Product Categorization and Color Prediction using Machine Learning techniques

Ronit Loke #23265555
School of Computing, Dublin City University, Ireland
Email: ronit.loke2@mail.dcu.ie

***Abstract*—This study explores the application of machine learning algorithms for predicting product attributes in the e-commerce sector. Utilizing large datasets of images and textual descriptions from an e-commerce website, various models were developed and tested. For predicting the top category ID, logistic regression proved most effective, particularly with textual data. The bottom category ID was best predicted using a Support Vector Machine (SVM) classifier, enhanced by a feature extraction and selection pipeline. Color prediction was accomplished using a Convolutional Neural Network (CNN) leveraging pre-trained ImageNet weights. The results of these experiments underscore the potential of machine learning techniques in enhancing product categorization and color prediction in e-commerce.**

I. Introduction

In the ever-evolving digital marketplace, Etsy stands out as a global platform that uniquely blends creativity with commerce, hosting an impressive array of handcrafted goods, vintage treasures, and distinctive items tailored to personal tastes and special occasions. With nearly 100 million active listings from over 5 million creative sellers, Etsy not only fosters a thriving community of creative buyers and sellers but also serves as a rich dataset for advanced machine learning research. This paper focuses on leveraging such a dataset, comprised of a curated subset of products from Etsy's extensive catalog, to address a significant challenge in e-commerce: the automated prediction of product attributes. Specifically, this study aims to develop robust machine learning models that can accurately predict the top and bottom category IDs, as well as primary and secondary color IDs, of products based on their available information. By applying sophisticated algorithmic techniques and utilizing a training dataset that captures the diverse range of Etsy's offerings, this research seeks to enhance the personalization and efficiency of the online shopping experience, ultimately striving to maximize the F1 score for each class of the product attributes in question. This introduction sets the stage for a detailed exploration of predictive modeling techniques in an e-commerce context, emphasizing the importance of accurate attribute classification in enhancing user experience and operational efficiency.

II. RELATED WORK

Machine learning and Natural Language Processing (NLP) have evolved significantly over the years. From rule-based and traditional statistical methods to advanced deep learning approaches, the field has seen a substantial transformation towards robust, dynamic methods. In [1] Early systems, precise for specific tasks, lacked the flexibility required for complex, diverse datasets. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) offered a more scalable framework for text classification but required extensive domain knowledge and intensive feature engineering. The introduction of neural networks marked a significant leap in handling the subtleties of human language. In this study [5] Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their derivatives like LSTM and GRU automated feature extraction, learning intricate patterns in data, thereby outperforming traditional models, especially in understanding context and semantic relationships within text. The shift towards transfer learning, particularly through pre-trained models such as BERT, ELMo, and GPT, revolutionized NLP by enabling models trained on extensive corpora to be fine-tuned for specific tasks. This approach significantly enhanced performance even with smaller datasets, marking a milestone particularly in environments where labeled data are scarce. Among these [4], DistilBERT, a compact variant of BERT, offers an optimal balance between performance and computational efficiency, retaining most of BERT's predictive power with fewer parameters and faster training times, making it suitable for resource-constrained applications. Active learning, a subset of machine learning, strategically selects the most informative data points for labeling, thus reducing the need for large labeled datasets and enhancing model performance. The Query by Committee (QBC) strategy, which involves selecting data points that a committee of models most disagree on, effectively reduces model bias and variance, enhancing the learning rate and robustness of models. In image and text classification, active learning has been employed to annotate data selectively, significantly improving classification accuracy with fewer labeled examples compared to traditional supervised learning. The Techniques like [2] convolutional neural networks combined with active learning have shown promising results, particularly in handling high-dimensional data such as images and text. Despite the advantages, both active learning and deep learning models face challenges such as selection bias, high computational costs, and the need for continual retraining as new data is annotated. Innovations in computational efficiency and algorithmic development are critical to mitigating these issues, facilitating broader adoption of these techniques. The integration of deep learning [3] with Naïve Bayes principles and the continuous development

of hybrid models that combine the probabilistic strengths of traditional algorithms with neural approaches represent promising research directions. Such models aim to leverage the benefits of both paradigms to enhance performance in various applications like e-commerce sentiment analysis and real-time text analysis.

