 2016	4	136
2	106473	Implanon	birth control	"this is my second implanon would not recommen...	1.0	May 11, 2016	6	140
3	184526	Hydroxyzine	anxiety	"i recommend taking as prescribed, and the bot...	10.0	March 19, 2012	124	104
4	91587	Dalfampridine	multiple sclerosis	"i have been on ampyra for 5 days and have bee...	9.0	August 1, 2010	101	74

- Our problem is to predict positive or negative sentiment from review text by first converting ratings into binary sentiment labels base on drug dataset

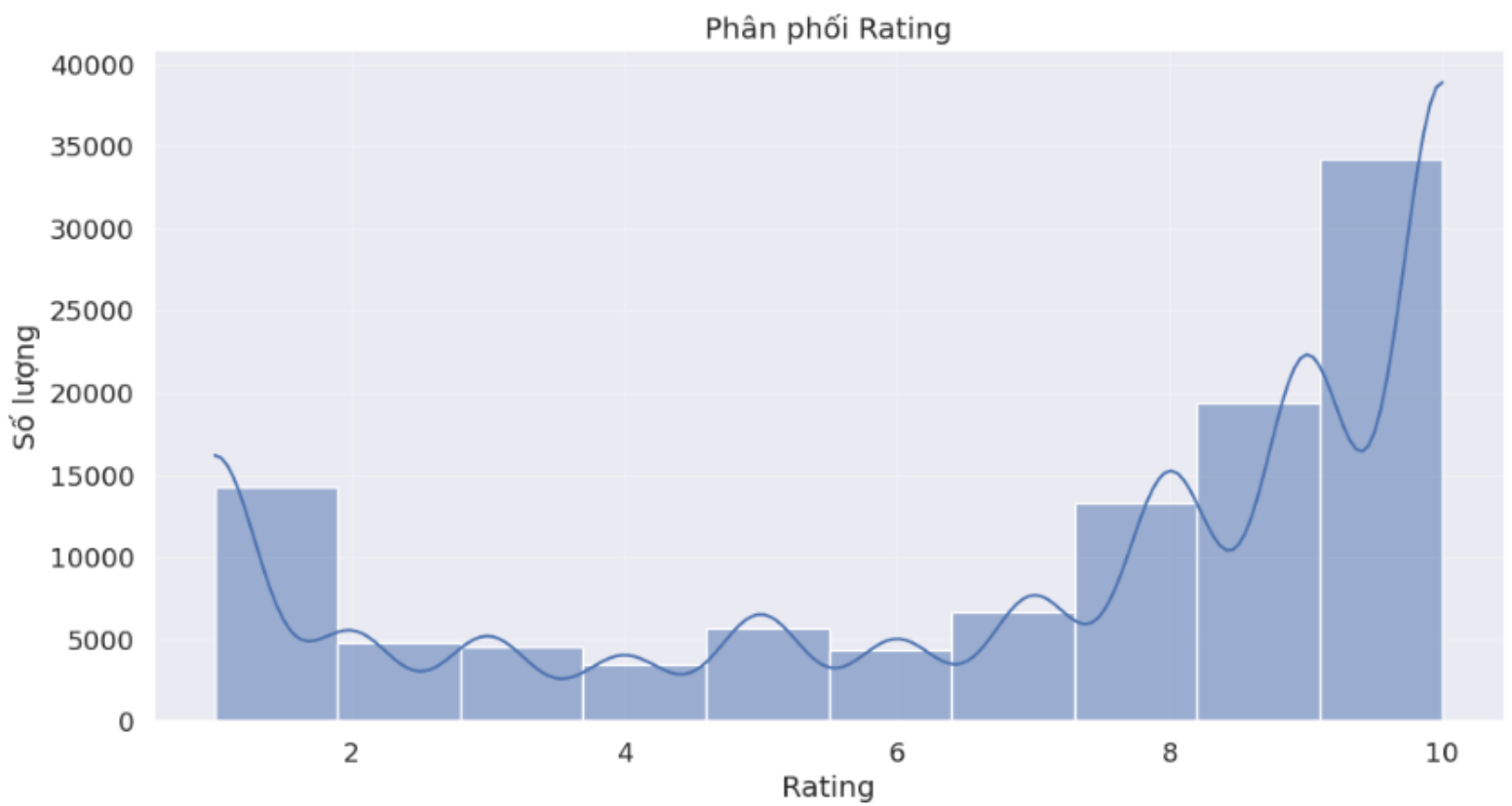
EDA - Heatmap correlation

The length of the review has the lowest correlation with all other features (0.03 with rating, 0.00 with usefulCount), so it should be dropped



