Banao - AI [Task 2]- 01/01

Ankit Kumar

Documentation Report: Task 2: Multi-Modal Data Analysis and Predictive Insights

Objective:

This task aims to utilize an open-source LLM to analyze and derive insights from a dataset that combines text data (such as product reviews) with other forms of metadata (like images or numerical data). You will extract meaningful features, analyze relationships, and build a predictive model. This task will evaluate your skills in multi-modal analysis, feature engineering, and predictive analytics.

1. Methodology

1.1 Data Preprocessing

To prepare the dataset for analysis, the following preprocessing steps were applied:

- **Removal of Special Characters and URLs**: All non-alphanumeric characters and URLs were stripped from the text to clean noisy data.
- **Whitespace Normalization**: Extra spaces were removed for uniformity.
- **Lowercasing**: All text was converted to lowercase to maintain consistency.
- **Tokenization and Stopword Removal**: The text was split into tokens, and common stop words (like "the," "is," etc.) were removed using the NLTK library.
- **Lemmatization**: Words were converted to their base forms using WordNetLemmatizer, ensuring uniformity and reducing dimensionality.

OUTPUT:

jupyter step2_preprocessed_text.csv Last Checkpoint: 47 minutes ago

File Edit View Settings Help

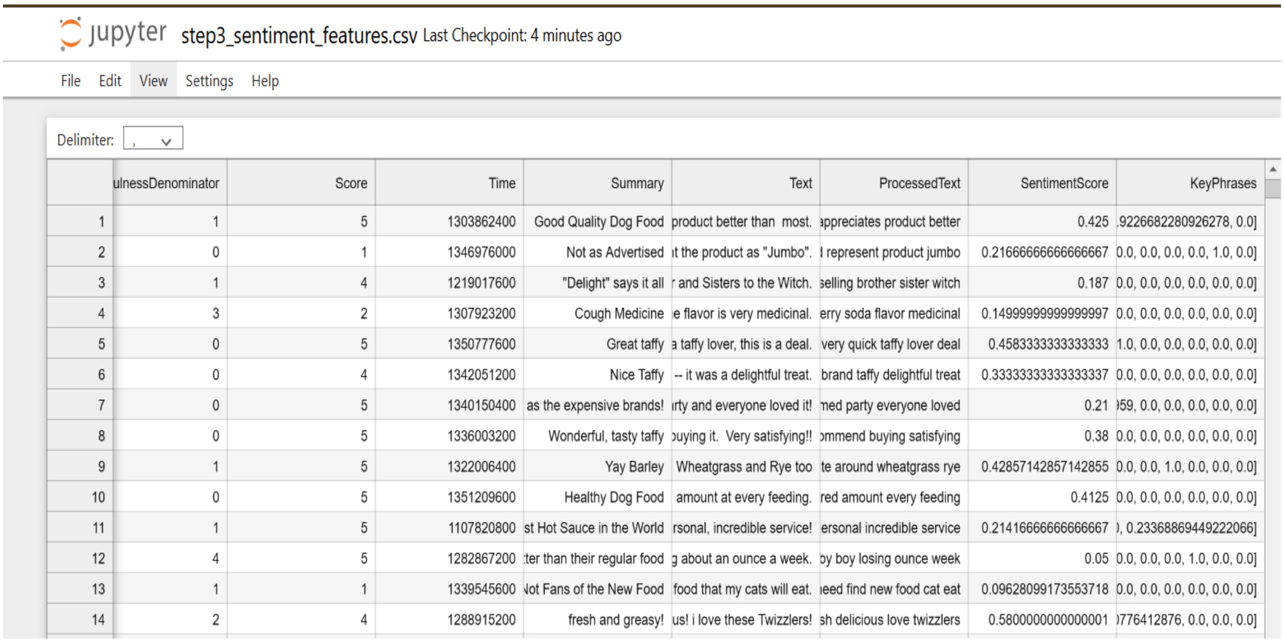
Delimiter: ,

	UserId	ProfileName	pfulnessNumerator	lpfulnessDenominator	Score	Time	Summary	Text	ProcessedText
1	AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	roduct better than most.	appreciates product better
2	ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	the product as "Jumbo".	l represent product jumbo
3	IWJIXXAIN	es "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	and Sisters to the Witch.	selling brother sister witch
4	C6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	flavor is very medicinal.	erry soda flavor medicinal
5	LF8GW1T	igham "M. Wassir"	0	0	5	1350777600	Great taffy	taffy lover, this is a deal.	very quick taffy lover deal
6	K1MGOEU	Twoapennything	0	0	4	1342051200	Nice Taffy	- it was a delightful treat.	brand taffy delightful treat
7	KFXXRU1	David C. Sullivan	0	0	5	1340150400	as the expensive brands!	ty and everyone loved it!	ned party everyone loved
8	VEQN31IQ	amela G. Williams	0	0	5	1336003200	Wonderful, tasty taffy	ying it. Very satisfying!!	mmend buying satisfying
9	9TZK0BBI	R. James	1	1	5	1322006400	Yay Barley	Wheatgrass and Rye too	te around wheatgrass rye
10	VZCCYT4	Carol A. Reed	0	0	5	1351209600	Healthy Dog Food	amount at every feeding.	red amount every feeding

1.2 Sentiment Analysis and Feature Extraction

- **Sentiment Analysis**:
  - Sentiment polarity scores were generated for each review using the TextBlob library. These scores range from −1 (negative sentiment) to +1 (positive sentiment), providing a numeric measure of customer sentiment.
- **Key Phrase Extraction**:
  - The TfidfVectorizer was used to extract the top 10 key phrases from the processed text. These phrases represent the most important terms contributing to the textual content.
- **Topic Modeling**:
  - A Latent Dirichlet Allocation (LDA) model was applied to uncover hidden themes within the reviews. The top three topics for each review were added as numerical features, representing the distribution of topics within the text.

OUTPUT:



	ulnessDenominator	Score	Time	Summary	Text	ProcessedText	SentimentScore	KeyPhrases
1	1	5	1303862400	Good Quality Dog Food	product better than most.	appreciates product better	0.425	.9226682280926278, 0.0]
2	0	1	1346976000	Not as Advertised	it the product as "Jumbo".	l represent product jumbo	0.21666666666666667	0.0, 0.0, 0.0, 0.0, 1.0, 0.0]
3	1	4	1219017600	"Delight" says it all	r and Sisters to the Witch.	selling brother sister witch	0.187	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
4	3	2	1307923200	Cough Medicine	e flavor is very medicinal.	erry soda flavor medicinal	0.14999999999999997	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
5	0	5	1350777600	Great taffy	a taffy lover, this is a deal.	very quick taffy lover deal	0.45833333333333333	1.0, 0.0, 0.0, 0.0, 0.0, 0.0]
6	0	4	1342051200	Nice Taffy	-- it was a delightful treat.	brand taffy delightful treat	0.33333333333333337	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
7	0	5	1340150400	as the expensive brands!	rty and everyone loved it!	ned party everyone loved	0.21	959, 0.0, 0.0, 0.0, 0.0, 0.0]
8	0	5	1336003200	Wonderful, tasty taffy	buying it. Very satisfying!!	ommend buying satisfying	0.38	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
9	1	5	1322006400	Yay Barley	Wheatgrass and Rye too	te around wheatgrass rye	0.42857142857142855	0.0, 0.0, 1.0, 0.0, 0.0, 0.0]
10	0	5	1351209600	Healthy Dog Food	amount at every feeding.	red amount every feeding	0.4125	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
11	1	5	1107820800	st Hot Sauce in the World	rsonal, incredible service!	ersonal incredible service	0.21416666666666667	), 0.23368869449222066]
12	4	5	1282867200	ter than their regular food	g about an ounce a week.	by boy losing ounce week	0.05	0.0, 0.0, 0.0, 1.0, 0.0, 0.0]
13	1	1	1339545600	Not Fans of the New Food	food that my cats will eat.	need find new food cat eat	0.09628099173553718	0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
14	2	4	1288915200	fresh and greasy!	us! i love these Twizzlers!	sh delicious love twizzlers	0.5800000000000001	0776412876, 0.0, 0.0, 0.0]