III. DATA SET

Our machine learning project utilizes a dataset provided by Etsy, comprising approximately 240,000 records in each of the training and test sets, formatted in Parquet files. The dataset includes detailed attributes of product listings such as product_id, title, description, and tags, alongside other categorical descriptors like type, room, craft type, recipient, and material. For our analysis, we focused on the product_id, title, description, and tags to train our models, as these fields contain rich textual content suitable for natural language processing, which is vital for our objective. The target outcomes for our model predictions are the top_category_id, bottom_category_id, and color_id, which are essential for improving product searchability and recommendation on the Etsy platform.

IV. DATA CLEANING

Data cleaning is an essential process in data analysis and machine learning to improve data quality, which significantly impacts the performance of predictive models. The primary objective of data cleaning is to address issues such as missing values, incorrect data formats, duplicate records, and outliers. Proper data cleaning ensures that datasets are accurate, consistent, and usable, facilitating better decision-making and more reliable analytical outcomes. Missing values posed a significant issue during the dataset analysis, as over 80% of the data in numerous categorical columns was missing. First I started with finding the missing values, that takes a dataset as input and performs several operations to identify and quantify missing data. It begins by converting common placeholders for missing data, such as empty strings, to a standard 'NaN' format recognizable by data processing libraries. The function then calculates the total count and percentage of missing values for each column and compiles these statistics into a DataFrame. This DataFrame is sorted by the count of missing values to help prioritize which columns require attention due to significant data absence.

The data is then processed by the second function, columns_dropped_containing_missing_values, which drops columns when the percentage of missing values surpasses a predetermined threshold—70% by default. This approach is used to eliminate columns that lack sufficient data to contribute meaningfully to analysis or predictive modeling, thereby streamlining the dataset for more efficient processing and potentially improving model accuracy. This methodical elimination of deficient data columns ensures that the remaining dataset is more robust and representative for analytical purposes.

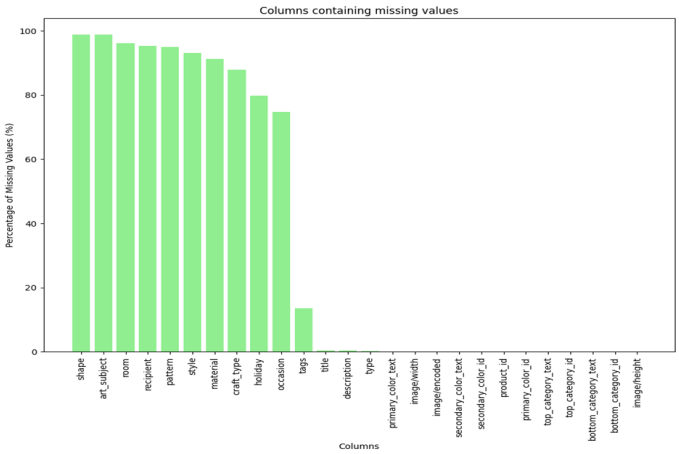

Figure 1. Visualization of Missing Data Analysis

V. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a fundamental step in understanding the underlying patterns and distributions within a dataset. The provided bar chart visualizes the distribution of product listings across various top-level categories on an e-commerce platform.