EDA - Rating Score

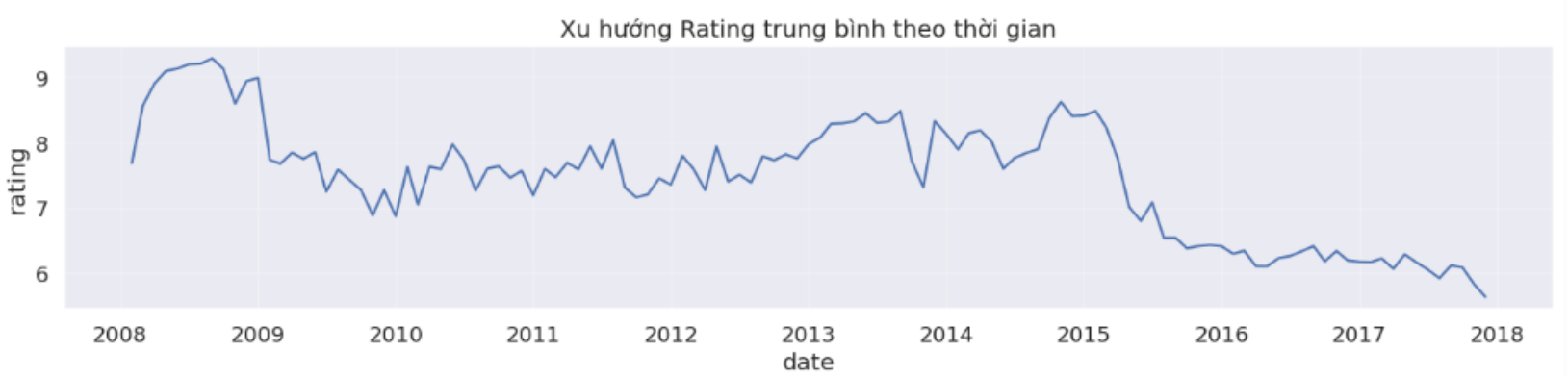
- Rating Distribution

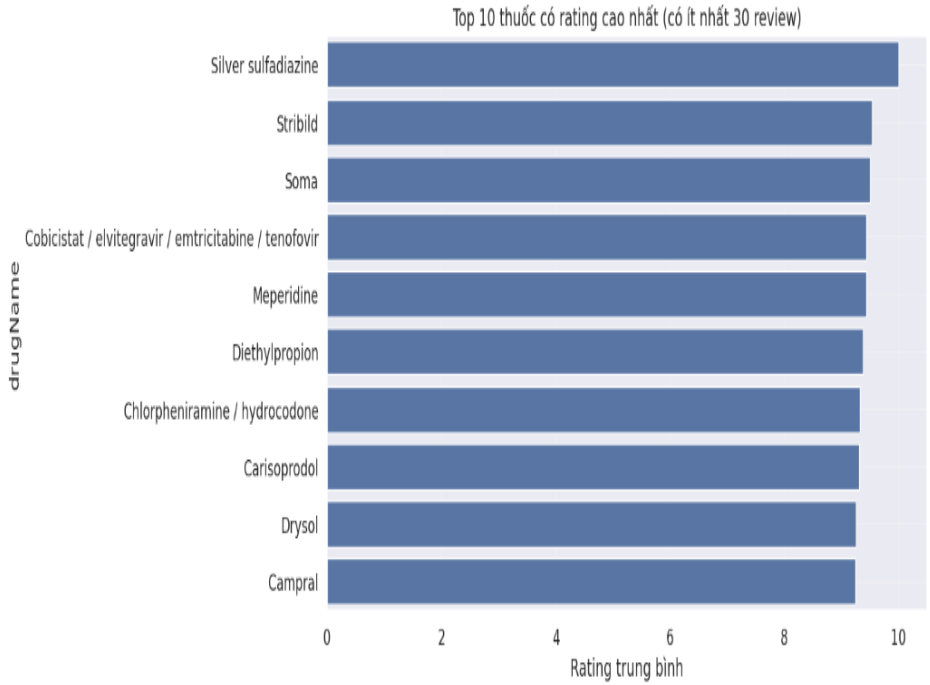


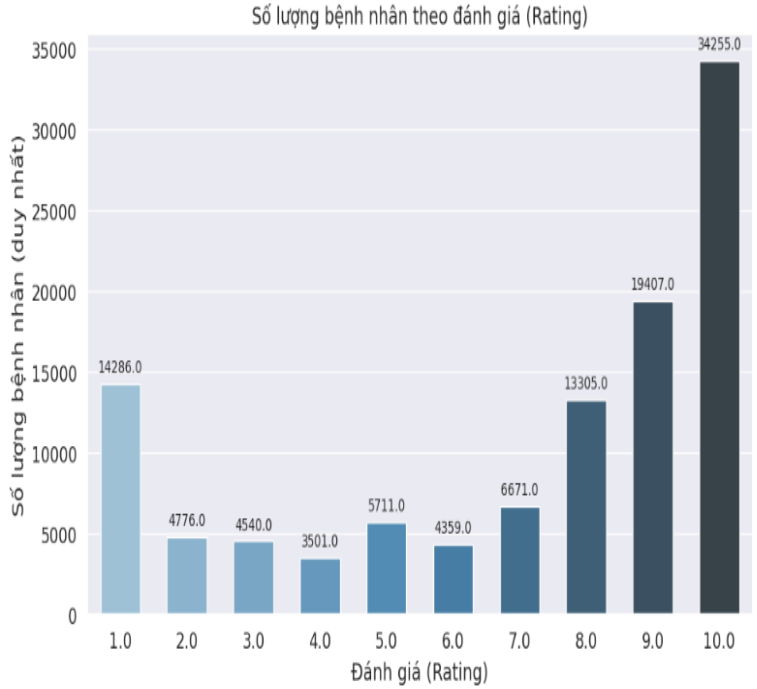
The rating distribution is highly imbalanced, with a significant concentration at the highest score (10), indicating a strong positive bias in user reviews.

