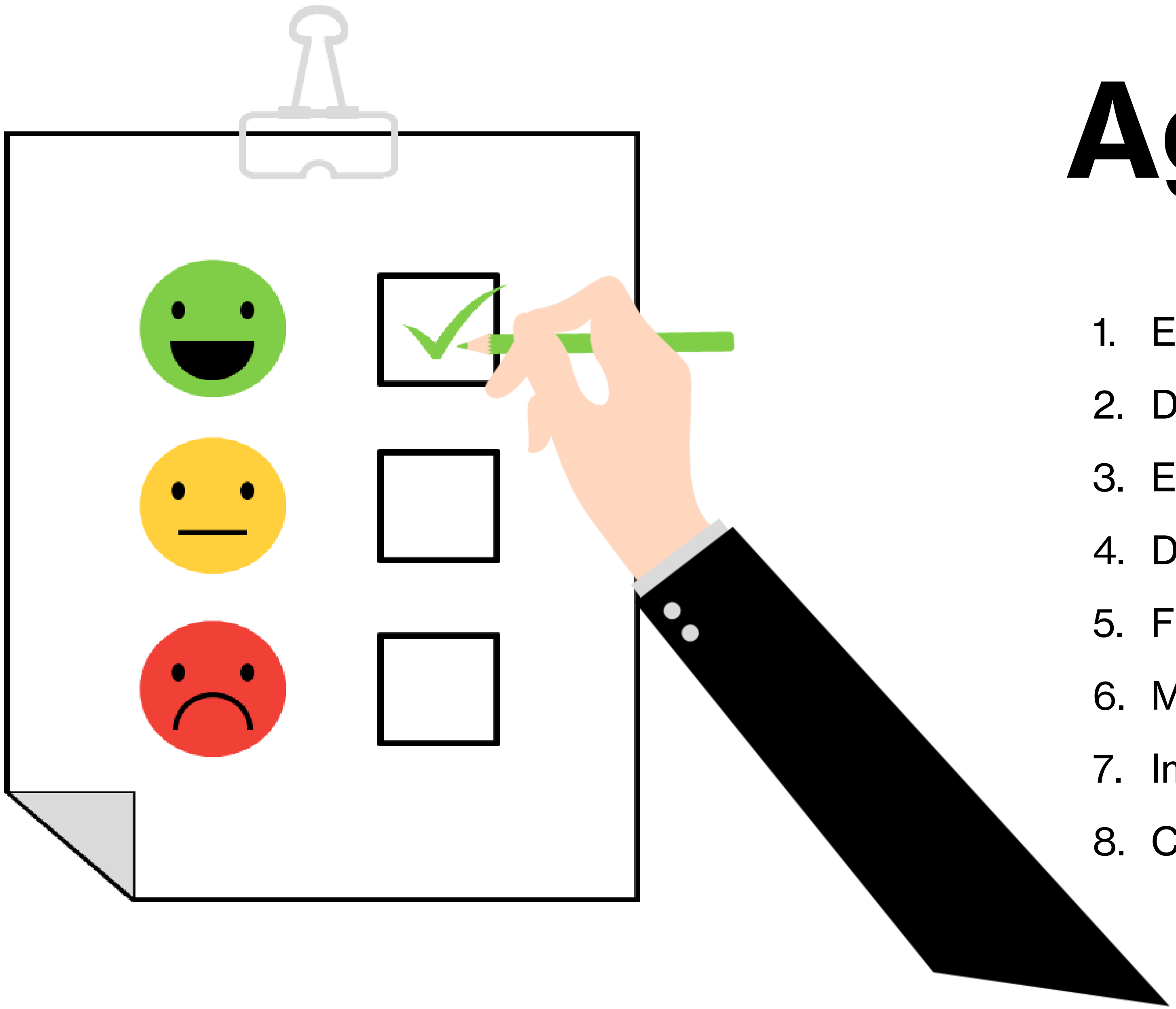
Understanding Customer Reviews

- Using NLP with Logistic Regression

Team Finch



Agenda

1. Executive Summary
2. Data Overview
3. EDA
4. Data Prepressing
5. Feature Engineering
6. Model Training
7. Improvements through LLM
8. Conclusion

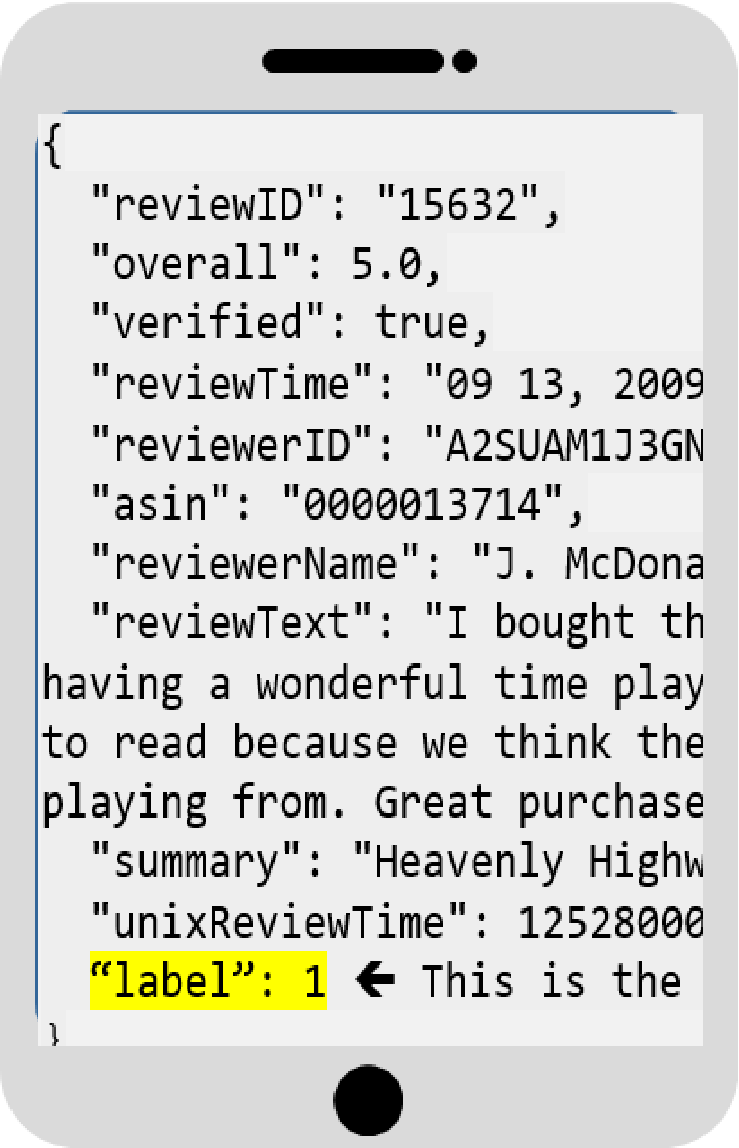
Executive Summary

- **Objective -** Explore and model customer reviews to find out which Amazon reviews are helpful

- **Methodology -** EDA and Data Preprocessing
  Feature Engineering
  Logistic Regression

