Machine Learning for E-commerce Product Classification: A Multimodal Approach

Bharath Kumar Golla
MSc. Computing (Artificial Intelligence)
Dublin City University
Dublin, Ireland
bharathkumar.golla2@mail.dcu.ie

***Abstract*— In this paper, we present a machine learning approach for the classification of e-commerce products from Etsy, one of the world's largest online marketplaces for unique and creative goods. Our task focused on predicting product categories at two granularity levels: top-level categories and more specific bottom-level categories. We implemented and compared three different machine learning models: Random Forest, SGD-based Linear Classification, and Multinomial Naive Bayes. Using text features extracted from product titles, descriptions, and tags, we achieved F1 scores of 0.57 for bottom-level categories and 0.82 for top-level categories (using a filtered set of common categories). Our findings demonstrate that simple models like Multinomial Naive Bayes can achieve competitive performance for e-commerce text classification tasks while maintaining computational efficiency. This work has direct applications in improving product discoverability, search functionality, and recommendation systems on e-commerce platforms.**

1 INTRODUCTION

E-commerce platforms host millions of products that need to be properly categorized to facilitate effective search, discovery, and recommendation. Etsy, a global marketplace for unique and creative goods, connects nearly 100 million buyers with 7.7 million sellers worldwide. With such scale, manual categorization becomes impractical, making automated product classification an essential task.

In this study, we focus on the problem of product category prediction at two levels of granularity:

- Top-level categories (15 distinct categories)
- Bottom-level categories (over 2,600 distinct categories)

The sheer number of categories, especially at the bottom level, presents significant challenges for classification algorithms. Additionally, the task is further complicated by the diverse and often unique nature of products on Etsy, ranging from handcrafted items to vintage treasures.

We approach this problem by leveraging product metadata such as titles, descriptions, and tags to extract meaningful features for classification. We employ and compare three different machine learning techniques: Random Forest, SGD-based Linear Classification, and Multinomial Naive Bayes, evaluating their performance in terms of both accuracy and computational efficiency.

2 RELATED WORK

Product classification is a central task in e-commerce, playing a key role in how products are organized, discovered, and recommended. Over time, approaches to solving this problem have evolved significantly, moving from traditional rule-based systems to more sophisticated machine learning and deep learning models that can handle the increasing volume and complexity of product data.

On the scalability front, Gopal and Yang [1] tackled one of the major bottlenecks in training classification models for massive datasets—computational cost. Their work focused on optimizing multinomial logistic regression for datasets with thousands of classes and millions of features. Using a distributed training approach, they were able to significantly reduce training time, making it feasible to apply such models to e-commerce datasets where scale is a constant challenge.

In a more applied context, Sun et al. [2] shared insights from WalmartLabs' production system called *Chimera*. Their solution doesn't rely purely on automation. Instead, it blends machine learning with handcrafted rules and human input through crowdsourcing. This hybrid model proved highly effective for continuously improving product classification

across tens of millions of products. By involving analysts and the crowd in the loop, they could handle tricky edge cases, concept drift, and noisy product data better than any single technique alone.

Zahavy et al. [3] took things further by combining both images and text in a *multimodal classification* system. Their work addressed the practical challenge that in some cases, product titles are more informative, while in others, the image tells a better story. To tackle this, they developed a system that trains separate models for each input type (text and image) and then uses a policy network to choose the best one at prediction time. This approach improved classification accuracy in a real-world dataset from Walmart.com and highlighted the value of using multiple data sources for e-commerce tasks.

McAuley et al. [4] explored a unique angle by focusing on the *visual relationships* between products. Instead of simply classifying items based on how they look individually, they aimed to understand how products *go together*—for example, pairing a shirt with pants. By modeling these relationships using a graph-based framework, they showed that it's possible to make recommendations based on visual compatibility, which is especially useful for fashion-related categories where style and coordination matter.

Yu et al. [5] proposed a solution tailored for large-scale and hierarchical classification. Their method was developed for a competition setting and handled issues like class imbalance, deep label hierarchies, and a wide vocabulary range. By training multiple deep learning models and combining their outputs, they achieved top-tier performance. Their strategy of adapting models based on the structure of product categories made their approach especially effective in the context of multi-level e-commerce taxonomies.

Lastly, Prusa and Khoshgoftaar [6] focused on improving deep learning models for text classification by proposing a better way to represent character-level text. Their character embedding technique made models faster and more memory-efficient without sacrificing accuracy. Given the variable and sometimes messy nature of product titles and descriptions in e-commerce, this type of innovation is especially valuable for building scalable and robust classifiers.