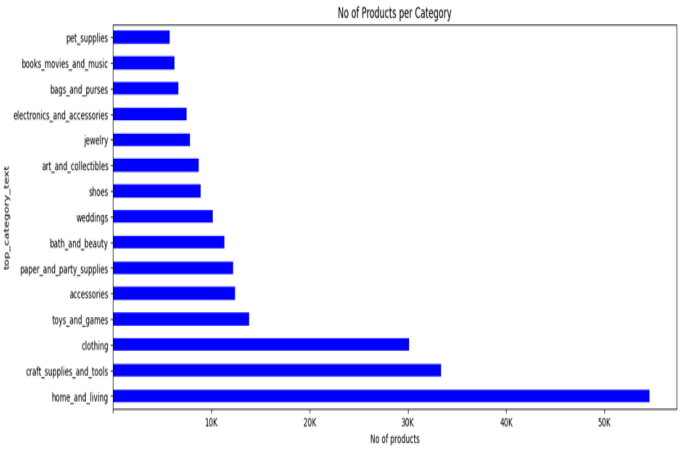

Figure 2. Visualization of no of products per category

The bar chart provides a clear depiction of product distribution across various categories within an e-commerce platform, essential for Exploratory Data Analysis (EDA). 'Home_and_living' emerges as the most populated category, significantly surpassing others like 'craft_supplies_and_tools' and 'clothing', which also show a strong presence. Conversely, niches such as 'pet_supplies', 'books_movies_and_music', and 'bags_and_purses' contain notably fewer listings, hinting at potential growth opportunities or specialized markets. This distribution suggests consumer demand patterns and could guide sellers on product placement and platform marketing strategies. The standardization in category naming points to high data quality, crucial for automated data processing and analysis. Overall, this visual EDA is a strategic asset for

platform stakeholders, providing insights into current trends and areas for potential expansion or increased competitive focus.

VI. DATA PREPROCESSING

Data preprocessing is crucial in machine learning to transform raw data into a structured format that algorithms can efficiently process. Techniques such as CountVectorizer and TF-IDF transform text into numerical representations, emphasizing words that are most indicative of content categories. The use of pipelines ensures consistent application of preprocessing steps and model training, preventing data leakage and errors. Feature selection through SelectFromModel reduces model complexity by discarding irrelevant features, enhancing both performance and interpretability. Normalization and LabelEncoder standardize numerical and categorical data respectively, ensuring uniformity and suitability for algorithmic processing, which is essential for achieving accurate and reliable model predictions.

A. CountVectorizer:

The CountVectorizer is a text transformation tool provided by scikit-learn that converts a collection of text documents to a matrix of token counts. This process is often referred to as "vectorization" of text data. It's used to transform text into a format that machine learning algorithms can understand, i.e., numerical data. Each column represents a unique word in the text corpus, and each row represents each document in the dataset, with values indicating the frequency of each word in the document.

B. TF-IDF (TfidfTransformer):

TF-IDF stands for Term Frequency-Inverse Document Frequency. It's a statistical measure used to evaluate how important a word is to a document in a collection or corpus. A word's relevance rises in direct proportion to how frequently it occurs in the document, however this is countered by the word's frequency in the corpus. While CountVectorizer only considers the frequency of words, TF-IDF also accounts for the importance of words. Common words that appear in many documents (e.g., "the", "is") are penalized. TfidfTransformer transforms a count matrix from CountVectorizer into a TF-IDF representation, making it more useful for machine learning models by highlighting the most relevant features.

C. Pipeline:

In scikit-learn, a Pipeline is used to assemble several steps that can be cross-validated together while setting different parameters. It applies a final estimator and a list of transforms one after the other. Intermediate steps of the pipeline must be transformers (i.e., must have fit and transform methods), and the final estimator only needs to have a fit method. Pipeline simplifies the process of writing and maintaining code for machine learning workflows. It ensures that the same sequence of steps is applied during training and prediction, avoiding common mistakes such as leaking data during data preprocessing.

D. SelectFromModel:

Is a feature selection technique that selects features based on the importance weights provided by any classifier or regressor that assigns weights to individual features (e.g., the coefficients in linear models). It is used to reduce the dimensionality of the input feature set, which can lead to reduced model complexity, shorter training times, and sometimes improvement in accuracy. In your case, LinearSVC is used within SelectFromModel to determine which features (words) contribute most to predicting the target variable and thus should be retained.

E. Normalization:

Normalization is a technique often used to scale numeric attributes in the dataset to a common scale without distorting differences in the ranges of values or losing information. In the context of image processing (as in the provided code), normalization typically involves adjusting pixel intensity values. In deep learning, especially in processing images, normalization helps in speeding up the learning process and leads to faster convergence. It typically involves scaling the pixel values to a range of 0 to 1 or normalizing pixel values based on the mean and standard deviation of the pixels.