- **Outcome** - Leaderboard AUC Score: 0.88905

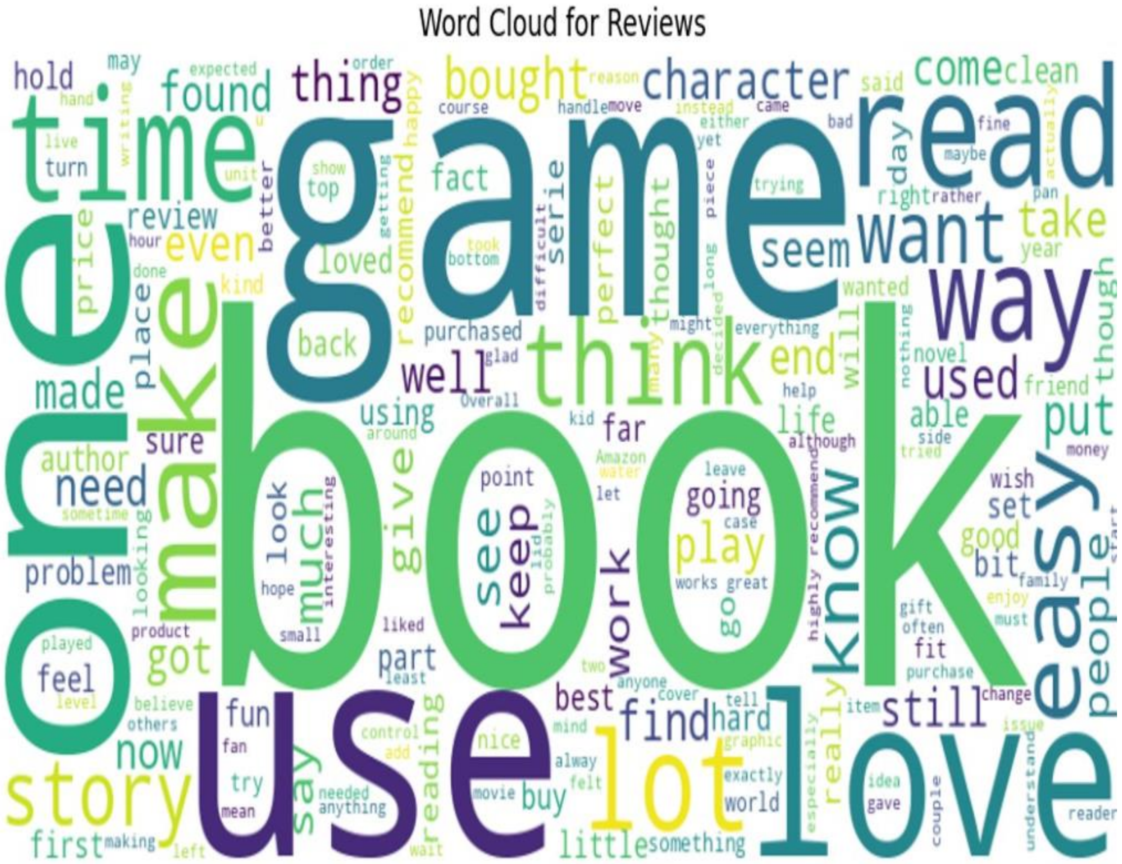
- **Conclusion** - **The timing and length of a review** play a crucial role in determining its helpfulness.



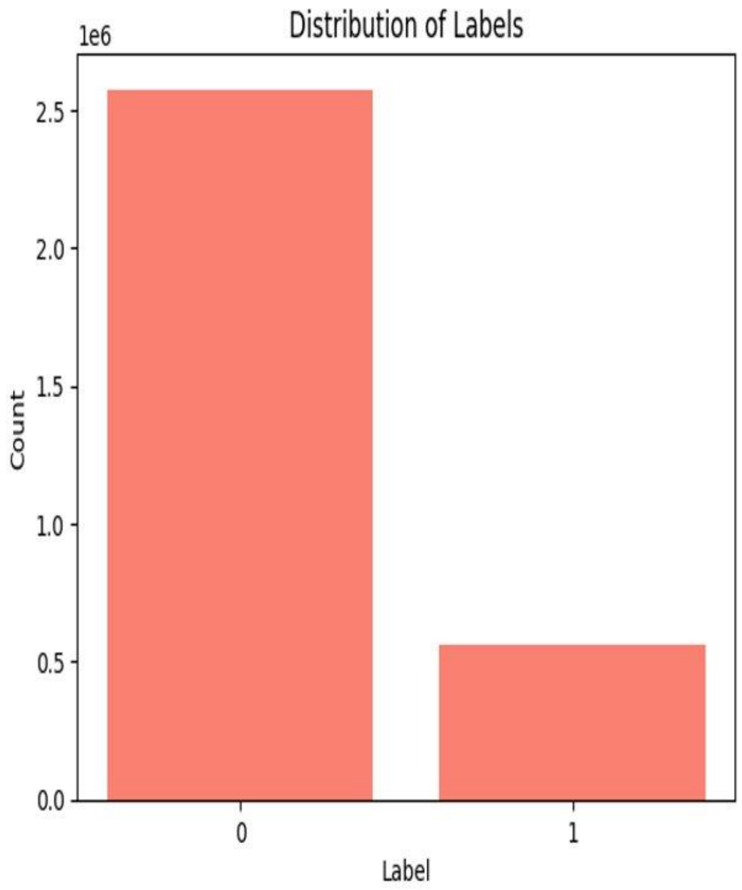
Data Overview

- **Total Reviews:** 3,138,710
- **Columns/Features :** 11
- **Time Span:** January 1998 to December 2017
- **# Unique Products:** 65,205
- **# Unique Reviewer ID:** 1,084,151
- **Missing Values:** reviewer name (217) & summary (351)
- **Average Review Length**: 408.76 characters
- **Longest Review Length:** 32712 characters
- **Shortest Review Length:** 1 character