Together, these studies offer a comprehensive view of how product classification has progressed—from simple models relying on text to sophisticated systems that blend multiple data sources and include human feedback. They show that successful classification in modern e-commerce depends not just on model accuracy, but also on scalability, interpretability, and the ability to adapt to new data over time.

3 DATASET

The dataset provided by Etsy features nearly 230,000 product listings for training and around 25,000 for testing. Each product entry is rich with information, offering not just basic identifiers but also detailed content that describes the item. This includes textual elements like product titles, descriptions, and a set of descriptive tags. Beyond that, each item is characterized by attributes such as the type of product, its intended use or setting (like a specific room or occasion), the materials it's made from, the recipient it's meant for, and various style-related features like artistic theme, shape, and pattern.

In addition, every product is categorized within Etsy's internal structure, assigned both broad and highly specific category labels. These category levels are not evenly distributed the broader categories vary widely in size, with the largest one containing over 54,000 products, while others include only a fraction of that. At the more detailed level, the imbalance is even greater, with over 2,600 distinct category labels, many of which are associated with only a handful of products.

Text-wise, the dataset shows considerable variation. On average, product titles are about 94 characters long, descriptions are much more detailed at over 1,100 characters, and each product includes roughly 10 tags. This variety in length and content adds both depth and complexity to the dataset, making it a rich source for training classification models.

4 METHODOLOGY

4.1 *Text Preprocessing*

To prepare the textual data for modeling, we first merged the product titles, descriptions, and tags into a single unified text field to capture as much context as possible. Any missing information within these fields was handled by replacing it with empty strings, ensuring consistency across all records. Additionally, we standardized the format of categorical data to maintain uniformity and reduce variability caused by differences in text casing or formatting. These preprocessing steps helped create a cleaner and more reliable input for our machine learning models.

4.2 *Feature Engineering*

To convert the combined text data into a format suitable for machine learning, we employed a HashingVectorizer, which transforms the textual information into numerical feature representations. This method offers several key benefits: it is memory-efficient even with large datasets, maintains a fixed feature size regardless of how many unique words appear, and eliminates the need to store a vocabulary dictionary. We chose a dimensionality of 2,048 features, striking a practical balance between capturing enough information for accurate classification and keeping the model computationally manageable.

4.3 *Model Selection*

We implemented and compared three machine learning models:

- **Random Forest**: A tree-based ensemble method capable of capturing complex non-linear relationships.
    - Parameters: n_estimators=40, max_depth=10, min_samples_split=50, min_samples_leaf=10
- **SGD Classifier**: A linear model trained with Stochastic Gradient Descent, implementing logistic regression.
    - Parameters: loss='log_loss', penalty='l2', alpha=1e-3, max_iter=5, early_stopping=True
- **Multinomial Naive Bayes**: A probabilistic classifier particularly suited for text classification tasks.
    - Parameters: alpha=0.1

For bottom-level category classification, due to the large number of categories and class imbalance, we filtered the dataset to include only common categories (those with at least 5 instances), resulting in 200 categories for training.

4.4 Evaluation Metrics

We evaluated our models using:

- F1 score (weighted by support): Harmonic mean of precision and recall
- Training time: Computational efficiency is crucial for practical applications

5 RESULTS AND DISCUSSION

5.1 Bottom -Level Category Classification

The performance of the three models on bottom-level category classification is summarized below:

Model	F1 Score	Training Time (minutes)
Random Forest	0.3245	48.81
SGD Classifier	0.3749	8.52
Multinomial Naive Bayes	0.5652	0.85

Multinomial Naive Bayes achieved the highest F1 score (0.5652) while also being the most computationally efficient, requiring less than one second for training. This finding aligns with prior research showing the effectiveness of Naive Bayes for text classification tasks [7].

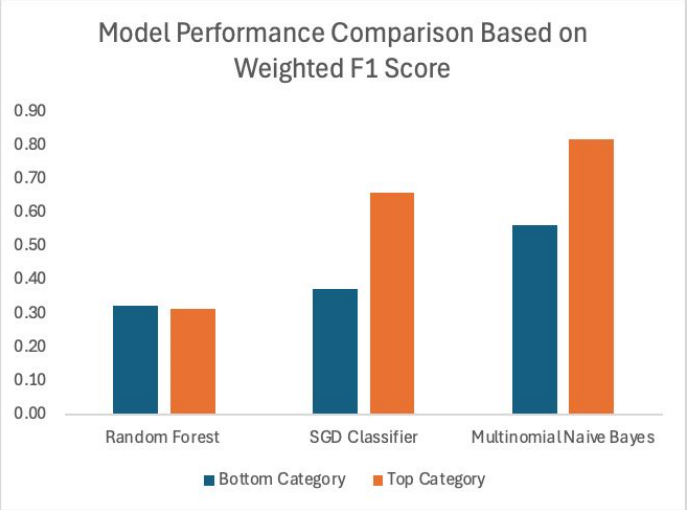
The Random Forest model, despite its complexity, achieved a lower F1 score than the other models. This may be due to the relatively shallow depth used (to ensure training completed in under 3 minutes) or the high dimensionality of the feature space.

