Classifying Product Feedback: Detecting Product Categories and Sentiment

Abstract

This project examines the classification of customer feedback by product category and sentiment using machine learning techniques. Two feature sets were generated using Amazon product reviews (Hou et al., 2024) from 10 product categories: one with part-of-speech (POS) tagging and another without. Features were generated using TF-IDF and n-grams, while selection was conducted using Chi-squared analysis to identify the most discriminative features. Classification models included Support Vector Machines (SVMs) and feed-forward neural networks (NNs), with NNs incorporating attention mechanisms for dynamic feature importance. Evaluation metrics, such as accuracy and F1-score, demonstrated that NNs generally outperformed SVMs. For category classification, top-K accuracy was used as an additional analysis metric to assess the performance of the best-performing NN, highlighting its ability to manage closely related classes. The study found limited impact of POS tagging on performance and proposes exploring top-K cross-entropy loss to further improve multiclass classification.

Introduction

Customer feedback is vast and unstructured, encompassing diverse formats such as open-text survey responses, social media comments, emails, and live interactions. Unlike structured data, organized into predefined categories, unstructured feedback comes in many forms and lacks a consistent format, making it challenging to analyze without specialized methods like natural language processing (NLP). Structuring and analyzing this data is ultimately an NLP problem, as it involves extracting patterns, classifying sentiments, and identifying themes within free-form text.

This project investigates techniques for structuring customer feedback, focusing on two key NLP tasks: sentiment detection and categorical classification. Sentiment detection seeks to determine whether feedback expresses positive or negative emotions, while categorical classification aims to assign feedback to specific themes or product categories. Using the Amazon 2023 Reviews dataset, this project classifies sentiment and product categories as a proxy for organizing unstructured feedback. The data is represented using established methods like TF-IDF and n-grams, which help highlight the significance and context of words. POS tagging is also tested to see whether understanding the grammatical roles of words improves classification accuracy. POS tagging is hypothesized to serve as a valuable feature for training data by providing grammatical context that distinguishes between word functions (e.g., verbs and nouns) and captures subtle linguistic relationships.

The project experiments with SVMs and NNs as classification models to evaluate these representations. NNs are hypothesized to outperform SVMs due to their ability to learn complex, nonlinear patterns and dynamically prioritize features through mechanisms such as attention. By comparing these approaches, this study aims to identify which techniques better address the challenges of sentiment and categorical detection. While this project uses product categories as a target, the broader objective is to explore techniques that can generalize to identify customer pain points, enabling businesses to deploy systems for organizing and understanding unstructured feedback effectively.

Related Work

Several key studies provide important insights relevant to this project. Gyawali et al. (2013) investigated the effectiveness of using POS tag n-grams in conjunction with character and word n-grams to identify an author's native language. Their research demonstrated that combining these features allows models to capture subtle linguistic patterns, significantly improving performance. This is particularly applicable to the classification of customer feedback, which often contains a variety of linguistic expressions. The findings suggest that including POS tag n-grams in classification tasks could help account for stylistic and grammatical variations, improving model performance.

Piskorski and Jacquet (2020) compared a TF-IDF-weighted character n-gram trained SVM to SVMs trained on embeddings such as GloVe, BERT, and FastText. Their findings showed that the TF-IDF model outperformed the embedding-based models, achieving macro and micro F1 scores of 83.5% and 92.4%, respectively. The study also observed that the TF-IDF model's performance steadily improved as more training data was utilized, demonstrating its ability to effectively leverage larger datasets. These findings underscore the continued relevance of traditional feature engineering techniques, which, in certain cases, can perform as well as or better than more complex embedding methods. This supports using TF-IDF and n-grams in this project to represent customer feedback.

Sun and Lu (2020) investigated attention mechanisms in neural networks for text classification, analyzing the gradient update process during training to understand how attention weights correlate with linguistic phenomena such as sentence structure and semantic focus. They introduced two key metrics in their study: polarity scores and attention scores. Polarity scores measure how strongly a word is associated with specific class labels, while attention scores evaluate the overall importance of a token within the dataset. They found that words with high polarity scores often received greater attention weights, making the model's decisions easier to understand. These findings are relevant to this project's use of neural networks, as attention mechanisms could improve both the accuracy of classification and the clarity of how feedback is categorized.

