JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Classifying Product Feedback: Detecting Product Categories and Sentiment

Shiv Palit

EN.605.744

Problem Statement

- Businesses generate massive volumes of unstructured customer feedback daily.
- Transforming this data into actionable insights is essential for enhancing products and boosting customer satisfaction.
- Classifying feedback by product categories presents challenges due to the diversity of language and expression styles.
- Accurate sentiment analysis requires capturing contextual meaning and subtle linguistic nuances.
- **Objective:** Develop a robust machine learning system to classify feedback and accurately predict sentiment.

Dataset

- **Amazon Reviews Dataset (2023)**
  - Comprehensive dataset containing product reviews from Amazon customers.
  - Review title, review body, review rating (1-5), product category.
- **Data Sampling & Preparation**
  - Reviews were sampled from 10 product categories:
    - All Beauty
    - Amazon Fashion
    - Appliances
    - Baby Products
    - Grocery and Gourmet Food
    - Arts, Crafts, and Sewing
    - Office Products
    - Pet Supplies
    - Industrial and Scientific
    - Tools and Home Improvement
  - 100 reviews per rating (1-5) were selected for each category, totaling 5,000 rows of feedback.
  - A new field was created, labeling ratings 1-2 as negative and 4-5 as positive for training the sentiment classifier.

Related Literature

- **Lahiri and Mihalcea (2013)**: *Native Language Identification: A Simple n-gram Based Approach*
  - Purpose: Explored character, word, and POS tag n-grams with TF-IDF for native language identification.
  - Relevance: Highlighted the synergistic use of POS tagging, TF-IDF, and n-grams for improving classification performance.
- **Petersen et al. (2022)**: *Differentiable Top-k Classification Learning*
  - Purpose: Proposed a differentiable approach to optimize models for multiple top-K values.
  - Relevance: Highlighted the importance of top-k accuracy as a metric to assess a classifier's performance.
- **Piskorski and Jacquet (2020)**: *TF-IDF Character N-grams vs. Embedding Models for Event Classification*
  - Purpose: Compared TF-IDF-weighted character n-grams with embedding models for classification.
  - Relevance: Validated TF-IDF and n-grams for nuanced text representation.
- **Sun and Lu (2020)**: *Understanding Attention for Text Classification*
  - Purpose: Investigated attention mechanisms for interpretability in text classification tasks.
  - Relevance: Inspired use of learnable attention weights in Neural Networks.

Approach

- **Preprocessing**
  - Text normalization using SpaCy (stopword removal, punctuation filtering, POS tagging).
  - Feature engineering with TF-IDF and n-grams (1–4 grams).
- **Feature Selection**
  - Chi-squared statistics applied to select discriminative features above a defined threshold.
- **Modeling**
  - SVM models optimized with grid search (POS and Non-POS features).
  - Neural networks with learnable attention weights for feature importance.
- **Evaluation**
  - Metrics: Accuracy, F1-score, and top-K accuracy for classification.
  - Experimentation with POS tagging and feature selection methods to optimize results.

Results – Sentiment SVM Base Model

	SVM Base Non POS			SVM Base POS			
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
neg	0.78	0.76	0.77	0.75	0.73	0.74	400
pos	0.77	0.79	0.78	0.73	0.75	0.74	400
accuracy			0.78			0.74	800
macro avg	0.78	0.77	0.77	0.74	0.74	0.74	800
weighted avg	0.78	0.78	0.77	0.74	0.74	0.74	800

Results – Sentiment Chi^2 Features SVM Optimized Model

	SVM Chi Feats Non POS			SVM Chi Feats Non POS			
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
neg	0.79	0.84	0.82	0.8	0.83	0.82	400
pos	0.83	0.78	0.81	0.83	0.79	0.81	400
accuracy			0.81			0.81	800
macro avg	0.81	0.81	0.81	0.81	0.81	0.81	800
weighted avg	0.81	0.81	0.81	0.81	0.81	0.81	800

Results – Sentiment Chi^2 Features NN Model

	NN Chi Feats Non POS			NN Chi Feats POS			
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
neg	0.81	0.87	0.84	0.82	0.86	0.84	400
pos	0.86	0.8	0.83	0.85	0.82	0.83	400
accuracy			0.84			0.84	800
macro avg	0.84	0.84	0.84	0.84	0.84	0.84	800
weighted avg	0.84	0.84	0.84	0.84	0.84	0.84	800