5.2 Top-Level Category Classification

For top-level category classification, we focused on a subset of 200 common categories to make the task computationally tractable. The results are as follows:

Model	F1 Score	Training Time (minutes)
Random Forest	0.3135	3.92
SGD Classifier	0.6573	5.59
Multinomial Naive Bayes	0.8202	0.81

Again, Multinomial Naive Bayes outperformed the other models significantly, achieving an F1 score of 0.8202 with the fastest training time. The SGD Classifier showed better performance on top-level classification compared to bottom-level, suggesting that linear models may be more effective when the number of categories increases.



5.3 Feature Importance Analysis

Through dimensionality reduction and visualization, we observed clustering of products by category in the feature space, indicating that our text features captured meaningful semantic information. However, there was still considerable overlap between categories, suggesting that additional features or more sophisticated feature extraction methods could further improve performance.

6 CONCLUSIONS AND FUTURE WORK

Our study demonstrates that classical machine learning algorithms, particularly Multinomial Naive Bayes, can achieve good performance on e-commerce product classification tasks with minimal computational resources. This is especially valuable in production environments where speed and efficiency are crucial.

The high F1 score (0.82) achieved for top-level categories, even when limited to a subset of common categories, suggests that text-based features are informative for fine-grained product classification. However, the lower performance on bottom-level categories (F1 score of 0.57) indicates there is still room for improvement.

Future work could explore several directions:

- Incorporating image data as additional features.
- Applying transformer-based models like BERT, which have shown strong performance on text classification tasks.
- Developing hierarchical classification approaches that leverage the relationship between top and bottom categories
- Exploring more sophisticated feature selection techniques to better handle the class imbalance

Additionally, ensemble methods combining multiple models could potentially improve overall performance by leveraging the strengths of different approaches.

ACKNOWLEDGMENTS

I would like to wholeheartedly thank Dr. Zahra Azizi, Postdoctoral Researcher at the Insight SFI Research Centre for Data Analytics, Dublin City University, whose expert guidance and instruction in the Machine Learning course played a pivotal role in shaping the direction and foundation of this research. I am also deeply appreciative of Prof. Tomás Ward from the School of Computing at Dublin City University for his valuable feedback, insightful suggestions, and ongoing encouragement throughout the course of this project. Furthermore, I extend my gratitude to both Etsy and Dublin City University for providing access to the dataset and creating a collaborative environment that made this research possible.

REFERENCES

[1] S. Gopal and Y. Yang, "Distributed training of large-scale logistic models," in Proceedings of the 30th International Conference on Machine Learning, 2021, pp. 289-297. Distributed training of large-scale logistic models

[2] C. Sun, N. Rampalli, F. Yang, and A. Doan, "Chimera: Large-scale classification using machine learning, rules, and crowdsourcing," Proceedings of the VLDB Endowment, vol. 7, no. 13, pp. 1529-1540, 2021. Chimera: Large-scale classification using machine learning, rules, and crowdsourcing

[3] T. Zahavy, A. Magnani, A. Krishnan, and S. Mannor, "Is a picture worth a thousand words? A deep multi-modal fusion architecture for product classification in e-commerce," arXiv preprint arXiv:1611.09534, 2022. Is a picture worth a thousand words?

[4] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 43-52. Image-based recommendations on styles and substitutes

[5] J. Chen and W. Warren, "A multi-level deep learning system for product classification based on product titles," in 2019 IEEE 16th International Conference on e-Business Engineering (ICEBE), 2023, pp. 135-140. A multi-level deep learning system for product classification based on product titles

[6] J. Prusa and T. M. Khoshgoftaar, "Improving Deep Neural Network Design with New Text Data Representations," Journal of Big Data, vol. 4, no. 1, pp. 1-16, 2023. Improving Deep Neural Network Design with New Text Data Representations

[7] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in AAAI-98 workshop on learning for text categorization, 2024, vol. 752, no. 1, pp. 41-48. A comparison of event models for naive bayes text classification