Finally, Petersen et al. (2022) proposed a differentiable top-k cros-entropy classification loss function, designed to optimize networks for multiclass classification. Their work showed that relaxing the top-k constraint not only improved top-5 accuracy but also led to enhancements in top-1 accuracy. This influenced the decision to additionally analyze the category classifiers in this project using top-3 accuracy, offering a more nuanced perspective on model performance in handling closely related product categories. This approach is particularly relevant when overlapping clases may obscure a single top prediction, yet still capture useful information in the top-3 predictions.

Discussion of Ideas

A combination of established research and practical consideration drove the approach in this project. One key decision was whether to use pre-trained embeddings models or traditional techniques like TF-IDF to represent the text. As highlighted by Piskorski and Jacquet (2020), TF-IDF with character n-grams can perform exceptionally well in specific contexts, even outperforming models based on embeddings like GloVe and BERT. While embeddings can offer broad applicability and powerful generalization across tasks, TF-IDF's ability to adapt to the language and structure of a given corpus makes it well-suited for classifying product reviews where context specific language is crucial to the task. Moreover, TF-IDF is computationally efficient and scales effectively, especially when compared to embeddings, which often require substantial resources for training and fine-tuning.

The decision to enhance text representation with POS tagging was driven by its ability to add grammatical structure to the features, distinguishing between functions such as verbs, nouns, and adjectives. As demonstrated by Gyawali et al. (2013), POS-tagged n-grams can capture stylistic and structural nuances, making them particularly effective in classification tasks involving diverse writing styles. For customer feedback, where tone and grammar often vary significantly, POS tagging has the potential to improve both the interpretability and precision of feature extraction. Similarly, n-grams ranging from 1 to 4 were used to tokenize the text, with the hope of enhancing the contextualization capabilities of the classifiers trained on this data.

Support Vector Machines were chosen as a candidate model for their strong track record in classification tasks. Piskorski and Jacquet (2020) validated the effectiveness of SVMs, showing their competitive performance in text classification. The ability of SVMs to employ kernel functions, such as the radial basis function, allows them to detect nonlinear relationships in data, making them a robust choice for handling complex feedback. Logistic regression and K-nearest neighbors were also considered as potential models for this research. However, they were eliminated early on as logistic regression can struggle with nonlinear patterns, and K-NNs are not trained to recognize specific patterns and instead rely on identifying similarities between data points, which may not always be present or well-represented in the training data.

Neural Networks have become central in NLP tasks and were chosen for their ability to capture nonlinear relationships and unique patterns within data. Sun and Lu (2020) demonstrated that attention mechanisms within neural networks not only improve performance but also provide interpretability by highlighting the features most relevant to a prediction. This dual capability makes NNs particularly appealing for analyzing customer feedback, where subtle patterns can be crucial for accurate classification.

The choice of evaluation metrics was straightforward. Standard metrics like accuracy, precision, recall, and F1-scores were used to assess the performance of the models in this project. Additionally, top-k versions of these metrics were applied to evaluate the product category classifiers. Inspired by Petersen et al. (2022), top-k metrics offer a broader perspective on model performance by accounting for cases where the correct label appears among the top predictions, even if it is not the top-ranked choice. This approach is particularly valuable for understanding how models handle closely related categories and identifying areas where performance can be improved.

Data Processing and Methodology

This study utilized the 2023 Amazon Reviews Dataset, encompassing customer reviews from 10 product categories. Each review included fields such as title, body, rating (1–5), and product category. For sentiment classification, ratings were grouped into binary classes: ratings 1-2 were labeled negative, and 4-5 were labeled positive. Reviews with a rating of 3 were discarded for the purposes of training the sentiment classifier as they were thought to be too ambiguous for the binary outcome. A total of 5,000 reviews were sampled, ensuring equal representation across product categories and sentiment classes to maintain balance for training and evaluation.