F. LabelEncoder:

LabelEncoder in scikit-learn is a utility class to help normalize labels such that they contain only values between 0 and n_classes-1. It is used for transforming non-numerical labels to numerical labels. Many machine learning algorithms cannot handle categorical labels directly. They require the labels to be numeric. This is especially true for classification tasks where target variables are categorical. LabelEncoder transforms these labels to numeric form, making it easier to apply and evaluate machine learning models.

VII. METHODOLOGY

The project harnesses a diverse array of advanced machine learning techniques and models, each meticulously chosen and tailored to address specific classification tasks within the realm of e-commerce analytics. The primary objective of this research is to accurately predict various attributes of products listed on an online platform. These attributes encompass the Top Category ID, Bottom Category ID, and Colour ID, which are pivotal for enhancing the user experience, streamlining product searches, and optimizing inventory management.

A. Top Category ID

1) Logistic Regression:

In its most basic version, this statistical model models a binary dependent variable using a logistic function; in more advanced versions, it can also represent multinomial outcomes. This model was chosen for its efficiency and effectiveness in binary classification problems, which can be extended to multiclass classification under certain frameworks (such as one-vs-rest), making it suitable for predicting categorical top category IDs.

2) Random Forest:

In order to perform classification (and regression) using ensemble learning, Random Forest builds a large number of decision trees during training and outputs the class that represents the mean prediction (or mode of the classes) of each individual tree. Random Forest was used because of its ability to handle overfitting, which is common with decision trees, especially in complex datasets with many features. Its ensemble approach ensures that it captures a broad range of patterns in the data.

3) Decision Trees:

A non-parametric supervised learning technique for regression and classification is called a decision tree. The objective is to build a model that, by utilising basic decision rules deduced from the data features, predicts the value of a target variable. Decision Trees were included in the methodology for their interpretability and ease of use. They are very intuitive and their decisions can be visualized and understood by non-experts, which helps in understanding feature importance.

B. Bottom Category ID

1) Support Vector Machine (SVM):

SVM was chosen for its effectiveness in high-dimensional spaces and its ability to use a subset of training points in the decision function (support vectors), making it memory efficient. The SVM model, especially with linear kernels, is well-suited for binary classification tasks and was applied here to differentiate between the various bottom category IDs based on textual data.

2) GridSearchCV:

It is a method that performs a methodical search over a specified parameter values for an estimator. It uses cross-validation to evaluate each individual model, which ensures the model's performance is not just a fluke of one split of the data. This approach was utilized to fine-tune model parameters for the Naive Bayes classifier to optimize its performance. It systematically works through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance.

C. Color ID

1) CNN Model (Convolutional Neural Network):

CNNs are a type of deep neural network that are typically used for image analysis. Their layers convolve with a dot product, such as a multiplication. The CNN model was selected for the task of classifying images by color ID due to its state-of-the-art performance on image classification tasks. CNNs can automatically detect the important features without any human supervision, making them ideal for image-based datasets.

VIII. EVALUATION

For the evaluation section of your report, you will detail the performance of the models used for classifying the Top Category ID, Primary Colour ID, and Secondary Colour ID. Each model's evaluation will focus on the methodology used, performance metrics achieved, and insights gleaned from the results.

A. Top Category ID

For predicting the Top Category ID, a logistic regression model was employed within a machine learning pipeline that included text preprocessing with CountVectorizer and TfidfTransformer. This approach converted product descriptions into a numerical format suitable for machine learning, enhancing feature recognition by focusing on the most informative words for category classification. The model was tuned using GridSearchCV to optimize parameters such as the regularization strength (C) and the application of the inverse document frequency (IDF) transformation. The best performing model achieved an F1 score of 0.8423, which reflects a robust balance between precision and recall in the multi-class classification setting. The high F1 score suggests that the logistic regression model, coupled with TF-IDF weighting, effectively captured the relationships between the words in the product descriptions and their corresponding categories. This indicates a strong predictive performance, making this approach suitable for similar text-based classification tasks.