Results – Category SVM Base Model

	SVM Base Non POS			SVM Base POS			
Category	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
All_Beauty	0.29	0.26	0.27	0.35	0.37	0.36	100
Amazon_Fashion	0.59	0.57	0.58	0.64	0.58	0.61	100
Appliances	0.61	0.56	0.58	0.29	0.58	0.39	100
Arts_Crafts_and_Sewing	0.22	0.41	0.29	0.33	0.33	0.33	100
Baby_Products	0.4	0.47	0.43	0.47	0.44	0.45	100
Grocery_and_Gourmet_Food	0.74	0.6	0.66	0.7	0.55	0.61	100
Industrial_and_Scientific	0.23	0.22	0.22	0.2	0.18	0.19	100
Office_Products	0.49	0.44	0.47	0.48	0.4	0.43	100
Pet_Supplies	0.64	0.49	0.55	0.6	0.48	0.53	100
Tools_and_Home_Improvement	0.37	0.28	0.32	0.3	0.24	0.27	100
accuracy			0.43			0.41	1000
macro avg	0.46	0.43	0.44	0.44	0.41	0.42	1000
weighted avg	0.46	0.43	0.44	0.44	0.41	0.42	1000

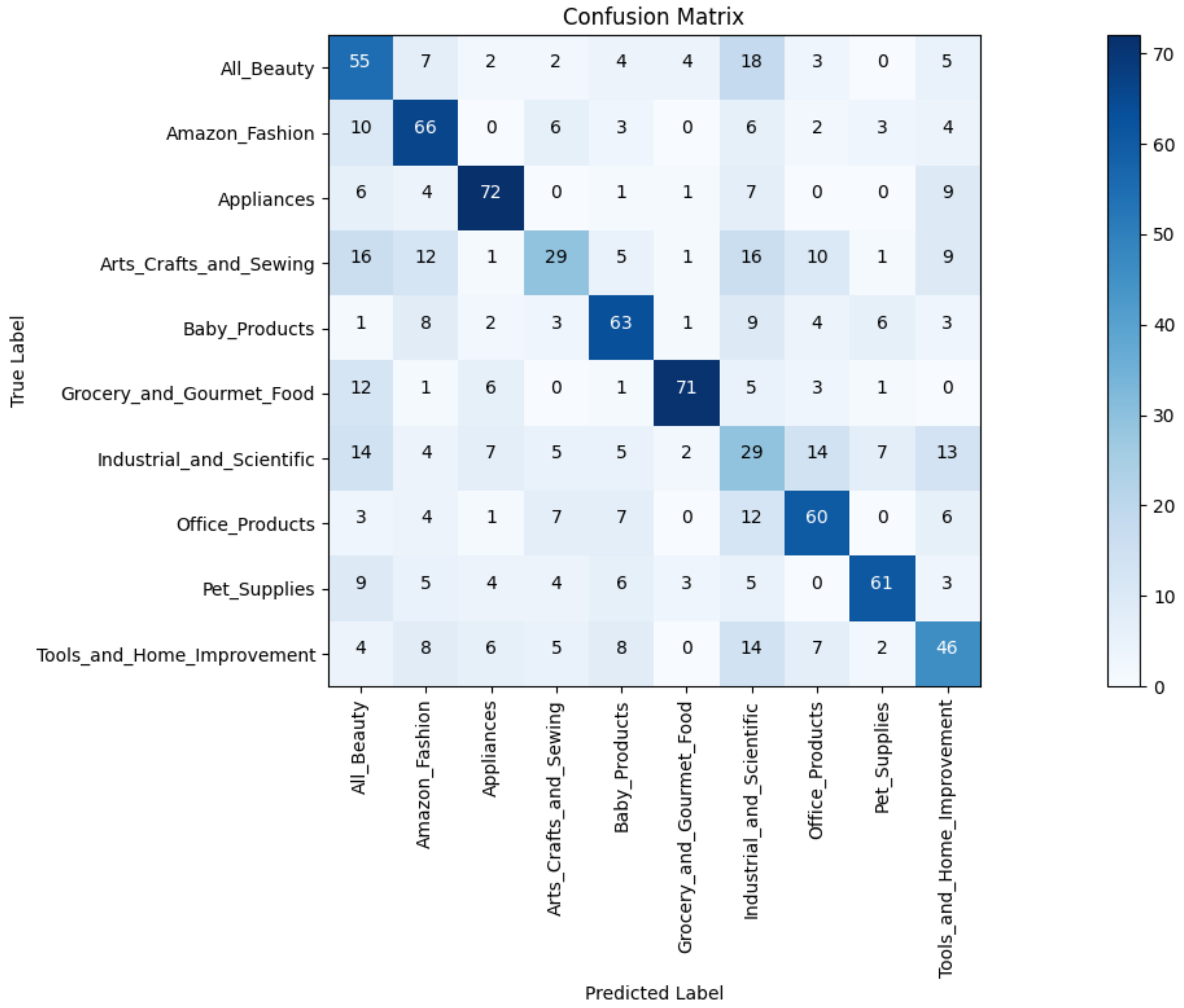
Results – Category Chi^2 Features SVM Optimized Model