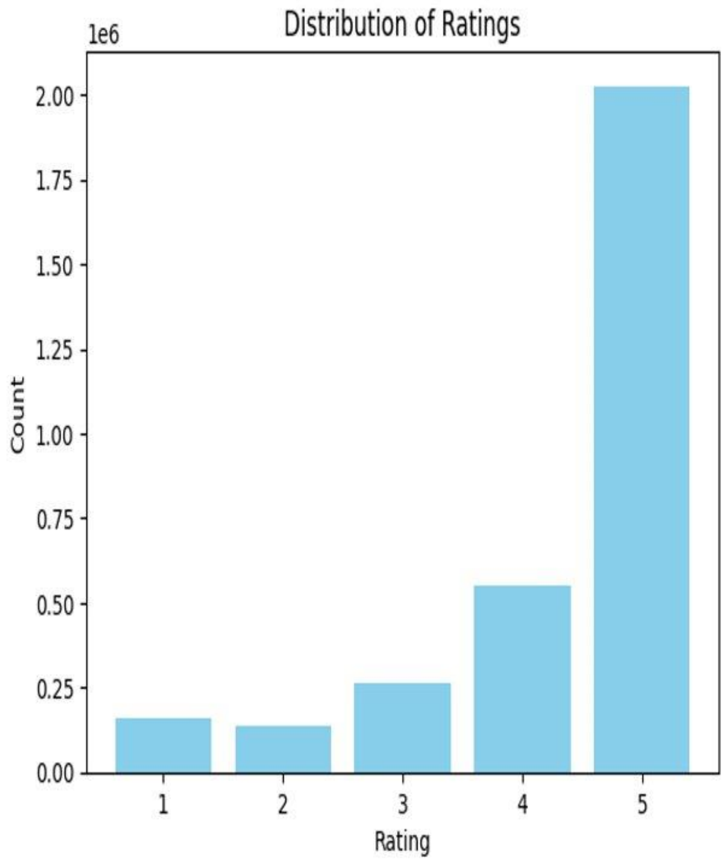
EDA – Word Cloud



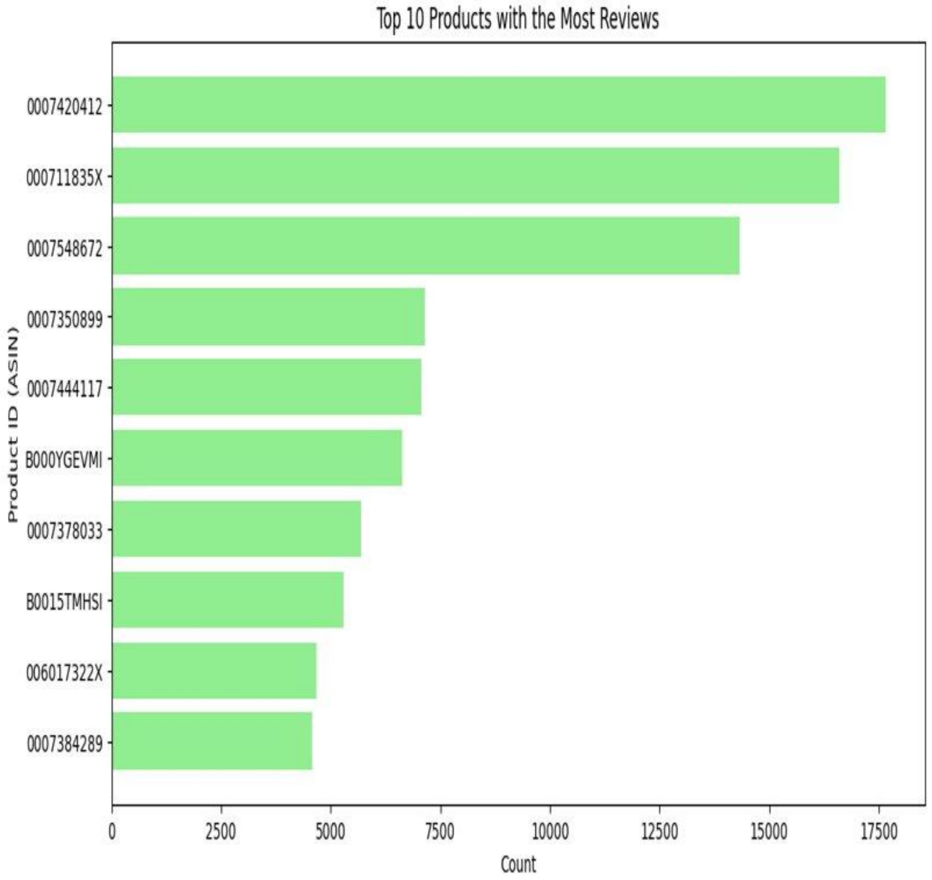
EDA



Distribution of Labels
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Count
0
1
Label

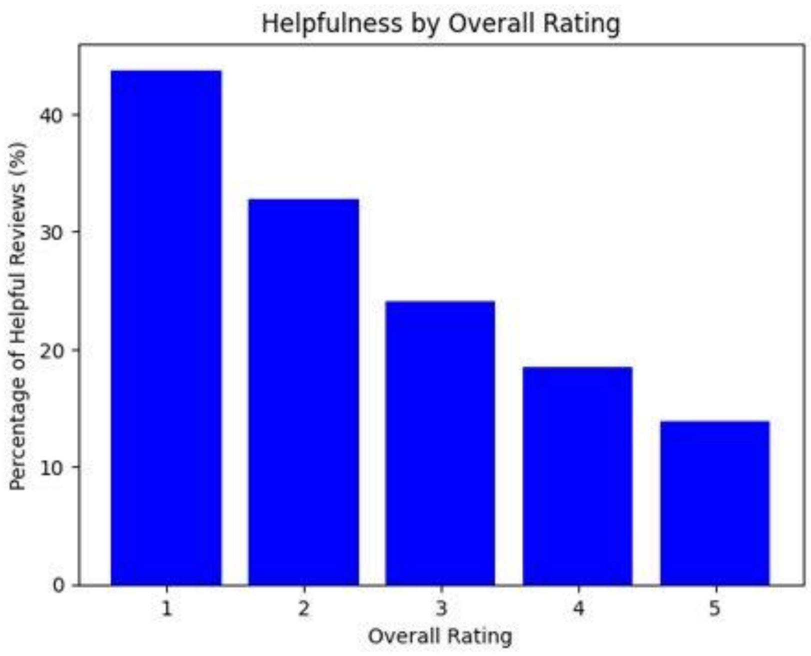


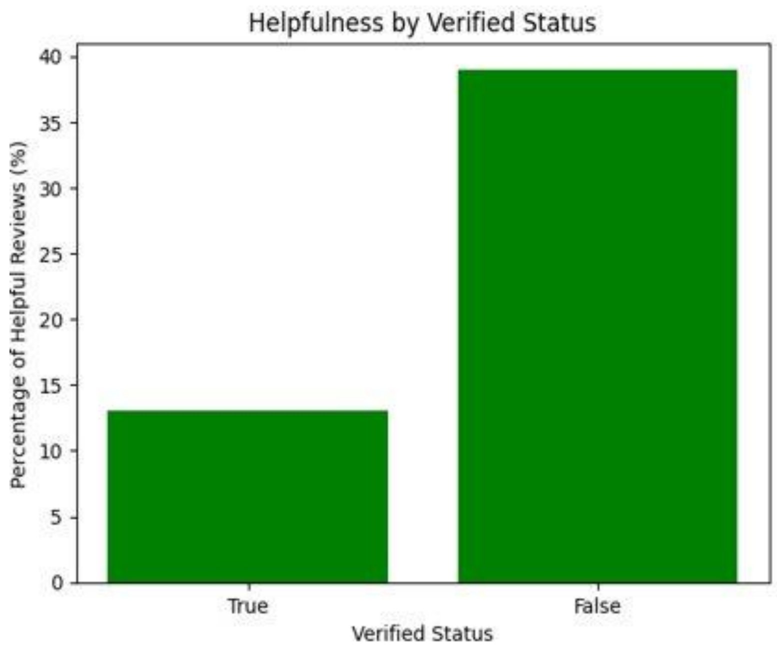
Distribution of Ratings
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Count
1
2
3
4
5
Rating

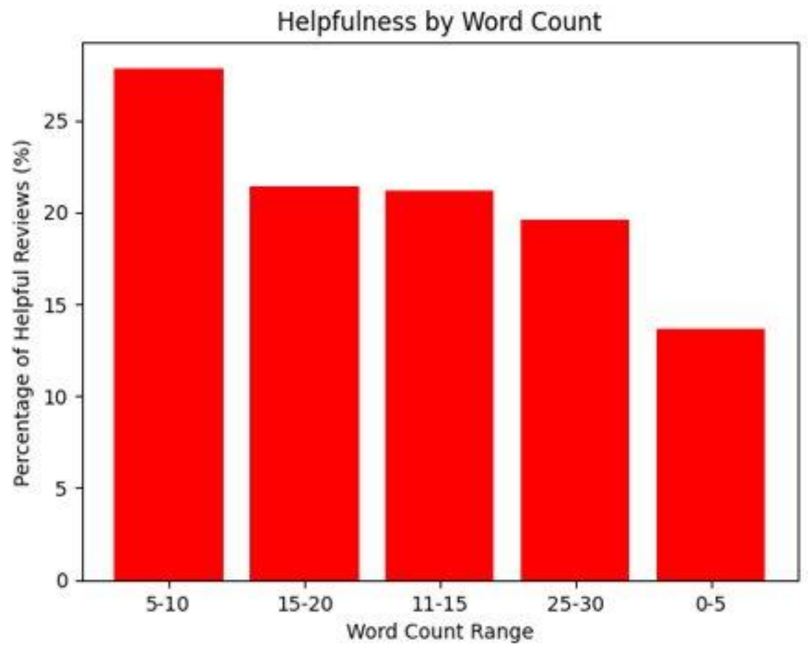


Top 10 Products with the Most Reviews
0007420412
000711835X
0007548672
0007350899
0007444117
B000YGEVMI
0007378033
B0015TMHSI
006017322X
0007384289
Product ID (ASIN)
0
2500
5000
7500
10000
12500
15000
17500
Count