Preprocessing involved text normalization using SpaCy (Honnibal et al., 2020), including stopword removal, punctuation filtering, and POS tagging. Two distinct feature sets were created: one incorporating POS tags and one without them. Features were represented using TF-IDF weighting applied to n-grams ranging from unigrams to four-grams. To optimize the feature space, Chi-squared

analysis was conducted to identify statistically significant features relevant to each classification task. This process helped reduce dimensionality while retaining the most impactful features for training.

Models were developed for both sentiment classification and product category classification tasks using SVMs and NNs. Initial modeling began with a base SVM trained on all features from both preprocessing pipelines. Following Chi-squared analysis, SVM models were optimized and retrained on the reduced feature sets to improve efficiency and performance. Feed-forward neural networks were then constructed and trained using the Chi-squared reduced feature set. These models incorporated attention mechanisms to dynamically emphasize critical features during training. Models were evaluated using both POS-tagged and non-POS tagged feature sets to assess the impact of POS tagging on classification tasks.

The performance of these models was evaluated using standard accuracy, F1-score, precision, and recall. For the product category classification task, metrics were adapted to a top-K framework, with K=3, for a more flexible analysis of model ranking performance. Under this framework, the model was rewarded if the true label appeared in the top-3 predictions. This approach was only applied to the best-performing category classifier to provide additional insights into its ability to manage closely related classes.

Results: Sentiment Classification

As outlined in the section above, each classification task began with training a simple SVM model to establish a baseline for the analysis. The model was configured with a linear kernel and a regularization parameter (C) of 1.0. While the results were not expected to be outstanding, they served as a benchmark for improvement in subsequent experiments. The results of the baseline model trained on the POS and non-POS tagged datasets are presented in Table 1. below. It is evident the model trained on the non-POS dataset demonstrated slightly better performance across all metrics, achieving an accuracy of 78%, compared to 74% for the POS-tagged dataset. The non-POS dataset showed balanced precision and recall for both positive and negative classes, leading to an F1-score of 0.77 for negative and 0.78 for positive. The POS-tagged model, while slightly underperforming the non-POS model, still provided consistent results, with an F1-score of 0.74 across both classes. These results suggest that POS tagging has minimal impact on model performance and may even hinder the model's ability to learn discriminative patterns in the data.

	SVM Base Non POS (Features = 5279)			**SVM Base POS (Features = 5263)**			
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

Table 1: Sentiment classification: base SVM

The baseline model was trained using all features extracted during the preprocessing steps, which for both sets was over 5,000 features. To reduce noise and allow for the SVM to better generalize patterns in the data, Chi-squared feature selection was performed. Only features with a Chi-squared value above 15 were selected. This value was selected as the threshold as it is highly significant ($p <$

0.01) for this experiment and because it still retained close to 300 features per set. Table 2. shows the 10 most importance features from each feature set.

Non-Pos	'great', 'not', 'love', 'return', 'good', 'easy', 'waste', 'perfect', 'not buy', 'easy to'
POS	'[PART]not', '[ADJ]great', '[VERB]love', '[ADJ]good', '[VERB]return', '[ADJ]easy', '[PART]not [VERB]buy', '[ADP]for', '[ADJ]easy [PART]to', '[ADJ]disappointed'

Table 2. Sentiment classification: Top 10 Chi-squared selected features

The SVM model was optimized using a grid search, which identified 1.0 as the optimal regularization parameter and a radial basis function as the kernel. The results of this model trained on the Chi-squared selected features are shown in Table. 3.

	SVM Chi Feats Non POS (Features = 271, $\chi^2 > 15$)			**SVM Chi Feats POS (Features = 269, $\chi^2 > 15$)**			
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

Table 3: Sentiment classification: optimized SVM with Chi-squared selected features

Overall, both models achieved the same accuracy (81%) and identical macro and weighted averages (0.81 for all metrics). The non-POS dataset slightly outperformed the POS-tagged dataset for the negative class, achieving a recall of 0.84 and an F1-score of 0.82 compared to 0.83 and 0.82 for the POS-tagged model. For the positive class, both models achieved an F1-score of 0.81, with the non-POS dataset showing a marginally lower recall (0.78) compared to 0.79 for the POS dataset. These results suggest that chi-squared feature selection is effective for dimensionality reduction while preserving and improving performance. However, POS tagging again shows limited benefit, as the non-POS dataset matches or slightly exceeds the POS dataset's performance across most metrics.

