Optimizing E-Commerce Product Categorization Using Textual Data

Omkar Sudhir Khaladkar
Dublin City University
Student ID: 23262297 Email: omkar.khaladkar2@mail.dcu.ie

***Abstract*—Online shopping has become a crucial part of daily life, revolutionizing how consumers access and purchase products. This paper explores an innovative machine learning approach to enhance the user experience on Etsy, a leading e-commerce platform known for unique, handmade, and vintage items. The primary challenge addressed is the efficient classification of products into categories such as 'top category', 'bottom category', 'primary color', and 'secondary color', crucial for navigating the vast and diverse offerings on Etsy. Our methodology involves analyzing product titles, descriptions, and tags using various machine learning techniques, including K-Means clustering, KNN Classifier, Multinomial NB, and Multi-label Logistic Regression, to predict customer preferences and authenticate products against potential fraud. The Multi-label Logistic Regression model demonstrated superior performance, achieving the highest accuracy in identifying genuine products, as evaluated by F1 scores and a detailed classification report. This research aims to simplify the shopping process on Etsy, enabling users to easily find and confidently purchase products that meet their preferences and ensure authenticity.**

Keywords - ***Etsy, Machine Learning, Multi-label Logistic Regression, Product Authentication, Fraud Detection, Classification Techniques, Product Categorization, Handmade and Vintage Items, Predictive Modeling, K Means Clustering, KNN Classifier, Multinomial NB, F1 Scores.***

I. Introduction

Online shopping has become a regular part of our daily lives and the way we buy things is constantly changing. The world of E-commerce is always evolving, with new trends and technologies emerging all the time. Digital marketplaces that focus on niche markets are on the rise, as exemplified by platforms such as Etsy. This website specializes in providing unique, vintage, and handmade items to its customers. These platforms provide a wide range of products, each with unique characteristics and classifications that challenge the traditional retail model.

To enhance the user experience and operational efficiency, navigating through Etsy's extensive and varied inventory requires sophisticated technological solutions. In this regard, our paper proposes a novel machine-learning approach aimed at improving product classification and authentication. Our model leverages advanced techniques that analyze both textual and visual data to predict customer preferences accurately. Furthermore, it ensures the authenticity of listed products, addressing key challenges in digital commerce.

Our approach involves utilizing pre-trained models for both text and image analysis and implementing various machine-learning frameworks to extract embeddings from product titles, descriptions, and images. These embeddings play a crucial role in accurately categorizing products into specific categories such as 'top category', 'bottom category', 'primary color', and 'secondary color', which ultimately leads to a more personalized and secure shopping experience.

The paper will demonstrate the effectiveness of our approach through rigorous testing with various machine learning models, presenting detailed analyses based on F1 scores and classification reports. The goal is to show that our model not only meets but exceeds the current standards for product classification and authentication, making Etsy a safer and more enjoyable shopping destination.

II. Related Work

In the digital age, e-commerce platforms like Etsy are continuously looking for innovative ways to improve the shopping experience for their users. The incorporation of machine learning into product categorization not only automates the process, making navigation more intuitive for customers, but also enhances various aspects of business operation such as customer segmentation, demand forecasting, and fraud detection. This technological integration aims to meet consumer demands more effectively and maintain a competitive edge by understanding buying behaviors for better-targeted marketing and personalized product recommendations.

Among the notable advancements in this field is the research as outlined [1] a system specifically for Adidas, utilizing deep learning techniques such as Convolutional Neural Networks and template matching to automate the categorization of fashion products. This system showcased a remarkable 90.8% accuracy in logo detection and proved its worth in handling real-world variations in product images, demonstrating how advanced machine learning techniques can be applied practically in a commercial setting to enhance product categorization.

Similarly, [2] introduces an intriguing twist to the conventional categorization process in their study, They employed machine translation techniques to convert product descriptions into sequences of category tokens, effectively transforming the product taxonomy into a more flexible and dynamic system. [2] not only improved categorization accuracy but also adapted seamlessly to new products, thereby presenting a significant advancement in making e-commerce platforms more adaptable and accurate.

Meanwhile, [3] demonstrated how supervised machine learning can be employed to significantly enhance the efficiency of product categorization. Their classifier, developed for grouping products into categories like 'Electronics' or 'Automotive,' achieved accuracy improvements from initial rates of around 70% to about 75% with further refinements. This highlights the potential of machine learning to reduce manual efforts in categorization, supporting more effective marketing and operational strategies within e-commerce settings.

The exploration of new methods for text categorization [5] further enriches this narrative by introducing term weighting methods that enhance the discrimination of terms across categories, thus significantly improving classification accuracy. These methods, particularly effective in multi-class settings, underscore the adaptability and potential of machine learning techniques to transform traditional business operations.