	SVM Chi Feats Non POS			SVM Chi Feats POS			
Category	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
All_Beauty	0.49	0.5	0.49	0.51	0.44	0.47	100
Amazon_Fashion	0.66	0.65	0.65	0.69	0.63	0.66	100
Appliances	0.51	0.63	0.57	0.54	0.61	0.57	100
Arts_Crafts_and_Sewing	0.31	0.56	0.4	0.26	0.55	0.35	100
Baby_Products	0.6	0.55	0.58	0.69	0.58	0.63	100
Grocery_and_Gourmet_Food	0.86	0.62	0.72	0.88	0.68	0.77	100
Industrial_and_Scientific	0.27	0.34	0.3	0.23	0.32	0.27	100
Office_Products	0.68	0.47	0.56	0.69	0.46	0.55	100
Pet_Supplies	0.86	0.54	0.66	0.82	0.51	0.63	100
Tools_and_Home_Improvement	0.47	0.34	0.4	0.52	0.34	0.41	100
accuracy			0.52			0.51	1000
macro avg	0.57	0.52	0.53	0.58	0.51	0.53	1000
weighted avg	0.57	0.52	0.53	0.58	0.51	0.53	1000

Results – Category Chi^2 Features NN Model

	NN Chi Feats Non POS			NN Chi Feats POS			
Category	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Support**
All_Beauty	0.45	0.58	0.51	0.49	0.5	0.49	100
Amazon_Fashion	0.56	0.62	0.59	0.69	0.64	0.66	100
Appliances	0.71	0.7	0.7	0.72	0.69	0.7	100
Arts_Crafts_and_Sewing	0.36	0.35	0.36	0.29	0.37	0.32	100
Baby_Products	0.62	0.61	0.61	0.61	0.63	0.62	100
Grocery_and_Gourmet_Food	0.81	0.74	0.77	0.86	0.76	0.81	100
Industrial_and_Scientific	0.32	0.3	0.31	0.17	0.22	0.19	100
Office_Products	0.57	0.58	0.57	0.63	0.53	0.58	100
Pet_Supplies	0.77	0.61	0.68	0.78	0.61	0.69	100
Tools_and_Home_Improvement	0.45	0.45	0.45	0.44	0.42	0.43	100
accuracy			0.55			0.54	1000
macro avg	0.56	0.55	0.56	0.57	0.54	0.55	1000
weighted avg	0.56	0.55	0.56	0.57	0.54	0.55	1000

Confusion Matrix - Category Chi² Features Non POS NN Model



Top K Scoring Methodology

- Objective:
  - Evaluate model performance by rewarding correct predictions within the top K ranked classes, rather than requiring the top prediction to match the true label
- Approach:
  - For each prediction, rank the classes by their scores (e.g., logits or probabilities).
  - Check if the true label is among the top K classes.
  - Reward the model for a correct classification if the true label appears in the top K.
- Advantages:
  - Accounts for the complexity of closely related classes.
  - Provides a more forgiving metric for multi-class classification tasks.

Results – Category Chi^2 Features NN Model (Top K=3)

	NN Chi Feats Non POS (True label in Top K=3)			
Category	**Precision**	**Recall**	**F1-Score**	**Support**
All_Beauty	0.63	0.81	0.71	100
Amazon_Fashion	0.72	0.8	0.76	100
Appliances	0.86	0.79	0.82	100
Arts_Crafts_and_Sewing	0.8	0.73	0.76	100
Baby_Products	0.77	0.75	0.76	100
Grocery_and_Gourmet_Food	0.9	0.84	0.87	100
Industrial_and_Scientific	0.62	0.72	0.67	100
Office_Products	0.78	0.74	0.76	100
Pet_Supplies	0.84	0.66	0.74	100
Tools_and_Home_Improvement	0.73	0.71	0.72	100
accuracy			0.76	1000
macro avg	0.76	0.76	0.76	1000
weighted avg	0.76	0.76	0.76	1000

Confusion Matrix - Category Chi^2 Features NN Model (Top K=3)