The same chi-squared selected feature set was used to train a feed-forward NN. The architecture was the NN was the same to train both sets: two hidden layers, one with 64 neurons and the next with 16. The results of this experiment is seen below in Table 4.

	NN Chi Feats Non POS (Features = 271, $\chi^2 > 15$)			**NN Chi Feats POS (Features = 269, $\chi^2 > 15$)**			
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

Table 4: Sentiment classification: NN with Chi-squared selected features

Accuracy for both datasets was identical at 84%, with matching macro and weighted averages across all metrics. Interestingly, as seen in the optimized SVM results, the non-POS dataset performed better on negative recall and positive precision, while the POS dataset performed better on negative

precision and positive recall. However, these differences are not significant enough to identify a clear advantage for either approach.

Results: Category Classification

The approach for classifying product categories from feedback followed the same process as sentiment classification. The baseline SVM model was trained on all features extracted during preprocessing to establish a benchmark. The results for this baseline model are presented in Table 5. The accuracy scores of 0.43 for the non-POS feature set and 0.41 for the POS feature set indicate marginally better performance without POS tagging. Among the categories, the best-performing class for the non-POS feature set is Grocery and Gourmet Food, with an F1-score of 0.66, while the POS feature set achieves its highest F1-score of 0.61 for Amazon Fashion and Grocery and Gourmet Food. Conversely, the Industrial and Scientific category is the worst-performing for both feature sets, with F1-scores of 0.22 for the non-POS set and 0.19 for the POS set, highlighting consistent challenges in classifying this category. These results suggest that while non-POS tagging slightly outperforms POS tagging overall, classification performance varies considerably by category.

	SVM Base Non POS (Features = 6644)			SVM Base POS (Features = 6687)			
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

Table 5: Category classification: base SVM

Once again, Chi-squared feature selection was applied using a threshold of 30. The top 10 features from each category were almost identical between the two feature sets, shown below.

All Beauty	'hair', 'skin', 'wig', 'shave', 'scent', 'brush', 'nail', 'shaver', 'face', 'eyebrow'
Amazon Fashion	'dress', 'wear', 'shirt', 'fit', 'bra', 'chain', 'waist', 'outfit', 'size', 'to'
Appliances	'filter', 'coffee', 'ice', 'water', 'fridge', 'dryer', 'refrigerator', 'part', 'stove', 'cup'
Arts Crafts and Sewing	'yarn', 'project', 'paint', 'bead', 'sewing', 'needle', 'color', 'mold', 'paper', 'backing'
Baby Products	'baby', 'diaper', 'stroller', 'seat', 'pillow', 'son', 'wipe', 'monitor', 'diaper bag', 'strap'
Grocery and Gourmet Food	'taste', 'flavor', 'tea', 'delicious', 'chocolate', 'sugar', 'sweet', 'taste like', 'stale', 'tasty'
Industrial and Scientific	'filament', 'nozzle', 'carpet', 'cleaner', 'unit', 'vacuum', 'test', 'pipe', 'print', '3d'
Office Products	'printer', 'pen', 'ink', 'cartridge', 'page', 'pencil', 'phone', 'print', 'label', 'scan'
Pet Supplies	'dog', 'cat', 'litter', 'food', 'pet', 'tank', 'dog love', 'collar', 'chew', 'cat not'
Tools and Home Improvement	'light', 'bulb', 'faucet', 'bright', 'lamp', 'battery', 'knife', 'tool', 'led', 'of light'

Table 6. Category classification: Top 10 Chi-squared selected features

This feature selection resulted in almost 886 features for the non-POS set and 685 features for the POS set. These were then used to train optimized SVM classifiers, which identified a 1.0 regularization parameter and radial basis function as the kernel. The results are presented in Table 7.