Collectively, [6] provides a compelling evidence of the transformative power of machine learning in e-commerce. By automating categorization and enhancing accuracy, these advanced techniques not only simplify the shopping process but also build a more reliable and engaging marketplace. As e-commerce continues to evolve, the integration of such technologies will undoubtedly play a critical role in shaping the future of online shopping, ensuring that platforms like Etsy remain at the forefront of the digital marketplace.

III. Data

The company Etsy has provided a dataset that contains a vast collection of their online marketplace listings [7]. This dataset currently includes around 100 million items from more than 5 million active sellers. For the purpose of machine learning, a portion of these listings is available as training data and testing data in a compressed file of 13GB.The Parquet files in this dataset have been organized into distinct folders for training and testing purposes. This setup is ideal for both training and evaluating machine learning models as it offers a well-structured framework. The dataset captures the richness of Etsy's product range through detailed textual and visual data. Textual content includes product identifiers, titles, descriptions, and tags, while visual content features encoded images with their dimensions—ideal for image-based machine learning tasks.

IV. Methodology

A. Exploratory Data Analysis

Next, we dive into the exploratory data analysis (EDA) to understand the structure and quality of our dataset more deeply:

Training Data: Contains 229,624 records across 26 columns.
Test Data: Includes 25,514 records in 18 columns.
Completeness: All attributes are fully populated with no missing values.
Categorization: The data features 15 unique classes for the Top Category and 2,609 classes for the Bottom Category. It also details around 19 distinct colors for both primary and secondary classifications.

B. Data Processing

The dataset required processing for enhancing the accuracy and precision analysis. We removed several columns including 'room', 'craf_type', 'recipient', 'material', 'occasion', 'holiday', 'art_subject', 'style', 'shape', and 'pattern' because they weren't crucial to our analysis. The 'type' column was also dropped as it provided little insight. For the remaining text data in the 'tags', 'titles', and 'descriptions' columns, we employed Natural Language Processing (NLP) techniques to refine the data:

We removed unnecessary elements like special characters, digits, URLs, extra spaces, and emojis to focus only on meaningful words, achieving a cleaner and more standardized dataset. Commonly used but less informative words, known as stopwords (e.g., 'the', 'an', 'is', 'a'), were eliminated to highlight more significant words in the text. Lemmatization was applied to convert words to their base forms, helping to normalize the text and reduce the vocabulary size. This simplification not only eases the computational process but also improves the model's accuracy by providing a more consistent input format. These cleaning steps were specifically applied to the 'title', 'description', and 'tags' columns, making them more effective for analyzing and classifying product information.

C. Feature Engineering

In feature engineering, we worked on refining our dataset. First, we cleaned up various columns by removing unnecessary information. Then, we merged these cleaned columns together to create a new one merged[title + description + tags] column. This new column held all the cleaned text from the original columns, such as titles, descriptions, and tags. We separated these texts with spaces in the new merged column. This consolidation helped us gather all relevant information into one place for further analysis.

To ensure we didn't lose our progress, we saved a backup of this cleaned data as a CSV file. This way, we could always refer back to the cleaned version if needed, without having to redo the entire cleaning process. This step was crucial for maintaining the integrity of our data and ensuring that we could confidently proceed with our analysis.

D. Model Building

The main aim of our model was to select the best model for classifying product IDs on Etsy. We tested four different models to see which one performs the best. To manage the large dataset efficiently, we broke it down into smaller batches. This batch processing strategy helped us process data faster and use less memory.

Before training the models, we converted the text data into a numerical format using a tool called CountVectorizer. It is valued for its simplicity and effectiveness in text analysis, particularly where the frequency of words directly impacts

classification. It maintains raw word counts, making it straightforward and computationally efficient, ideal for large datasets or limited processing resources. In our research, we tested various machine learning models tailored for different tasks, including the KNN Classifier, Multinomial Naive Bayes (NB), and Multi-label Logistic Regression.

The K-Nearest Neighbors (KNN) Classifier is celebrated for its simplicity, requiring no model training, making it highly intuitive and easy to implement. It's extremely versatile, handling various data types and adaptable with different distance metrics. KNN is also robust to model assumptions, capable of effective multi-class classification, and performs well with sufficient representative data. However, It is sensitive to noisy data and significantly affected by the curse of dimensionality, leading to reduced performance as the number of features increases.

Multinomial Naive Bayes is particularly good for categorizing text, as it analyzes the frequency of words under the assumption that all features in the data are unrelated.However, the independence assumption of Naive Bayes can sometimes lead to significant errors, especially in cases where the relationship between attribute values is strong, which is common in text data. Meaning that while it can tell which class is more likely, the actual probability values can be misleading.

We chose Multi-label Logistic Regression (MLR) because of its ability to manage several categories for each data point, a crucial feature for tackling complex classification problems where each instance may fit into multiple categories at once. This model uses logistic functions to calculate the likelihood of each category, making it a strong tool for handling datasets with multiple dimensions. The flexibility and effectiveness of Multi-label Logistic Regression are especially beneficial in situations where traditional models may not account for the relationships between different labels.