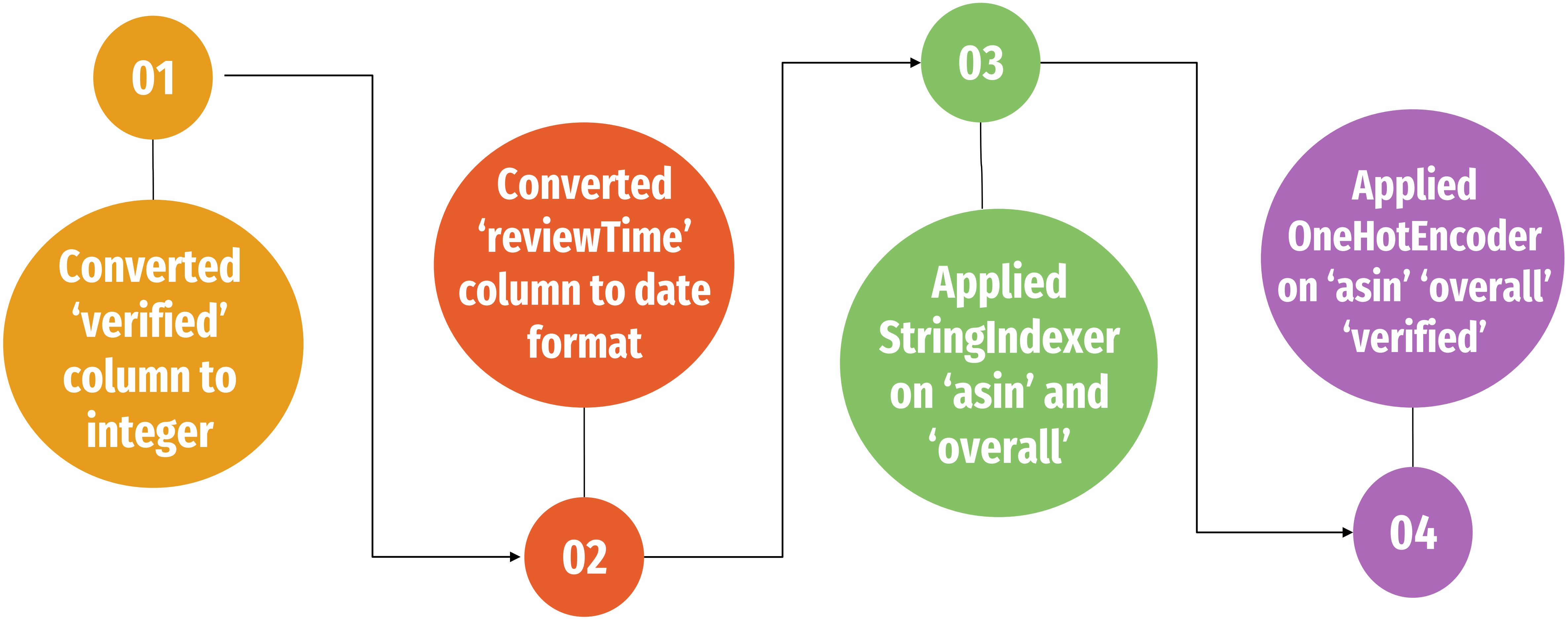
EDA


Helpfulness by Overall Rating
Percentage of Helpful Reviews (%)
0
10
20
30
40
1
2
3
4
5
Overall Rating


Helpfulness by Verified Status
Percentage of Helpful Reviews (%)
0
5
10
15
20
25
30
35
40
True
False
Verified Status

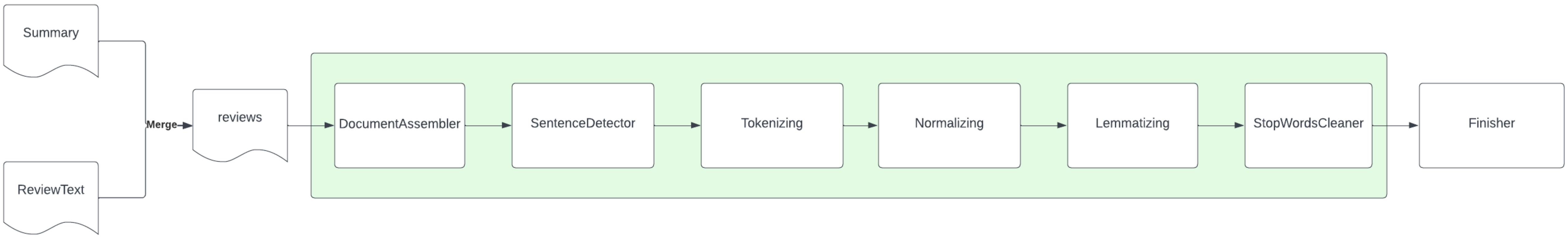

Helpfulness by Word Count
Percentage of Helpful Reviews (%)
0
5
10
15
20
25
5-10
15-20
11-15
25-30
0-5
Word Count Range

Data Prepressing (Non-Review Text & Summary)

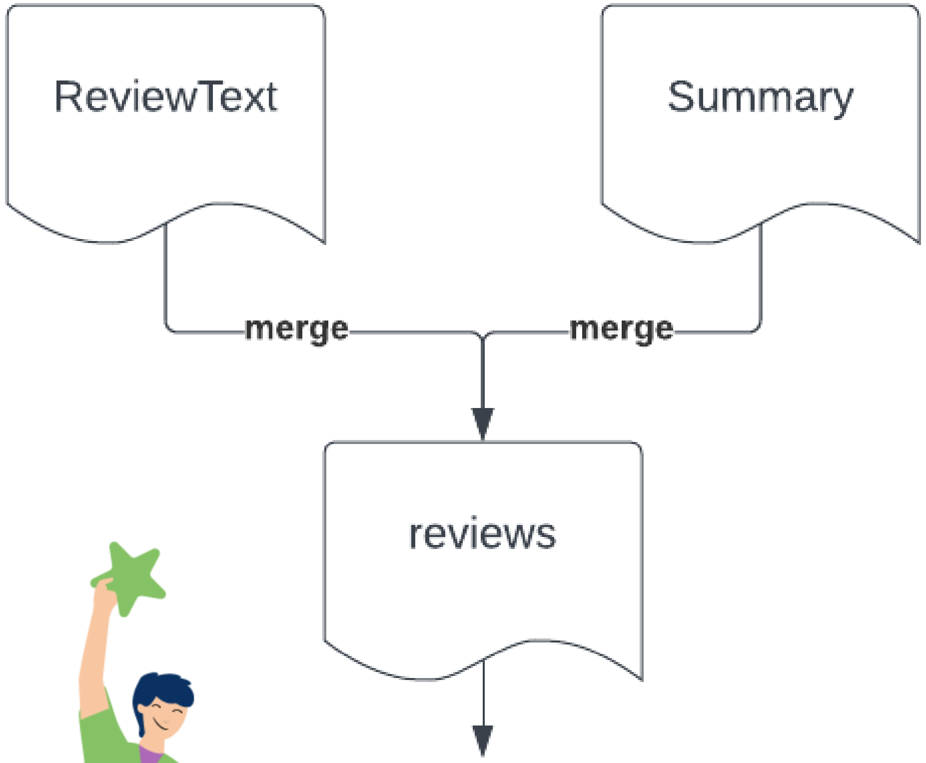


Data Prepressing *(Review Text & Summary)*

Like a narrow-down tunnel, we try to concentrate on the most significant words in the paragraph, filtering out noise and highlighting core content.



