AMAZON PRODUCT REVIEW SENTIMENT ANALYSIS

Presented by Ruthy Yao

OVERVIEW

- Summary
- Business Problem
- Data Overview
- Methodology
- Results
- Insights & Recommendations

SUMMARY

- Applied ML models to predict customers' sentiment based on the text of their reviews.
- Recommend Random Forest model (RFM) due to its superior performance in
  - Detecting the negative sentiment- of all the negative reviews, RFM can correctly predict 82% of them; RFM can minimize the error of misclassifying a negative review to 6%.
  - Identifying the most contributing words towards a particular sentiment class - identified "taste" and "freshness" are what customers care most for our products
- Implemented word vectorization using NLP models to convert the texts to numerical before loading to the ML models.

BUSINESS PROBLEM

- Large volume, unstructured data making manual analysis time-consuming and inefficient.
- The leadership team needs to quickly interpret the sentiments, to understand customers feedback about our products, adjust business strategies accordingly.
- Need to automate the process of sentiment analysis using ML and NLP techniques to gain real-time insights, improve product offerings and enhance customer satisfaction.

DATA OVERVIEW

- 233,325 pieces of review over the last 12 months.
- Imbalanced data with 75% are positive class.
- Each review includes a summary and a body
  - Decided to use the summary considering the computation cost
- Texts contain
  - stop words, such as “a”, “the”, “in”, “not”, etc.
  - words of same root with variations, such as “tasty” and “tasti”, “advertised” and “advertise”
  - punctuation: “.”, “,” ,”!” ,”/”
  - mix of upper case and lower case.

METHODOLOGY

Word embedding

- Using TD-IDF model which consider both the word frequency and the rarity of the words across the entire dataset.

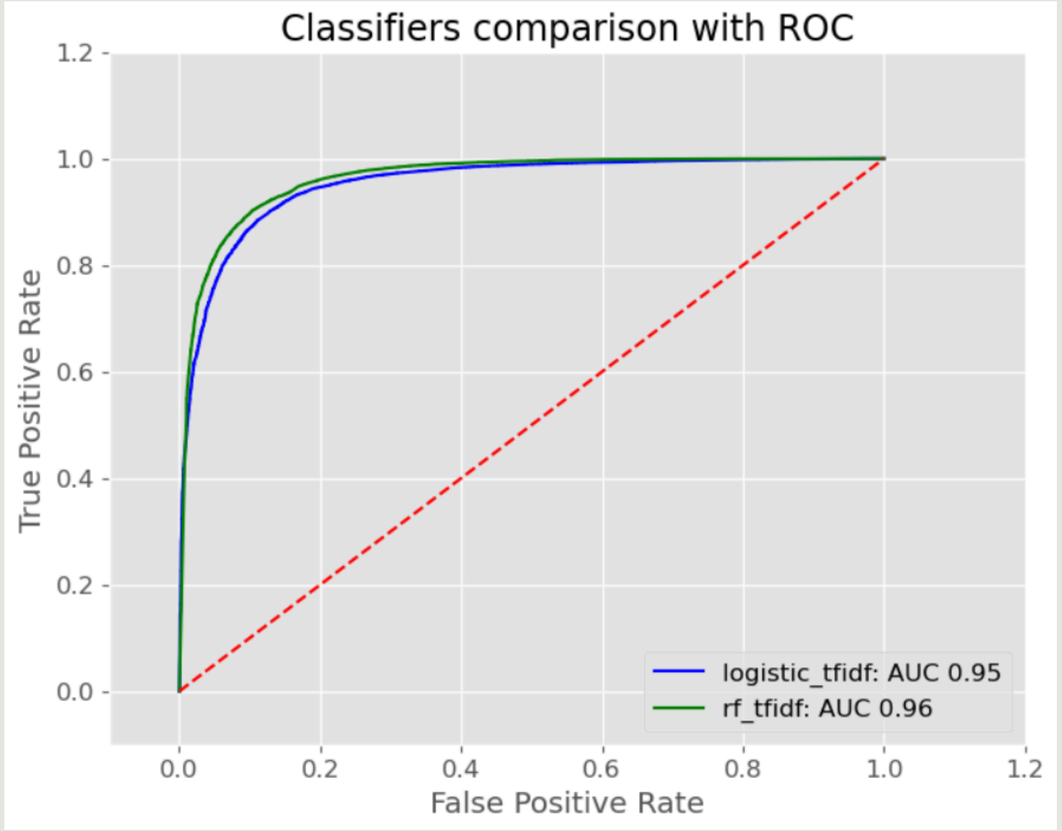
Predictive Analysis

- Applied ML model to predict the sentiment;
- Compare and evaluate the models using ROC-AUC graph and recall/precision scores.

RESULT

Random Forest Model (RFM) triumphs over logistic regression

- RFM achieves higher AUC: 0.96 vs 0.95 in logistic regression



RESULT

RFM is better at detecting the negative sentiment and minimize the misclassification error.

Recall score for negative class (The percentage of the negative reviews that the Model did predicted as negative):

- Random Forest: 0.82
- Logistic Regression: 0.74

Precision score for positive class (The percentage of all predicted positive reviews that are actually positive):

- Random Forest: 0.94
  - misclassification error: 6%
- Logistic Regression: 0.92
  - misclassification error: 8%

RESULT

RFM’s feature importance technique allows us to identify the key words that contribute most to a particular sentiment class.

- Those words such as "delicious", "tasty", "yummy", "nasty", "disgust", "yuck", "stale" appearing on top of the list indicates that taste, freshness are the area that accounts major part of the customer satisfaction and customer sentiment.

	feature	importance
14072	not	0.041359
9017	great	0.032323
2227	best	0.016128
5474	delici	0.015583
12290	love	0.013970
8653	good	0.012694
5790	disappoint	0.011953
3175	but	0.010054
6740	excel	0.009513
23322	yummi	0.009310
1796	bad	0.006497
21128	too	0.006225
1688	awesom	0.006016
14772	ok	0.005589
23302	yum	0.005305

INSIGHTS & RECOMMENDATIONS

- **Product Improvement**: Since most of the negative sentiment is around the "taste" and "freshness", it would be worth focusing on improving the taste of our products and how to better preserve the food to maintain the freshness.
- **Customer Satisfaction**: Track sentiment trends over time to see if new product launches are being well received.
- **Marketing Strategy**: Highlight positive reviews to promote top-performing products.

Thank You

- Email: zejia.yao@gmail.com
- LinkedIn: https://www.linkedin.com/in/ruthy-yao-b3258b25/
- GitHub: https://github.com/RuthyYao