1.3 Feature Engineering

- **Feature Integration**:
  - Text-based features (e.g., sentiment scores, key phrases, and topic distributions) were combined with numerical metadata, such as HelpfulnessNumerator, HelpfulnessDenominator, and Score.
- **Resulting Features**:
  - A comprehensive feature set was created, including TF-IDF scores, sentiment polarity, topic distributions, and helpfulness metrics, to support predictive modeling.

jupyter step4_feature_engineering.csv Last Checkpoint: 35 minutes ago

File Edit View Settings Help

Delimiter: ,

	TF-IDF_br	TF-IDF_coffee	TF-IDF_flavor	TF-IDF_good	TF-IDF_great	TF-IDF_like	TF-IDF_love	TF-IDF
1	0.0	0.0	0.0	0.28251160878973214	0.0	0.26243195645242534	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.7353328908446626	0.6777060864726179	0.0	0.0	0.0	
5	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
6	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
7	0.0	0.0	0.7170520545084224	0.0	0.6970196203301959	0.0	0.0	
8	0.0	0.0	0.7353328908446626	0.6777060864726179	0.0	0.0	0.0	
9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
10	0.0	0.0	0.0	1.0	0.0	0.0	0.0	

jupyter step4_feature_engineering.csv Last Checkpoint: 35 minutes ago

File Edit View Settings Help

Delimiter: ,

	TF-IDF_product	TF-IDF_taste	SentimentScore	HelpfulnessNumerator	HelpfulnessDenominator	Score	Topic_0	Topic_1
1	226682280926278	0.0	0.425	1	1	5	0.6598165854420358	0.058913402583574206
2	1.0	0.0	0.21666666666666667	0	0	1	0.7777368475649299	0.1111132988241486
3	0.0	0.0	0.187	1	1	4	0.3333333333333333	0.3333333333333333
4	0.0	0.0	0.14999999999999997	3	3	2	0.1120876704244595	0.5763371570074214
5	0.0	0.0	0.4583333333333333	0	0	5	0.7776164258597197	0.11121495137718275
6	0.0	0.0	0.33333333333333337	0	0	4	0.16666926629074819	0.6666487248271558
7	0.0	0.0	0.21	0	0	5	0.442201056784265	0.44661679542557714
8	0.0	0.0	0.38	0	0	5	0.1120876704244595	0.5763371570074214
9	0.0	0.0	0.42857142857142855	1	1	5	0.6620907913669324	0.16881932135327202
10	0.0	0.0	0.4125	0	0	5	0.11198828705554942	0.11586215015916469

2. Insights Derived from Predictive Modeling

2.1 Model Training and Evaluation

- **Model Used**: A Gradient Boosting Regressor was selected for its ability to handle complex, non-linear relationships between features.
- **Performance Metrics**:
  - **Mean Squared Error (MSE)**: This metric quantified the average squared difference between predicted and actual scores.
  - **R-squared (R^2)**: This score measured the proportion of variance in the dependent variable explained by the model.

OUTPUT:

jupyter step5_predictions.csv Last Checkpoint: 35 minutes ago

File Edit View Settings Help

Delimiter: ,

	SentimentScore	HelpfulnessNumerator	HelpfulnessDenominator	Topic_0	Topic_1	Topic_2	Actual	Predicted
1	0.3875	0	0	0.11154412206140754	0.1139977555859301	0.774458122352662 4	5	4.448482838638957
2	148148148148148	0	0	0.530622269033933	0.3941903337506183	0.07518739721544879	5	4.892292744630752
3	812500000000001	0	2	0.16667184741021268	0.16860299701816459	0.6647251555716228	3	2.7803206337414244
4	617424242424242	0	1	0.1893844901343463	0.25852736326246184	0.5520881466031918	2	3.569058943052258
5	350340136054418	0	2	0.3336544239607461	0.03417382261672286	0.6321717534225311	5	3.369842498335606
6	0.0	3	3	0.11111275012938755	0.777766416364047	0.1111208335065654	4	3.615198665279468
7	466666666666666	2	2	0.7748995636698852	0.11246481815475416	0.11263561817536064	5	4.928278807019383
8	.144520757020757	0	0	0.3333333333333333	0.3333333333333333	0.3333333333333333	5	4.298549610889591
9	516666666666665	0	1	0.03740106692064949	0.6519313300849071	0.3106676029944435	4	3.939768316041185
10	142857142857142	1	1	0.7763580873721653	0.11177125281771283	0.11187065981012195	5	4.07402867114899

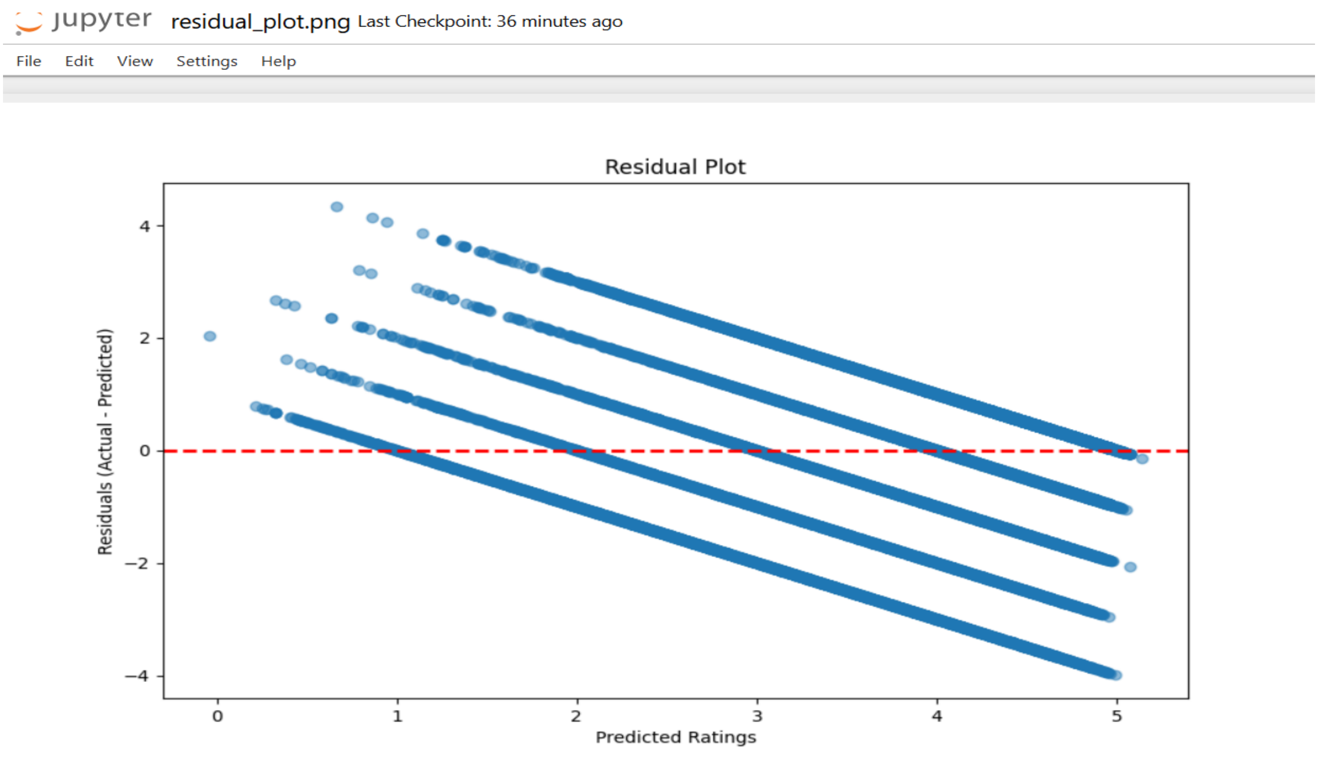
jupyter model_performance.txt Last Checkpoint: 30 minutes ago