	SVM Chi Feats Non POS (Features = 886, $\chi^2 > 30$)			SVM Chi Feats POS (Features = 685, $\chi^2 > 30$)			
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

Table 7: Category classification: optimized SVM with Chi-squared selected features

There was a significant lift in performance across the board. The accuracy of the non-POS feature set increased to 52%, and the POS-tagged feature set increased to 51%, exemplifying the benefits of feature selection for classification tasks. Notably, the Grocery and Gourmet Food category, still the best-performing class for the non-POS feature set with an F1-score of 0.66, improved further to 0.72. Similarly, the POS feature set saw its strongest category, Amazon Fashion, increase from an F1-score of 0.61 to 0.66. The Industrial and Scientific category, which had been the worst-performing category for both feature sets, also showed improvement, with F1-scores increasing from 0.22 to 0.30 for the non-POS set and from 0.19 to 0.27 for the POS-tagged set.

Next, a NN with two hidden layers, the first with 128 neurons and the next with 32, was trained using the Chi-squared selected features. The results of this model are presented in Table 8.

	NN Chi Feats Non POS (Features = 886, $\chi^2 > 30$)			NN Chi Feats POS (Features = 685, $\chi^2 > 30$)			
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

Table 8: Category classification: NN with Chi-squared selected features

With the NN model trained on the Chi-squared reduced feature sets, there were further gains in performance across both feature sets. The overall accuracy increased to 55% for the non-POS set and 54% for the POS-tagged set, showing consistent improvement over the SVM models. For the non-POS feature set, Grocery and Gourmet Food remained the best-performing category, achieving an F1-score of 0.77, up from 0.72 with the SVM. Similarly, the POS feature set also saw its top-performing category, Amazon Fashion, improve from 0.66 to 0.69. However, Industrial and Scientific continued to be the most challenging category for both feature sets, with only slight improvements. The F1-score increased to 0.31 for the non-POS set and decreased to 0.19 for the POS-tagged set, indicating persistent difficulties in distinguishing this category due to overlapping language or limited representative features in the data.

To gain a deeper understanding of the Neural Network model's performance, particularly its ability to handle closely related product categories, it was evaluated using a top-K framework, with K=3, as a more relaxed assessment method. The results of this evaluation framework are presented in Table 9.

	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

Table 9: Category classification: NN with Chi-squared selected features using Top-K framework

The top-K framework demonstrated significant improvements in the model's ability to detect product categories. While this does not necessarily reflect its capacity for exact match classification, it does indicate the model's strength in accurately ranking the true class within its top-3 predictions. With a top-K accuracy of 76% and identical macro and weighted average F1-scores, the model clearly exhibits an ability to learn and apply the right patterns for distinguishing between categories. Notably, the Industrial and Scientific category showed substantial improvement, achieving an F1-score of 0.67, confirming that previous challenges were likely due to overlapping features and ambiguity between classes.

The prediction results for both the top-1 and top-3 performance frameworks are presented as confusion matrices in Figure 1. These visualizations provide valuable insight into the challenges the model faces, particularly due to class similarities and overlapping features. The top-1 confusion matrix highlights the areas where the model struggles to correctly classify certain categories, such as Industrial and Scientific, which often exhibits feature overlap with other technical categories. However, when the evaluation framework was relaxed to top-3 predictions, the confusion matrix reveals a marked improvement in the model's ability to rank the true class within the top predictions. This shift demonstrates the model's capacity to identify relevant patterns even when exact matches are challenging.