V. Model Evaluation

During our model evaluation phase, we utilized KMeans Clustering to better understand the data, calculating a Mean Squared Error (MSE) score for each category. The top_category_id returned the lowest MSE, indicating more accurate model performance, likely due to the smaller number of classes making predictions more straightforward. In contrast, the bottom_category_id recorded a higher MSE, attributable to its greater number of classes which complicates accurate classification. We then processed and fed data into models to classify four types of IDs: top_category_id, bottom_category_id, primary_color_id and secondary_color_id. This approach allowed us to refine our models to adapt effectively to the specific characteristics of each category, thereby enhancing overall prediction accuracy and understanding the complex relationships within the data. Table 1. shows the evaluation of the top_category_id of all the machine learning models.

Among the various models we tested, including the KNN Classifier, Multinomial Naive Bayes, and Multi-label Logistic Regression, it was the Multi-label Logistic Regression that stood out for its superior accuracy of 90%. This model was particularly effective at predicting the 'top_category_id,' which involves categorizing items into their broadest groups. The success of the Multi-label Logistic Regression highlights its ability to efficiently handle datasets where each item can be associated with more than one label simultaneously. Its performance shows that it can discern complex patterns and relationships within the data, making it highly suitable for tasks that require the classification of items into multiple categories at once.

TABLE I
Model Performance Metrics for top_category_id

Model Name	Accuracy	F1 Score	Recall
KNN			
Train	0.8089	0.8300	0.8089
Validation	0.7079	0.7099	0.7079
MultiNomialNB			
Train	0.8469	0.8479	0.8469
Validation	0.7456	0.7459	0.7456
MLR			
Train	0.9972	0.9972	0.9972
Validation	0.9030	0.9026	0.9030

VI. Limitations and Future Work

We initially attempted to create a single model to predict three different features at once. However, we faced challenges in accurately processing image data, which affected our ability to correctly predict the primary and secondary colors of products. To improve our results, we plan to refine our data sampling techniques to increase the model's accuracy and reduce discrepancies between training and validation results. Additionally, enhancing our image processing methods using image/encoded feature, such as predicting colors from central image pixels and aligning these with predefined color categories in our dataset, will be crucial. With more meticulous data cleaning and preparation, we aim to achieve more consistent and reliable predictions.

VII. Conclusion

In our study focused on improving the Etsy shopping experience using only textual data, we leveraged the CountVectorizer tool and batch processing to optimize memory storage and workflow efficiency. The Multi-label Logistic Regression model stood out, demonstrating exceptional precision and accuracy in predicting Etsy's 'top_category_id', bottom_category_id. This model was particularly effective due to its ability to manage complex datasets where items belong to multiple categories, capturing intricate patterns within the data effectively. Following closely, the model also showed good performance in predicting the 'bottom_category_id'. However, for the 'primary_color_id' and 'secondary_color_id', the accuracy and F1 scores were notably lower, indicating challenges in achieving similar levels of precision and accuracy for color-related predictions. Moving forward, we plan to refine our text processing techniques and explore additional methodologies to enhance model performance across all categories, particularly

focusing on improving color predictions to further enrich the Etsy user experience, allowing shoppers to more reliably find and purchase products that meet their specific preferences.

REFERENCES

[1] L. Donati, E. Iotti, G. Mordonini, and A. Prati, "Fashion Product Classification through Deep Learning and Computer Vision," *Applied Sciences*, vol. 9, no. 7, Art. no. 7, Jan. 2019, doi: 10.3390/app9071385.

[2] M. Y. Li, S. Kok, and L. Tan, "Don't Classify, Translate: Multi-Level E-Commerce Product Categorization Via Machine Translation." arXiv, Dec. 13, 2018. doi: 10.48550/arXiv.1812.05774.
.

[3] S. Shankar and I. Lin, "Applying Machine Learning to Product Categorization".

[4] R. C. Morales-Hernandez, J. G. Jaguey, and D. Becerra-Alonso, "A Comparison of Multi-Label Text Classification Models in Research Articles Labeled With Sustainable Development Goals," IEEE Access, vol. 10, pp. 123534–123548, 2022, doi: 10.1109/ACCESS.2022.3223094.

[5] D. Wang and H. Zhang, "Inverse-Category-Frequency Based Supervised Term Weighting Schemes for Text Categorization," *Computing Research Repository - CORR*, vol. 29, Dec. 2010.

[6] Anolytics, "E-commerce Machine Learning: Product Categorization," Anolytics. Accessed: Apr. 20, 2024. [Online]. Available: https://www.anolytics.ai/blog/e-commerce-machine-learning-product-classification-insight/

[7] Etsy Ireland - Shop for handmade, vintage, custom, and unique gifts for everyone, Etsy. Accessed: Apr. 20, 2024. [Online]. Available: http://www.etsy.com/