File Edit View Settings Help

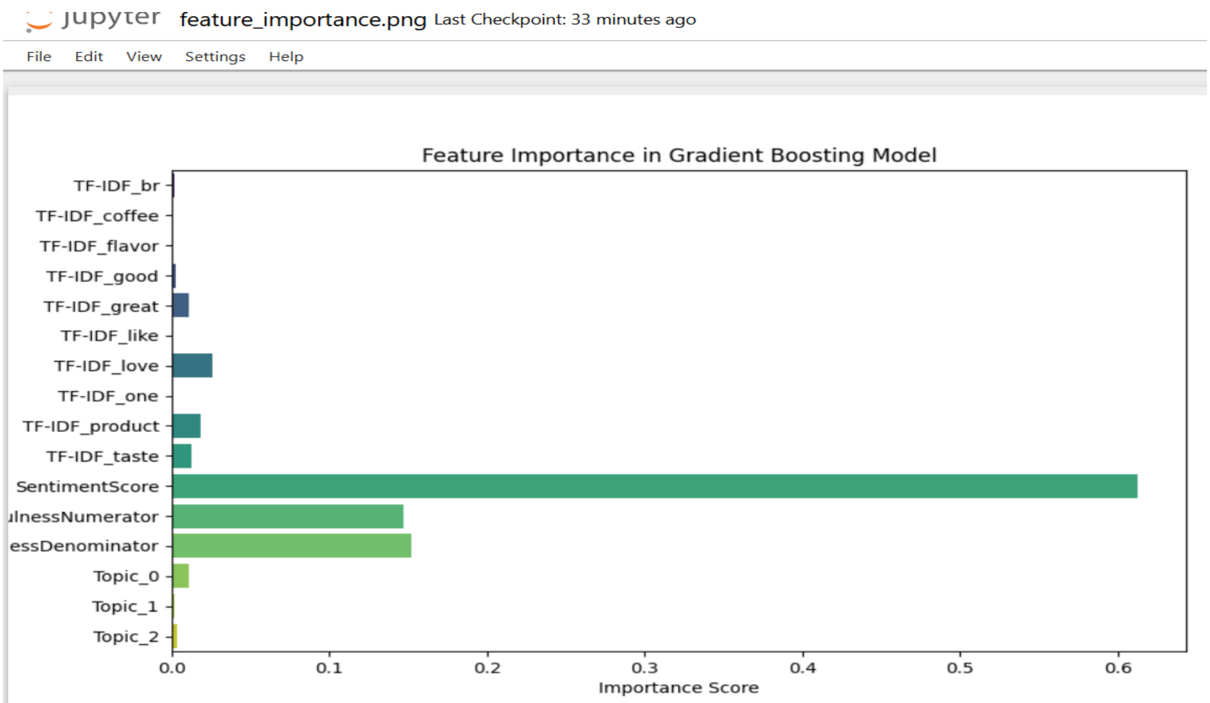
```
Mean Squared Error: 1.128465315247665
R-squared: 0.33727144219110805
```

2.2 Key Observations

- **Feature Importance**:
    - Sentiment scores and topic distributions emerged as significant predictors of product ratings, highlighting the importance of textual data in understanding customer opinions.
    - Numerical metadata (e.g., helpfulness metrics) also played a critical role in influencing predictions.