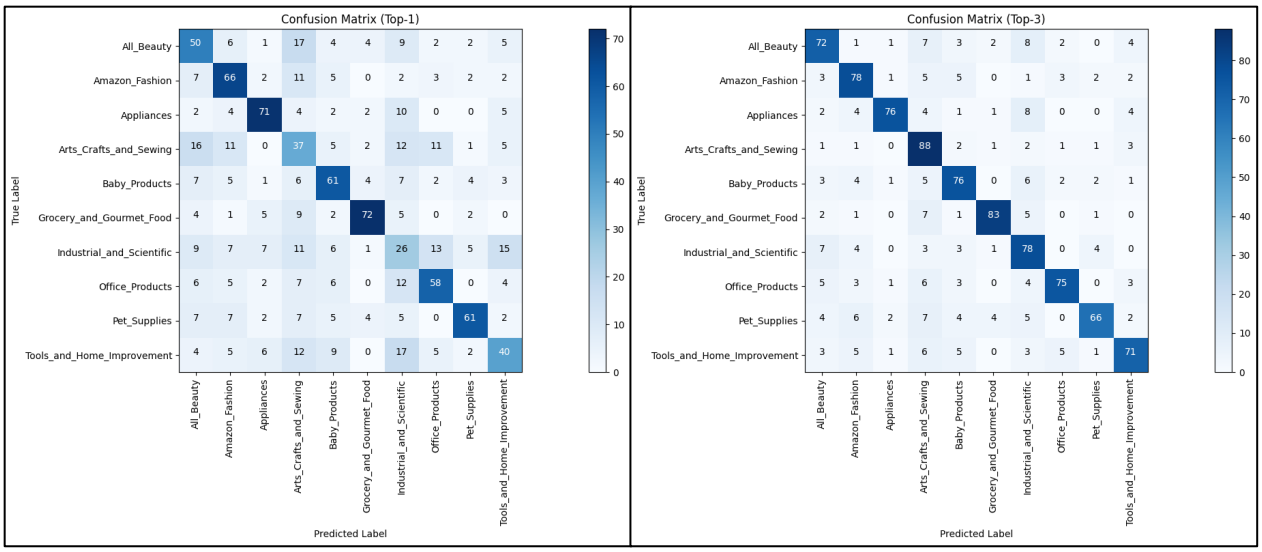

Figure 1: Category classification: NN with Chi-squared selected features results confusion matrix

Conclusion

Understanding and effectively classifying customer feedback is crucial for businesses to enhance customer experience and make data-driven decisions. This study highlights the challenges and opportunities in structuring unstructured customer reviews, focusing on sentiment and product category classification. By employing TF-IDF and n-grams as feature representations, it provides a pathway to navigate the complexities of diverse feedback data. The findings emphasize the need for efficient methods to extract actionable insights from large volumes of customer feedback.

The results of this study underline several key insights. Feature selection using Chi-squared analysis proved invaluable in reducing dimensionality and improving model performance. However, contrary to initial hypotheses, the inclusion of POS tagging did not enhance the models and even detracted from their effectiveness. Neural networks outperformed SVMs in both sentiment and category classification tasks, demonstrating their superior ability to capture non-linear relationships and feature interactions. These outcomes confirm the hypotheses regarding the strength of neural networks and the limited utility of POS tagging in this context.

Building on the findings from the top-K framework, future research could explore optimizing models specifically for top-K accuracy. As suggested by Petersen et al., training neural networks with differentiable top-K cross-entropy loss could further enhance their ability to manage closely related classes. Additionally, expanding the dataset and incorporating advanced pre-trained language models could be investigated to improve performance, especially for categories with high overlap or sparse features. These advancements would pave the way for even more robust and scalable solutions in customer feedback analysis.

References

- Binod Gyawali, Gabriela Ramirez, and Thamar Solorio. 2013. Native Language Identification: a Simple n-gram Based Approach. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–231, Atlanta, Georgia. Association for Computational Linguistics.
- Felix Peterson, Hilde Kuehne, Christian Borgelt, and Oliver Deussen. 2022. Differentiable Top-K Classification Learning. In *Proceedings of the International Conference on Machine Learning*, pages 17656–17668. PMLR.
- Jakub Piskorski and Guillaume Jacquet. 2020. TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 26–34, Marseille, France. European Language Resources Association (ELRA).
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. DOI: 10.5281/zenodo.1212303.
- Xiaobing Sun and Wei Lu. 2020. Understanding Attention for Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online. Association for Computational Linguistics.
- Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging Language and Items for Retrieval and Recommendation. arXiv preprint arXiv:2403.03952.