EDA - Rating Score

- Related features with rating score

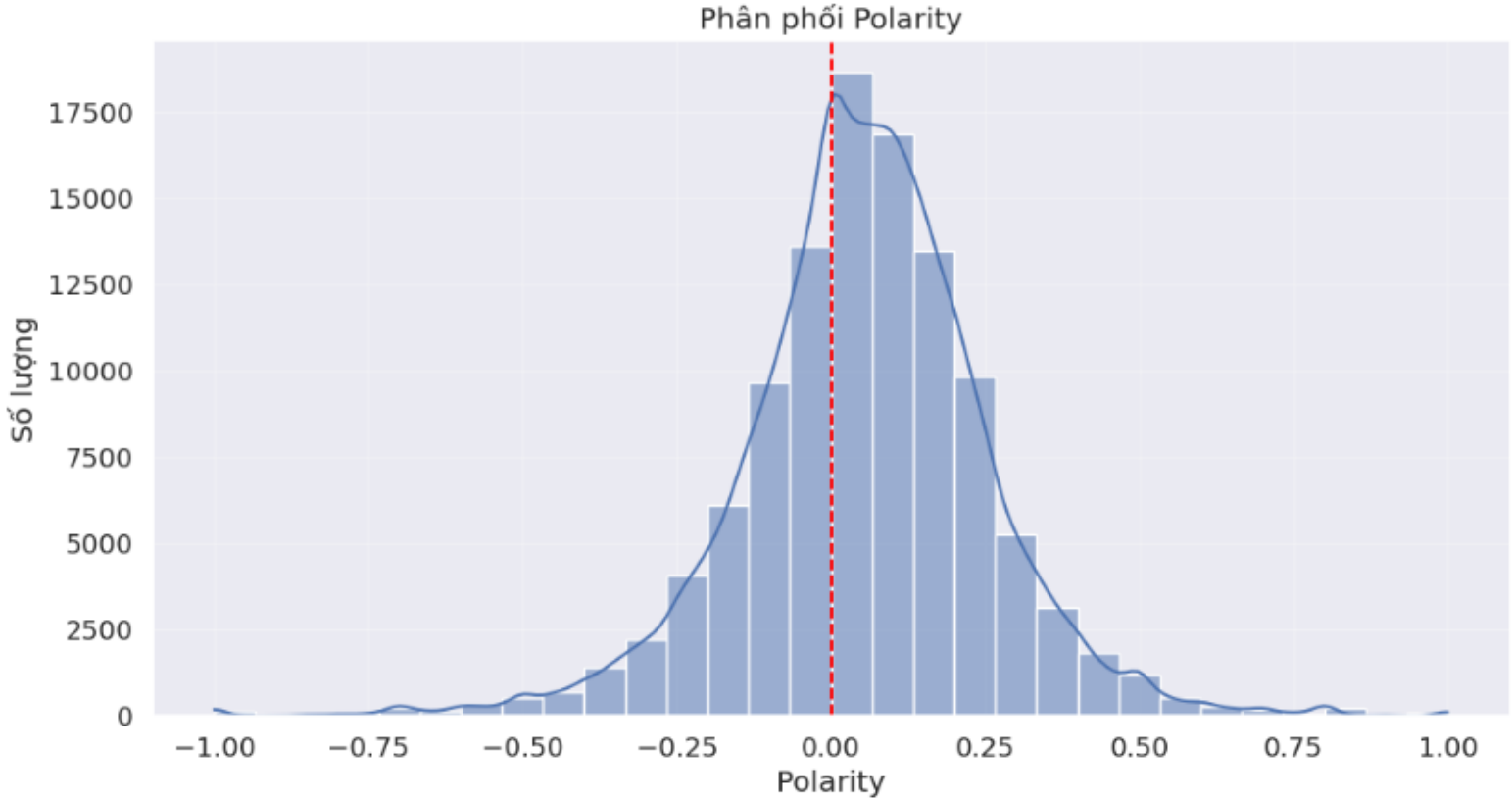






EDA - Polarity index

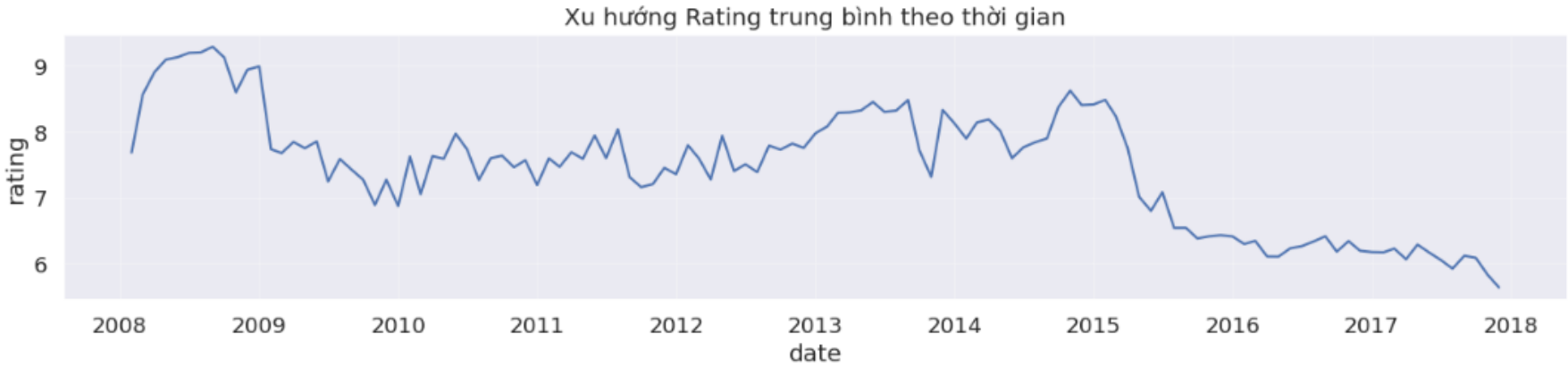
- Polarity index is a measure in sentiment analysis to determine how positive or negative a text is base on review text.