B. Bottom Category ID

In the evaluation of the Bottom Category ID, a Support Vector Machine (SVM) model integrated with a feature selection process using LinearSVC was utilized to optimize feature relevance via SelectFromModel. Text data was vectorized and transformed through CountVectorizer and TfidfTransformer, with batch processing applied to manage large datasets effectively. Parameter optimization through GridSearchCV selected 'clf_C': 1, 'feature_selection_estimator_C': 1 as optimal, resulting in an F1 score of 0.6081 on the validation set. This score, though moderate, revealed significant variability in category-specific performance, suggesting that some categories might benefit from further model refinement or additional feature engineering to enhance prediction accuracy.

C. Primary Colour ID

The prediction of the Primary Colour ID was approached using a convolutional neural network (CNN), designed to process image data. The model architecture included several convolutional layers and max-pooling layers to extract features from images, followed by fully connected layers to perform classification based on these features. The training process was monitored using a custom training loop with a learning rate scheduler to enhance convergence. The best model achieved a Test Accuracy of 0.4126 after 20 epochs, indicating the model's ability to correctly identify the primary colour from product images about 41% of the time. While the accuracy achieved may seem moderate, it highlights the challenges and complexities involved in image-based classification tasks, especially in distinguishing among various colours where subtle differences can significantly affect classification accuracy. Improvements might include using more complex models, increasing the dataset size, or employing more sophisticated image augmentation techniques to enhance model robustness.

D. Secondary Colour ID

Similar to the Primary Colour ID, the Secondary Colour ID classification was tackled using a CNN tailored for image data. The model used the same architecture and training strategy as the Primary Colour ID model, ensuring consistency in methodology across similar tasks. The secondary colour classification resulted in a Test Accuracy of 0.2252. The learning curves and validation accuracy indicated a stable training process but highlighted the difficulty of achieving high accuracy in this specific task. The lower accuracy compared to the primary colour model underscores the increased difficulty of accurately predicting secondary colours, which may be less prominent in images. Future work could explore multi-label classification techniques, deeper networks, or even domain-specific enhancements in image preprocessing to improve performance.

IX. Conclusions

This research explored the utility of machine learning techniques in predicting key product attributes on an e-commerce platform, achieving significant outcomes. Logistic Regression effectively predicted Top Category ID, achieving an impressive F1 score, highlighting the potential of combining traditional models with advanced text preprocessing. The SVM model used for Bottom Category ID displayed moderate effectiveness, pointing to the nuanced challenges of fine-grained categorization. Conversely, the use of CNNs for color identification showed the complexity involved in processing visual data, with results suggesting a need for further refinement. Overall, the study successfully demonstrated how different machine learning approaches could be tailored to enhance the accuracy and efficiency of product categorization and color prediction, fundamentally improving the e-commerce browsing experience.

X. Future work

Future research can expand upon the findings of this study by exploring several promising directions. Enhancing model performance through advanced algorithms, such as deep learning and ensemble methods, could further improve accuracy, particularly for categories and colors that proved challenging. Investigating more sophisticated image augmentation techniques would likely aid CNNs in better generalizing from the training data to real-world applications. Incorporating multi-label classification approaches could more accurately reflect the complexity of real-world products. Additionally, expanding the dataset and integrating real-time adaptive learning mechanisms could help maintain model relevance over time. Finally, leveraging user feedback to refine model predictions could personalize user experiences, fostering higher engagement and satisfaction. Such advancements will drive forward the capabilities of machine learning in transforming e-commerce platforms into more user-centric and intelligent marketplaces.

References

[1] M. Lekha K. Sakthi and P. A. Manesha. Exploring Active Machine Learning Techniques to Boost Classification Accuracy in Image and Text Models. pages 16876–16892, 2023.

[2] V. Gaur T. Singla and D. K. Misra. Comparison between multinomial naive bayes and multi-layer perceptron for product review in real time. pages 2111–2116, 2022.

[3] K. Taneja and J. Vashishtha. Comparison of Transfer Learning and Traditional Machine Learning Approach for Text Classification. *IEEE Access*, pages 94040–94052, 2022.

[4] Venkatesh and K. V. Ranjitha. Classification and Optimization Scheme for Text Data using Machine Learning Naïve Bayes Classifier. pages 235–238, 2018.

[5] Y. Zhang and X. Zheng. Development of image processing based on deep learning algorithm. 2022.