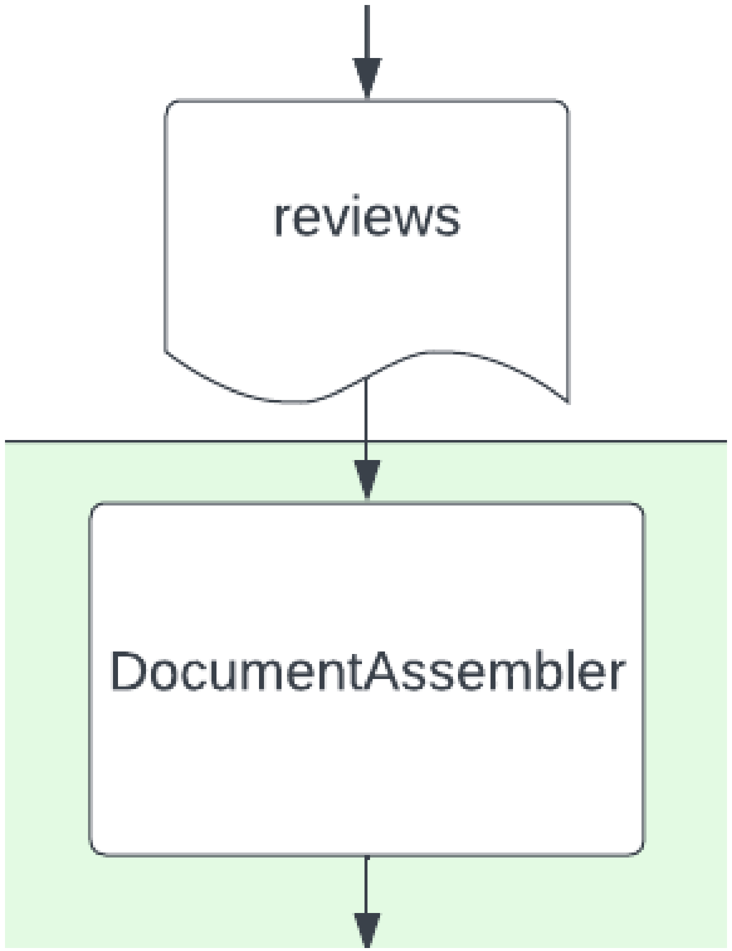
Paragraph ⟶ sentence ⟶ words ⟶ simplify words



Review Text	Summary
So much better than plastic mug types--keeps coffee warm and doesn't stain. We bought it because Cook's Country rated it tops.	Recommend

↓

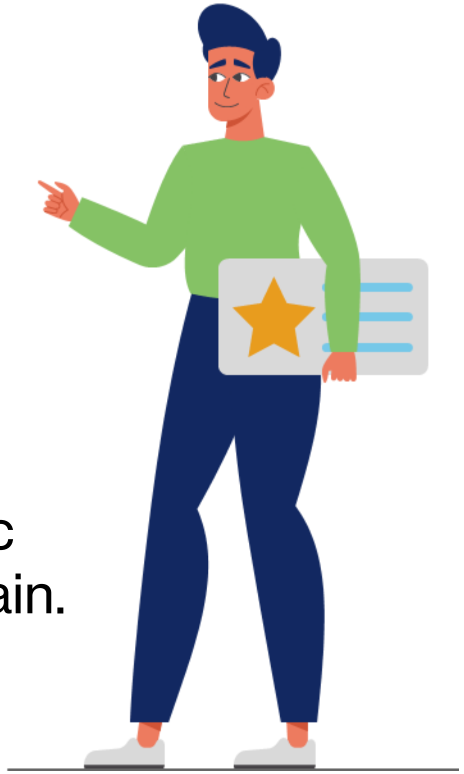
So much better than plastic mug types--keeps coffee warm and doesn't stain. We bought it because Cook's Country rated it tops. Recommend

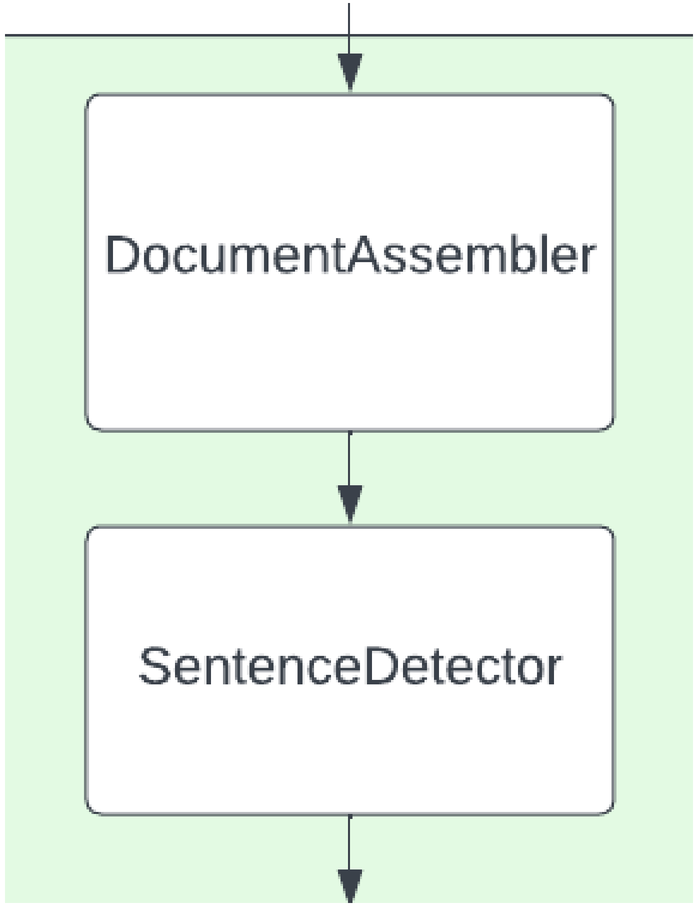


So much better than plastic mug types--keeps coffee warm and doesn't stain. We bought it because Cook's Country rated it tops. Recommend

↓

[{document, 0, 134, So much better than plastic mug types--keeps coffee warm and doesn't stain. We bought it because Cook's Country rated it tops. Recommend, {sentence -> 0}, []}]