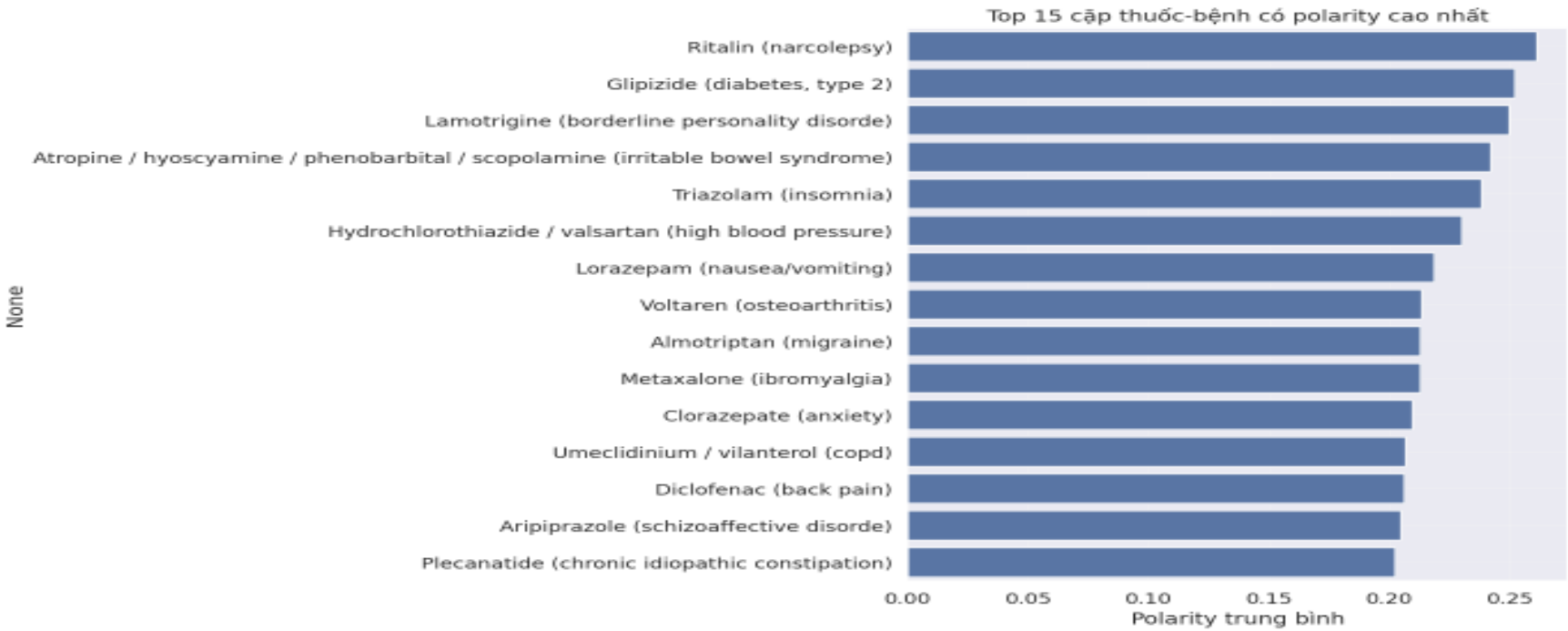


The bell-shaped chart shows that most polarity values cluster around 0, indicating neutral or mixed reviews. However, the distribution slightly leans positive, as the area on the right side of zero appears slightly larger.