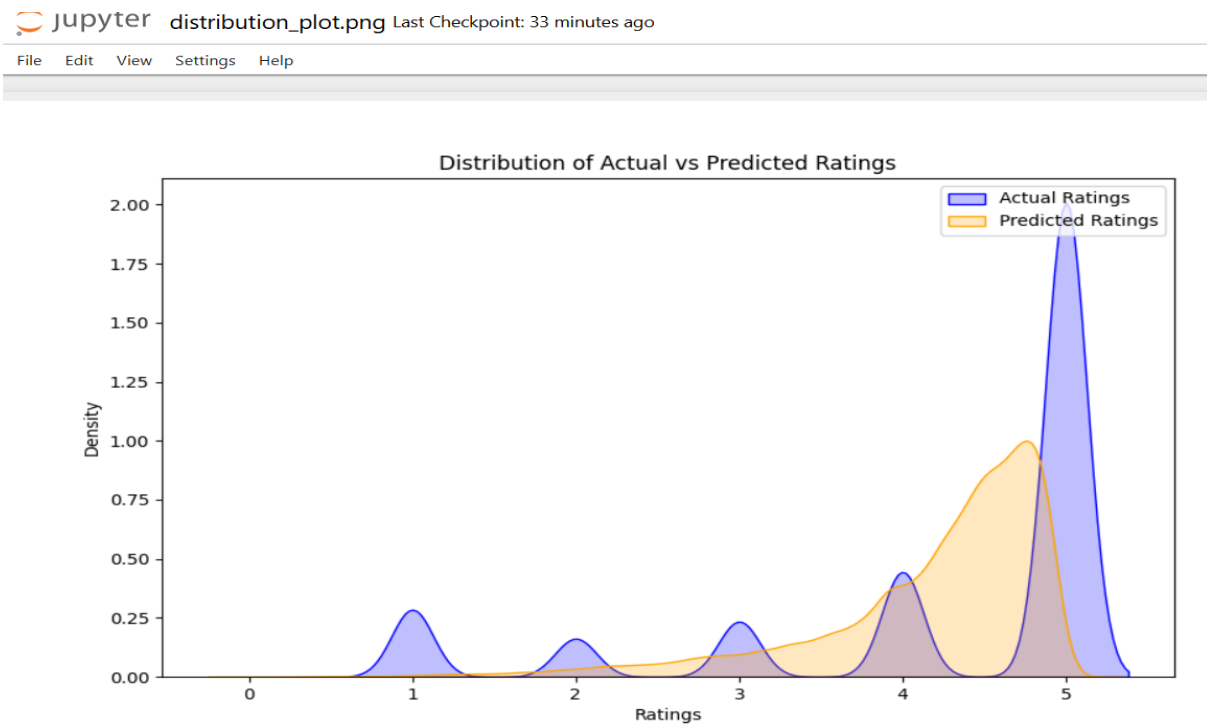


- **Trends in Data**:
    - Products with highly positive sentiment scores were strongly correlated with higher ratings.
    - Topic modeling revealed recurring themes (e.g., product quality, delivery experience) that align with customer satisfaction levels.



2.3 Implications for Product Strategies

- Products with consistently negative sentiment and low helpfulness scores should be prioritized for improvement.
- Insights from key phrases and topics can inform targeted marketing strategies and product enhancements.
- Trend analysis over time can identify seasonal patterns in customer feedback, aiding inventory and promotional planning.



3. Reflections on Multi-Modal Analysis

Effectiveness of Text and Metadata Integration

- **Enhanced Insights**:
  - Combining textual sentiment, themes, and numerical metadata enabled a more holistic understanding of customer behavior and preferences.
  - The integration of multi-modal features enhanced the model's predictive performance, as reflected in the R^2 score.
- **Challenges**:
  - Handling high-dimensional text data required careful selection of key features to prevent overfitting.
  - The lack of image data limited the scope of multi-modal analysis, though text and metadata alone proved highly informative.

Future Improvements

- Incorporating temporal data (e.g., review timestamps) could uncover valuable trends.
- Exploring advanced models like Transformer-based architectures may further improve text understanding and prediction accuracy.
- Adding product categories could provide deeper insights into cross-category trends and patterns.