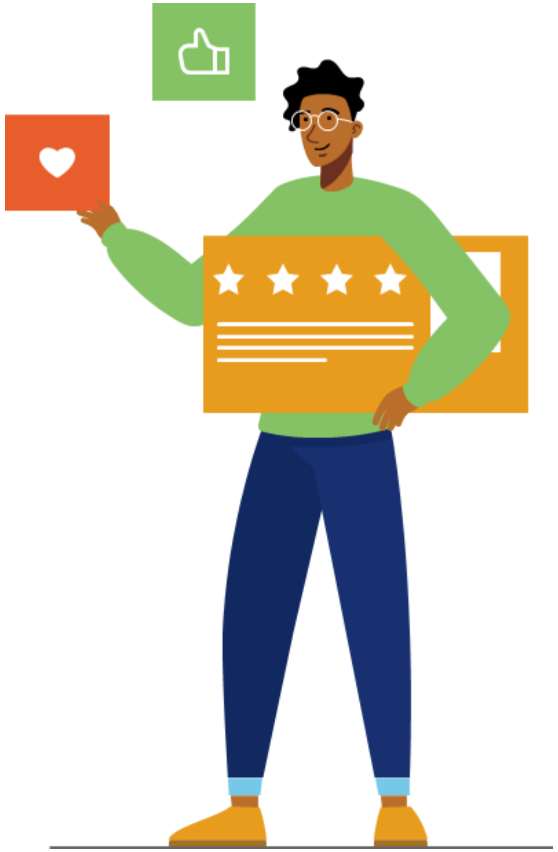


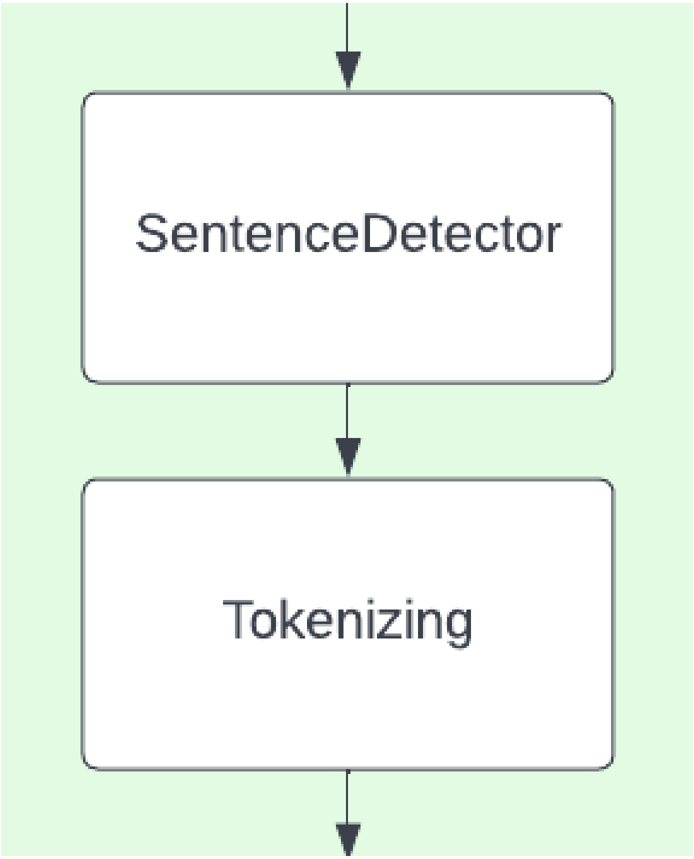


[{document, 0, 134, So much better than plastic mug types--keeps coffee warm and doesn't stain. We bought it because Cook's Country rated it tops. Recommend, {sentence -> 0}, []}]

↓

[{document, 0, 74, So much better than plastic mug types--keeps coffee warm and doesn't stain., {sentence -> 0}, []}, {document, 76, 134, We bought it because Cook's Country rated it tops. Recommend, {sentence -> 1}, []}]

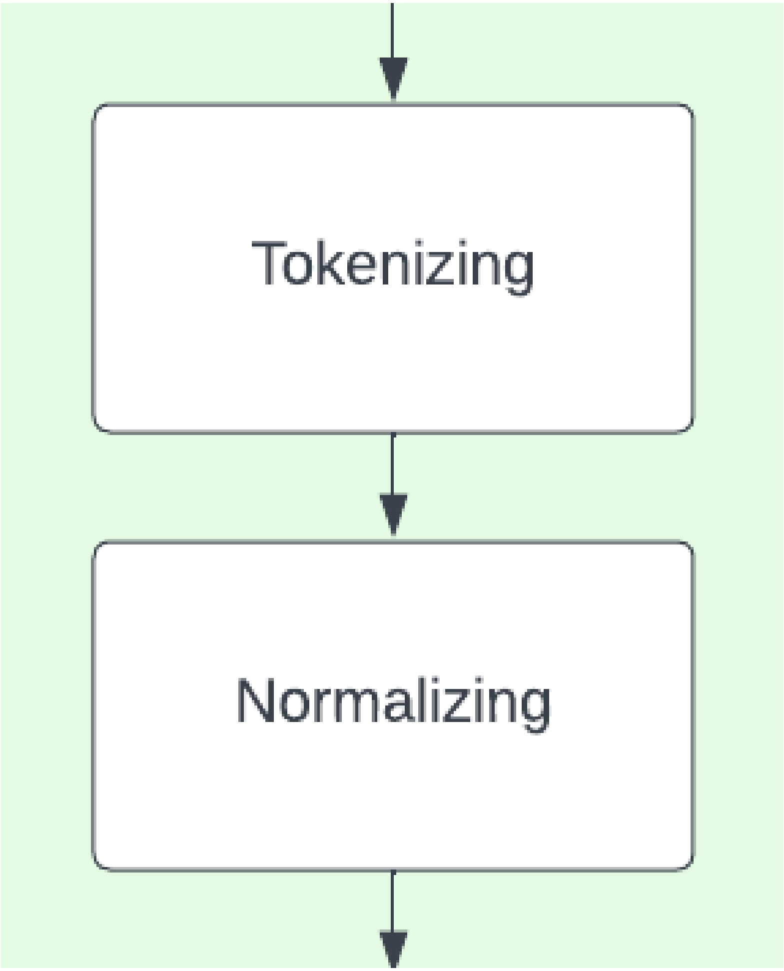




[{document, 0, 74, So much better than plastic mug types--keeps coffee warm and doesn't stain., {sentence -> 0}, []}, {document, 76, 134, We bought it because Cook's Country rated it tops. Recommend, {sentence -> 1}, []}]

↓

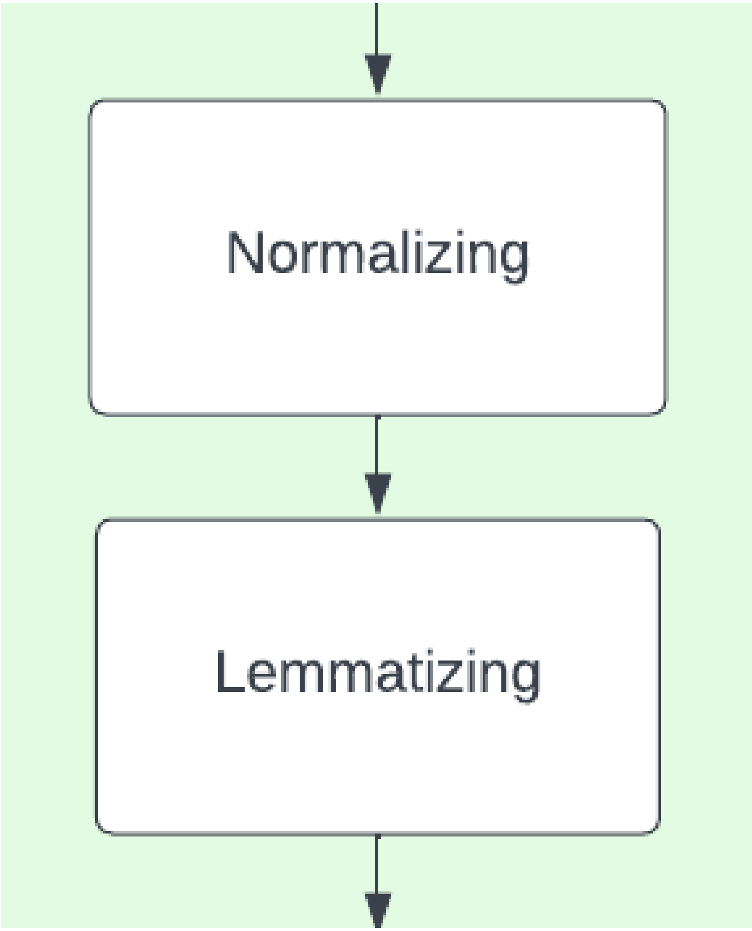
[so, much, better, than, plastic, mug, types--keeps, coffee, warm, and, doesn't, stain., we, bought, it, because, cook's, country, rated, it, tops.recommend]