EDA - Polarity index

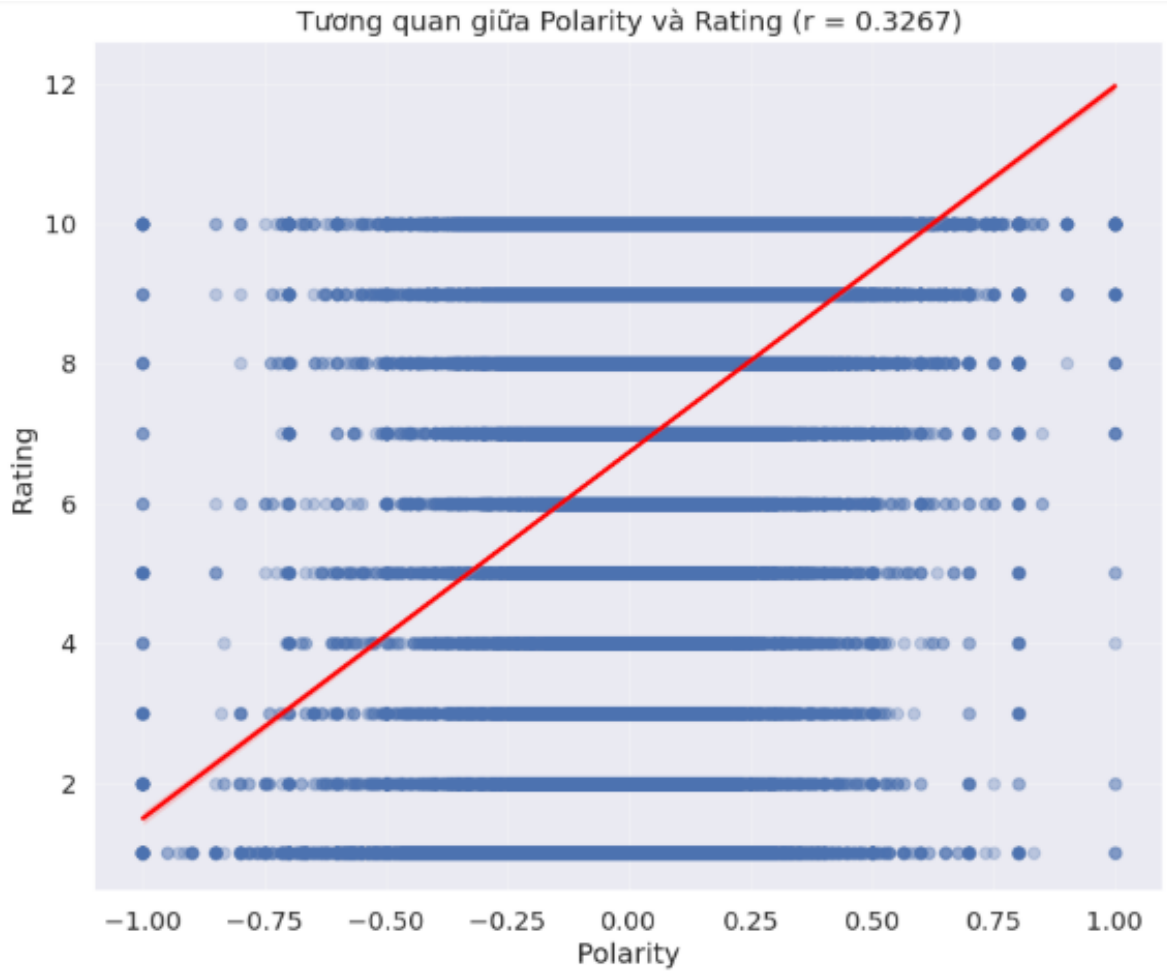
- Related feature with polarity index





EDA - Rating and Polarity relationship

There is a weak positive correlation($r = 0.3267$)between polarity and rating,suggesting that more positive reviews tend to have higher ratings.