[so, much, better, than, plastic, mug, types--keeps, coffee, warm, and, doesn't, stain., we, bought, it, because, cook's, country, rated, it, tops.recommend]

↓

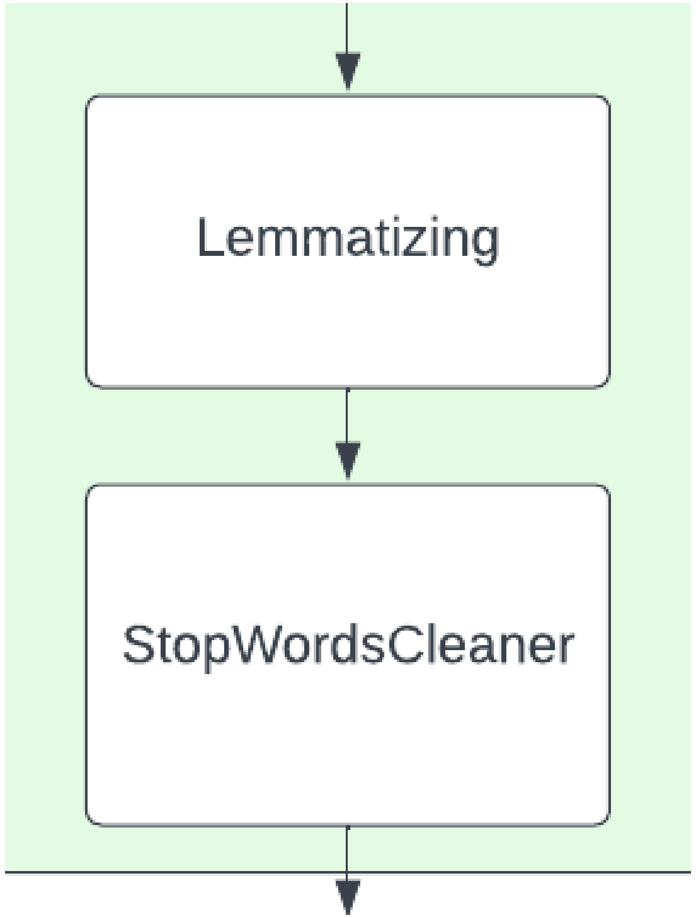
[so, much, better, than, plastic, mug, typeskeeps, coffee, warm, and, doesnt, stain, we, bought, it, because, cooks, country, rated, it, topsrecommend]



[so, much, better, than, plastic, mug, typeskeeps, coffee, warm, and, doesnt, stain, we, bought, it, because, cooks, country, rated, it, topsrecommend]

↓

[so, much, well, than, plastic, mug, typeskeeps, coffee, warm, and, doesnt, stain, we, buy, it, because, cook, country, rate, it, topsrecommend]

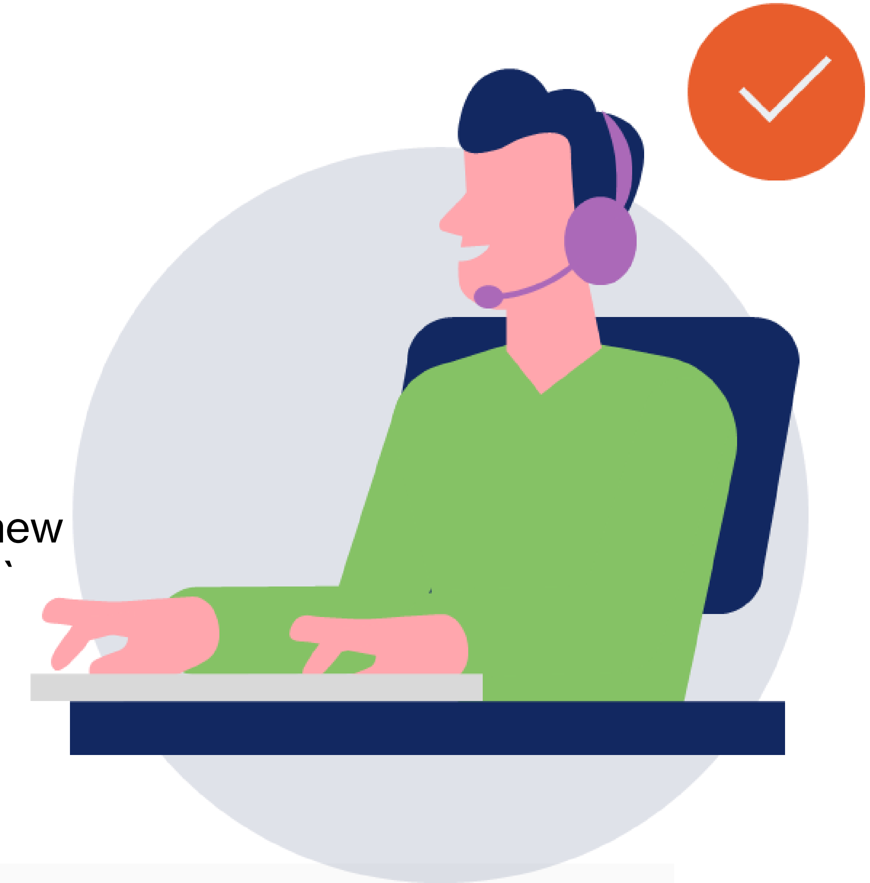


[so, much, well, than, plastic, mug, typeskeeps, coffee, warm, and, doesnt, stain, we, buy, it, because, cook, country, rate, it, topsrecommend]

↓

[much, well, plastic, mug, typeskeeps, coffee, warm, doesnt, stain, buy, cook, country, rate, topsrecommend]

Feature Engineering



1. **Calculate Days Since reviewTime:**
   Adds new column named "days" to the DataFrame. The values in this column represent the **number of days** since the date given in the `reviewTime` column to the current date.

2. **Calculate Length of reviewText:**
   Adds a new column named "len" to the DataFrame df. The values in this new column represent the **length (number of characters)** of the `reviewText` column for each row.

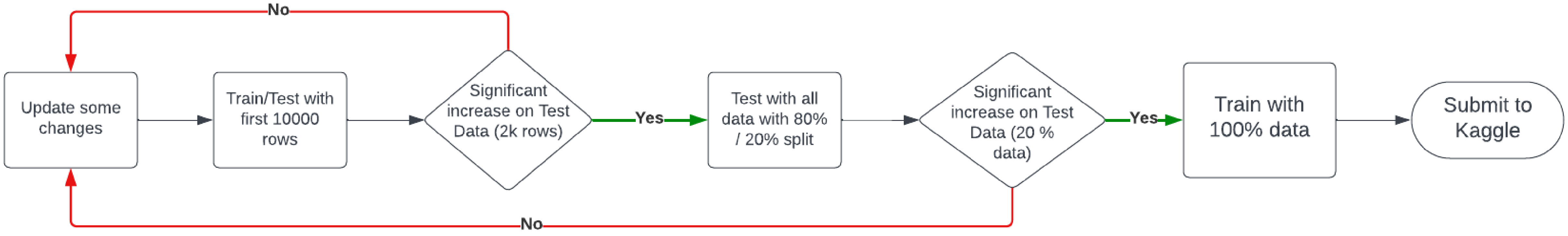
3. **Calculate TF-IDF of reviewText:**

```
# Calculating TF-IDF
review_tf = CountVectorizer(inputCol="reviewTokenFeatures", outputCol="reviewRawFeatures", vocabSize=10000, minTF=1, minDF=50, maxDF=0.
40)
review_idf = IDF(inputCol="reviewRawFeatures", outputCol="reviewIDF")
```

Model Training

```
LogisticRegression(maxIter=500, regParam = 0.01,
elasticNetParam=0)
```

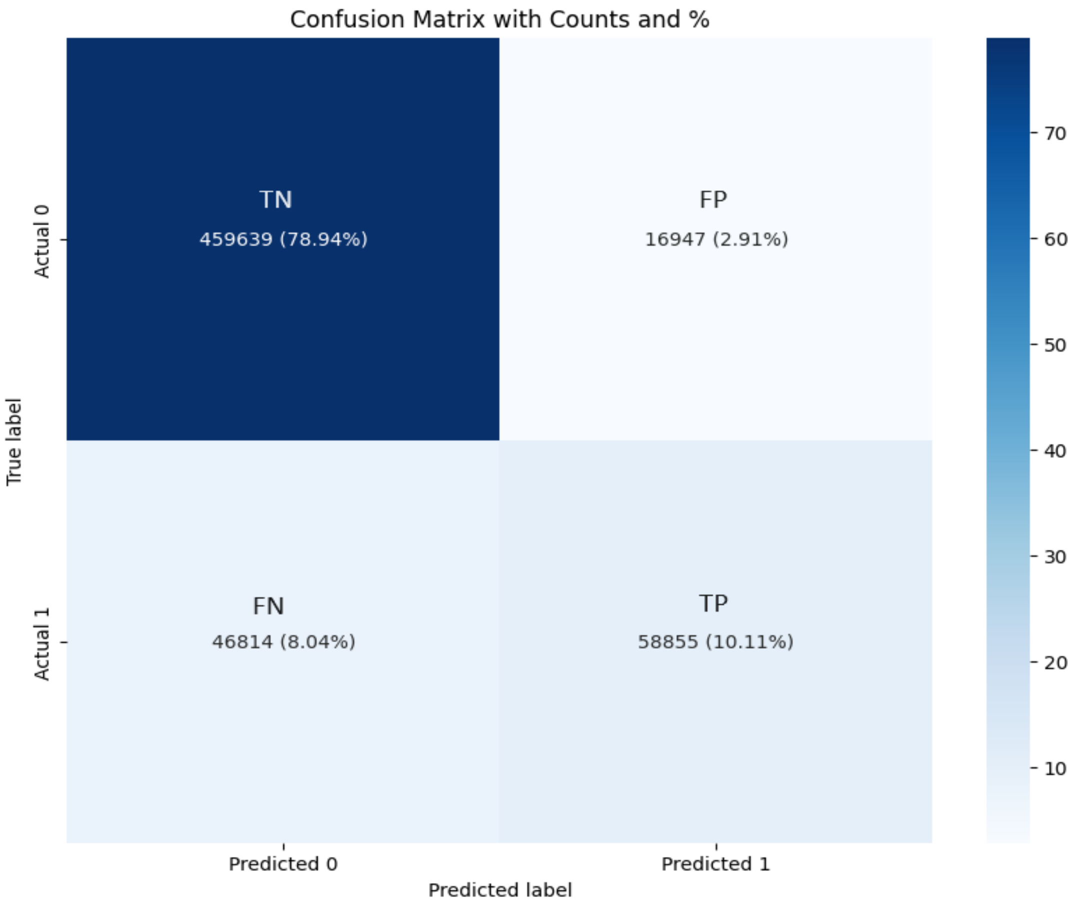
Test Pipeline for fast iteration:



Model Output

- Test AUC: 0.92307
- Kaggle AUC: 0.88905
- Precision Score: 0.7767
- Recall Score: 0.5569
- F1 Score: 0.6489

Confusion Matrix:



Improving Customer Review Analysis with LLM

Feature Engineering Insights:

- **Date Difference**: Add a feature representing the number of days since the review was written.
- **Review Length**: Calculate and incorporate the length of the review text.
- ***Source Query to LLM***: "What do people usually consider when reading a customer review? What features can determine a review's helpfulness?"

Data Cleaning Strategy:

- Implement procedures to identify and handle duplicates, missing values, and inconsistencies in the dataset.
- ***Source Query to LLM***: "Given our dataset's columns, how would you recommend cleaning the data, particularly concerning duplicates?"

Data Pre-processing for NLP:

- Techniques for transforming textual data into a format suitable for logistic regression, including tokenization, vectorization and Lemmatization.
- ***Scenario Presented to LLM***: "Assuming you're an NLP expert, how would you pre-process data for predicting a review's label using logistic regression?"

Coding Simplification:

- Aim for modular, efficient, and readable code structures.

Model Comparison

	Baseline Model	Best Model
Model	Logistic Regression	Logistic Regression
Data Pre-processing	1. Tokenized 2. Stop words removal 3. Vectorized 'review Text' (bag of words)	1. Tokenized 2. Normalized 3. Lemmatized (LLM) 4. Stop words removal 5. TF-IDF
Feature Engineering	1. Review Length	1. Review Length 2. Date Diff(LLM) *(Number of days since the review was written)*

Conclusions

1. **Length and Content of Reviews Matter**:
    - Lengthier reviews tend to be more helpful. Additionally, reviews mentioning "Book" and "Game" are common.
    - **Business Insight**: Amazon could incentivize users to write comprehensive reviews, especially for popular categories like books and games. This could be done through badges, recognition, or even discounts.
2. **Verified vs. Unverified Reviews**:
    - Unverified reviews are generally seen as more helpful.
    - **Business Insight**: Amazon might want to investigate the quality and relevance of unverified reviews. It's possible that these reviews provide unique insights or perspectives not covered by verified purchasers.
3. **Review Ratings**:
    - Lower-rated reviews are more helpful.
    - **Business Insight**: It's essential for Amazon to ensure a balanced display of positive and negative reviews. Negative reviews, when genuine, can help in building trust as they show transparency.
4. **Model's Performance and Application**:
    - The logistic regression model achieved a decent performance in predicting review helpfulness.
    - **Business Insight**: Amazon can leverage such models to automatically highlight or prioritize reviews that are likely to be deemed helpful, enhancing the user shopping experience.
5. **Continuous Improvement**:
    - The team presented further improvements through LLM on data cleaning, feature engineering, and preprocessing. Amazon should invest in continuous data science R&D to refine and improve models, ensuring they remain relevant and effective over time.

Thank you