Data Pipeline

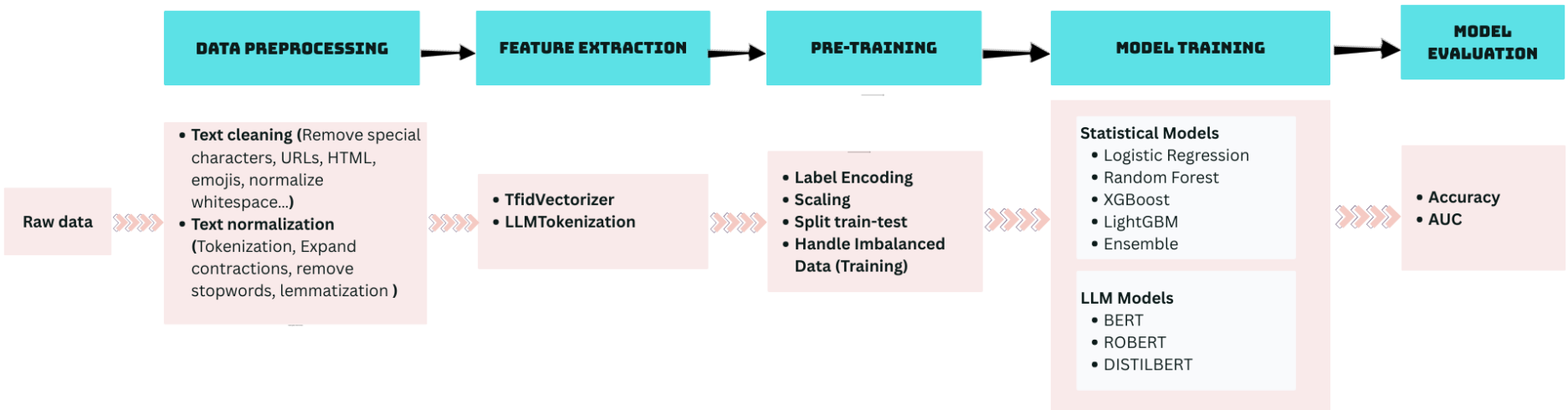


Figure: Overall NLP Model Pipeline

Text Preprocessing

- Convert all text to lowercase to ensure consistency.
- Remove HTML tags, email addresses, URLs, emojis, and special characters.
- Correct common spelling errors and remove duplicate entries.
- Tokenize sentences into words using NLTK.
- Expand contractions (e.g., won't $\rightarrow$ will not), normalize whitespace.
- Remove stopwords (e.g., "the", "is", "in") to reduce noise.
- Apply lemmatization to reduce words to their base form (e.g., "running" $\rightarrow$ "run").
- Detect named entities, sarcasm, and perform synonym replacement.

Stat-Model Setup

- Convert ratings into binary classes: negative (< 5) and positive ($>= 5$).
- Apply TF-IDF vectorization to transform text into numerical features.
- Use this setup to build a basic text classification problem.
- Split the dataset into training and test sets before applying the pre-trained model.

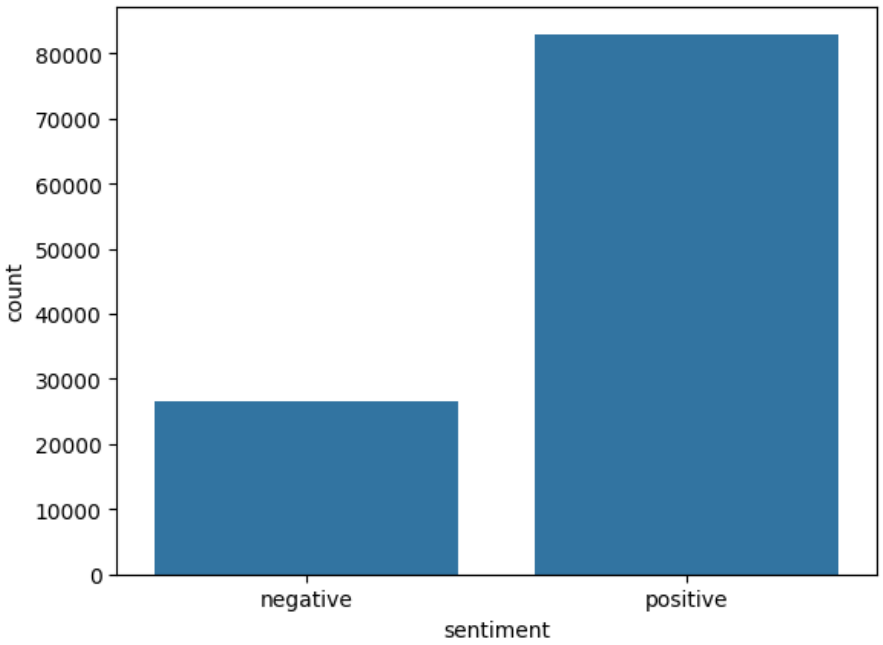


Figure: Sentiment Distribution

Stat-Imbalanced Data Handling

- Use Random OverSampling (ROS) to replicate minority class samples.
- Apply SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic samples.
- Resampling is performed only on the training set to avoid data leakage.

Stat-Training Models

- Train multiple classification models on TF-IDF features:
  - Logistic Regression
  - Random Forest
  - XGBoost
  - LightGBM
- Evaluate individual model performance on the test set.
- Combine models using ensemble techniques:
  - Voting Classifier (hard/soft voting)
  - Stacking Classifier (meta-model approach)
- Select the best-performing model for deployment

LLM-Pre-trained

- Use the `token_clean_contracted` column as input for the language model.
- Convert rating into binary labels: positive ($>= 5$) and negative (< 5).-
- Split the dataset into (80/20) train and test sets before fine-tuning the model.

LLM-Imbalanced Handling

- **Multi-technique class balancing approach:**
  - Three-pronged text augmentation for minority class:
    - WordNet synonym replacement
    - Back-translation (EN→FR→EN)
    - Easy Data Augmentation (random word swap)
  - Weighted loss function calculation based on class distribution

LLM-Model Training

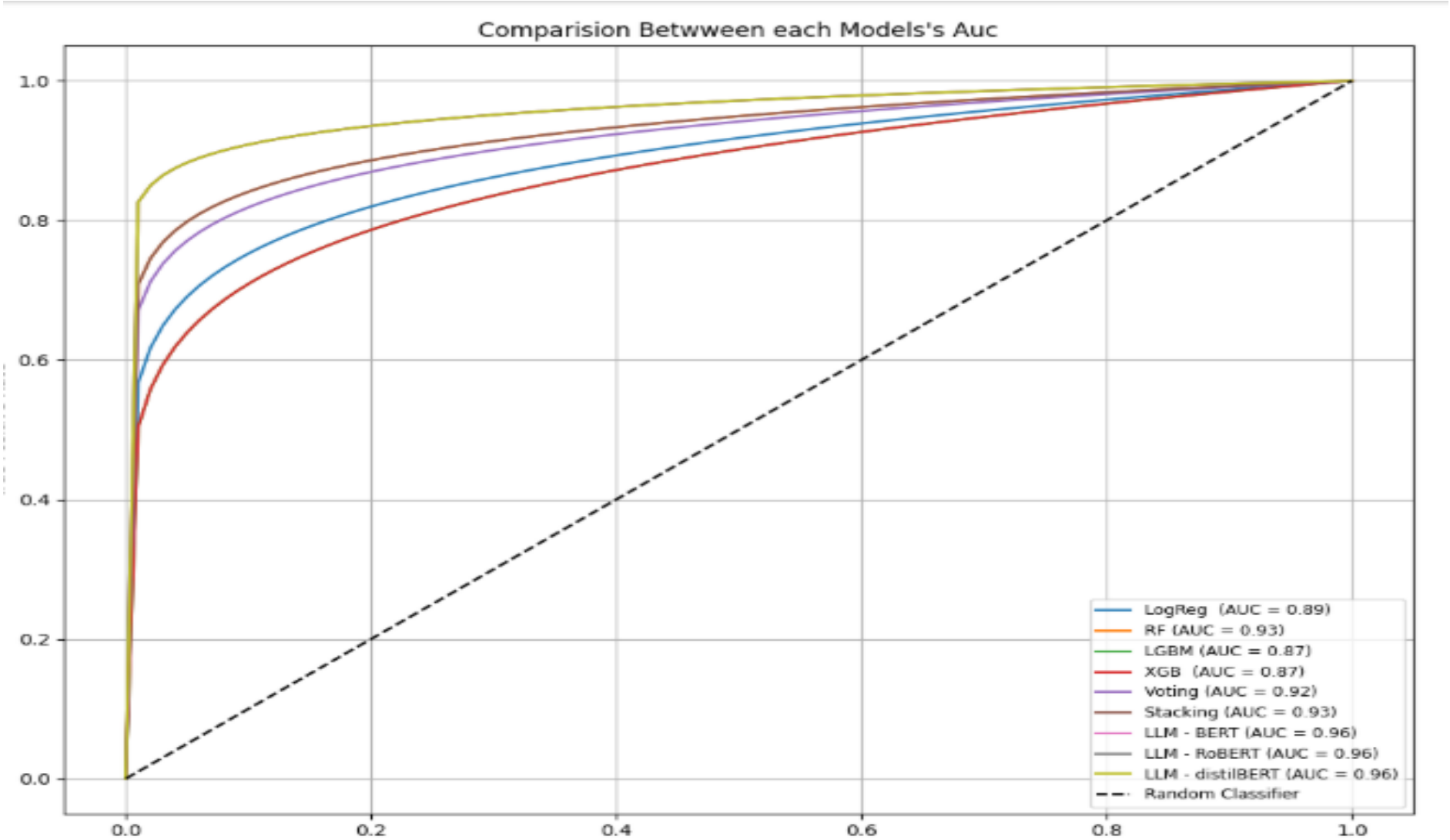
- **Dataset preparation:**
  - Each model tokenizes with consistent parameters
  - TensorDataset construction with encoded inputs and labels
  - WeightedRandomSampler for additional sampling balance
  - DataLoader with appropriate batch sizes and a weighted sampler to ensure balanced training despite class imbalance
- **Model training pipeline:**
  - Three pre-trained models: RoBERTa, BERT, DistilBERT
  - Using a unified training pipeline with Focal Loss to better handle class imbalance.
  - Consistent evaluation metrics : Accuracy, AUC

Results

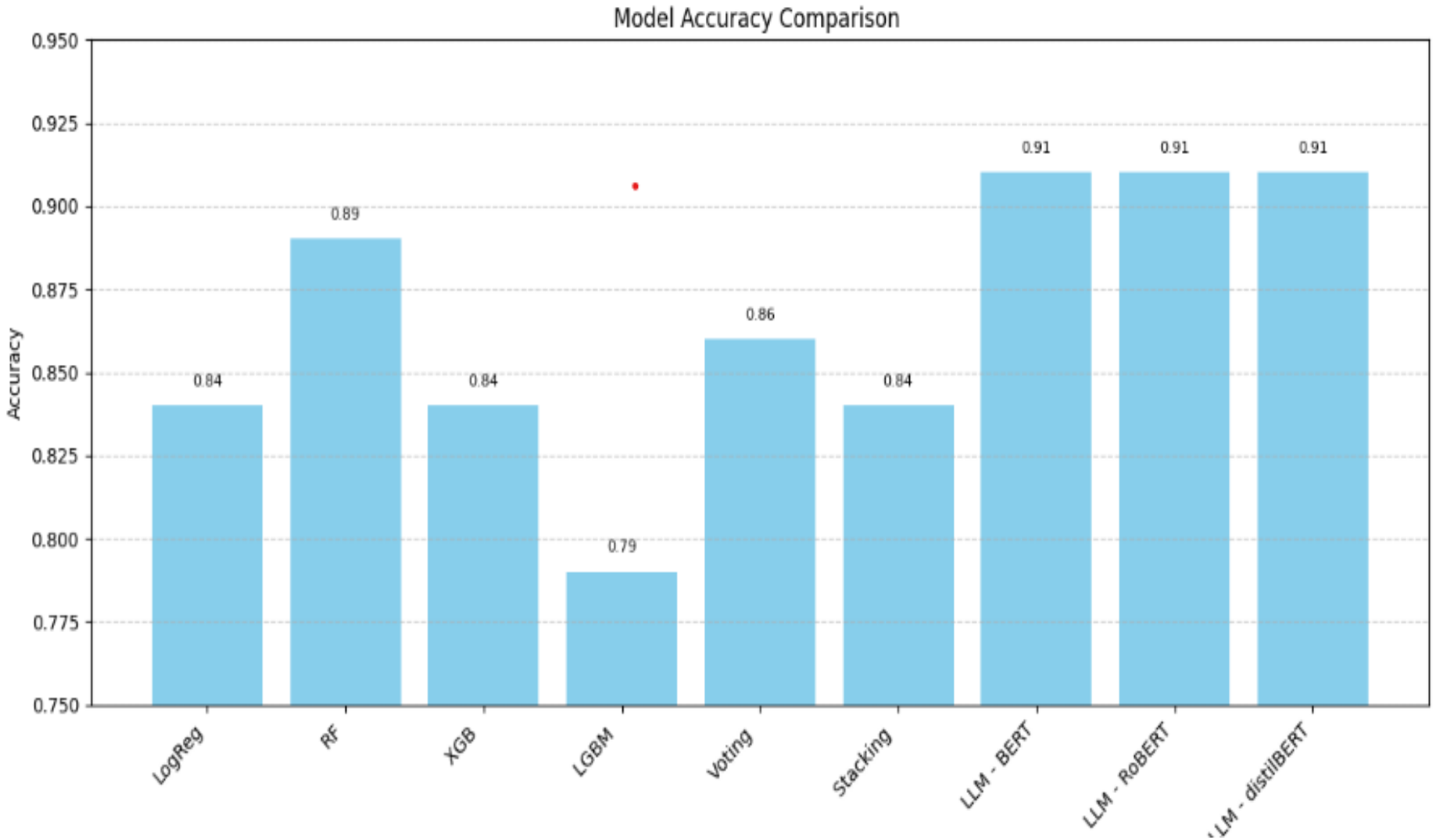
	Model	Accuracy	Auc
Statistical Model	Logistics Regression (ROS)	**0.89**	0.89
	Random Forest (ROS)	**0.89**	**0.93**
	LightGBM (ROS)	0.79	0.87
	XGBoost (ROS)	0.84	0.87
	Logistics Regression (SMOTE)	0.82	0.89
	Random Forest (SMOTE)	**0.89**	**0.93**
	LightGBM (SMOTE)	0.79	0.87
	XGBoost (SMOTE)	0.79	0.87
	Voting Classifier	0.86	0.92
	Stacking Classifier	0.84	**0.93**
LLM	LLM - BERT	**0.91**	**0.96**
	LLM - RoBERT	**0.91**	**0.96**
	LLM - distilBERT	**0.91**	**0.96**

Table: **Model Results**

Results In chart



Results In chart



Conclusion

- **Statistical Models**
  - Random Forest and ensemble models (Voting, Stacking) perform well.
  - LightGBM and XGBoost show weaker performance, especially with imbalanced data.
- **LLM**
  - Large language models (LLMs) demonstrate outstanding performance in sentiment analysis that effectively capture context and emotional nuances in drug review texts.
  - RoBERTa delivers the best results, while DistilBERT is a lightweight and efficient alternative.
- **Balancing Techniques**
  - Statistical Model
    - ROS is better suited for text data, avoiding the creation of semantically unnatural synthetic samples.
  - LLM
    - Combining synonym replacement, back-translation, word swap, and weighted loss improves class balance and data diversity for better learning.

Future Work

- **More data:** Collect drug-related data for multi-task learning (e.g., classification, side effects, efficacy).
- **More features:** Add drug images, dosage charts, user behavior for richer prediction.
- **Better models:** Fine-tune LLMs on medical texts (PubMed, Medline, ClinicalTrials) for domain understanding.

Thank you!

Questions?