Modelling Product Category Hierarchies with Textual Features: An Etsy Case Study

Naveen Singh, *ML-Labs*, School of Computing, Dublin City University, Ireland

Abstract—This project tackles the problem of multi-class classification of e-commerce product listings on the Etsy platform, aiming to predict both high-level (top category id) and granular (bottom category id) categories using product metadata. A basis pipeline incorporated TF-IDF features together with Logistic Regression models during development for both tasks. Text fields such as title, description, and tags were combined and vectorized for input. The goal was to maximize F1 score on a hidden test dataset. The following document provides details about processing choices and model selection and evaluation criteria along with upcoming developments for future improvement.

I. INTRODUCTION

ETSY is a large-scale e-commerce platform featuring diverse product listings with structured and unstructured metadata. Classes of product listings using an accurate hierarchical structure leads to performance enhancements in search capabilities and individual customer recommendations as well as data analytics potential. This project focuses on building machine learning models to predict:

- top category id (15 classes)
- bottom category id (2600+ classes)

Natural language processing techniques help handle product data because of its primarily textual nature to build a productive classification system. The goal involves achieving maximum statistics performance on the macro and weighted F1 metrics across the two classification tasks.

II. RELATED WORD

The field of product categorization and multi-class text classification evaluates classical together with modern deep learning methodology. The earlier systems applied Naive Bayes and Support Vector Machines models as well as bagof-words features. The combination of TF-IDF with Logistic Regression forms a stable technique for product metadata analysis because of its strength in handling rare features found in these sparse inputs.

Research teams have used BERT and DistilBERT transformers to enhance classification performance for short texts such as search queries and product titles. The research by Peeters et al. (2020) applied BERT which underwent finetuning to complete product matching tasks and produced excellent F1 scores exceeding 0.9 on expansive datasets. Two.

variety of domain-specific variants called eComBERT and Sentence-BERT have become specialized for matching and categorization applications.

Transformers operate with great power but demand substantial processing capabilities. The classical model (just like the report uses) delivers prompt training while maintaining explanatory abilities together with reliable performance on structured datasets which benefit from properly designed features like TF-IDF vectors.

III. EXPLORATORY DATA ANALYSIS

- Total rows in training dataset: 229,624
- Unique product_id: 229,624
- Number of top categories: 15
- Number of bottom categories: 2,609

Top Category Distribution:

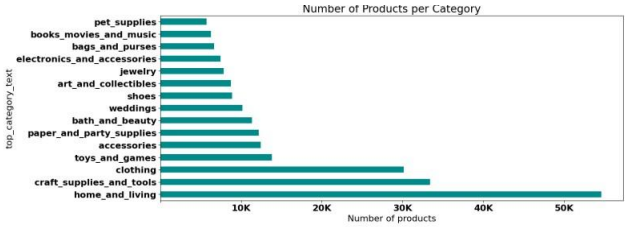

Fig. 1. Number of Products per Category (top category)

- home_and_living is the most common top category (50K items)
- Followed by craft_supplies_and_tools, clothing, and toys_and_games

Bottom Category Distribution:

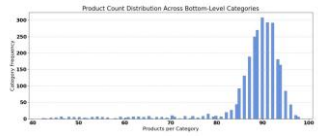

Fig. 2. Product Count Distribution Across Bottom-Level Categories

- M Most bottom-level categories have fewer than 100 products
- Sharp imbalance observed; long-tail distribution challenges model generalization

Missing data analysis

The style material craft_type and occasion fields show more than 70% missing values that introduce excessive noise to the data unless extensive imputation approaches suitable for text-based models are used.

The model excluded structured fields including room and both color text fields as well as recipient and room information during this round of development because they contained full or partial completion. The structured categorical features need encoding methods such as one-hot encoding or embedding representations to become part of the integration framework. The addition of these fields would require a multimodal modeling strategy to unite structured and textual data but surpassed the model development capabilities of TF-IDF + Logistic Regression.

The model underwent development as a straightforward text-based classification system by concentrating exclusively on unstructured text data contained in the title, description and tags fields. The sparse additional fields might have reduced predictive model performance by providing inconsistent or uninformative information. The selection of title field along with description and tags represented the main textual elements since these key fields contained less than 1% missing data points while having central functions in product descriptions. The textual data within these fields maintains the most accurate description along with meaning regarding each product so they suit the text-based classification approach best.

IV. METHODOLOGY

Data Cleaning and Feature Engineering

For the textual feature engineering, title, description, and tags were concatenated into a single text field referred to as combine_text. All missing values in these fields were replaced with empty strings to ensure consistency. The combined text was then converted to lowercase to standardize input data. Tags were formatted into space-separated strings in cases where they were stored as lists, ensuring uniformity across all records. No stemming or lemmatization was applied, following findings in related work that suggest minimal performance gains from these processes when using TF-IDF representations.

Model Pipeline

The model pipeline was designed to transform the clean text data into numerical features and classify products into categories. This pipeline consisted of two main stages:

- The TF-IDF vectorization process eliminated English stop words while analyzing unigram and bigram word sequences (n-grams of length one and two) and extracted 15,000 of the most frequent words in the dataset. The dimensionality of vocabulary received a restriction which supported both memory optimization and processing performance enhancement.
- The model used Logistic Regression as its classification method with distinct configurations for maximizing performance in different tasks. To handle the top_category_id model the classifier used balanced class weights together with the liblinear solver because this setup works well with smaller multi-class data. The saga solver was chosen for the bottom-level category model (bottom_category_id) because it scales effectively between its many classes despite its unbalanced distribution. The bottom-level model received a regularization strength parameter (C=1.2) which enabled slightly less rigorous regularization thus giving the model the capacity to accommodate more complex patterns in the data. Each model was set with 1000 maximum iterations as its stopping criterion to reach convergence.

Evaluation Strategy

The evaluation process required a 90/10 stratified split of dataset subsets according to the bottom_category_text. The stratification procedure preserved the category proportions between both subsets since this method was especially necessary given how numerous low-level classes skewed the data distribution.

The project ran two independent models which operated separately from one another.

- The model developed predictions for the top-level categories through its output known as top_category_id.
- The second analytic approach concentrated on predicting the detailed bottom-level categories which correspond to bottom_category_id.

The assessment of both models happened through macro and weighted F1 score evaluations. Because it maintains equivalent evaluation of all classes no matter their frequency the macro F1 score functions well when assessing predictions of underrepresented categories. Each class in the weighted F1 score receives weights that are proportional to their support level expressed through true instances.

V. RESULTS

The model designed to predict the top-level categories exhibited robust performance. It achieved a F1 score of around 0.86. These results indicate that the model performed consistently across the different top-level categories and was able to handle both frequent and less frequent categories well.

The bottom-level category classification, as anticipated, proved to be more challenging. The model attained F1 score of about 0.42.This disparity between top- and bottom-level category performance reflects the difficulty posed by the large number of bottom categories and their highly imbalanced distribution. While the model succeeded in capturing general patterns among the more frequent bottom categories, it struggled with the long tail of rare classes. The following table summarizes the F1 scores for both classification tasks:

Classification Task	F1 Score	Precision	Recall
Top Category ID	0.8603	0.8617	0.8606
Bottom Category ID	0.5907	0.6212	0.6033

To further support these quantitative findings, **t-SNE visualizations** of the TF-IDF embeddings were generated for both classification tasks. These visualizations offer a qualitative view of how the model separates different categories in lower-dimensional space.

For the **top categories**, distinct clustering patterns were observed for several categories like *bath_and_beauty* and *home_and_living*, reflecting clearer boundaries in their textual descriptions. In contrast, the **bottom categories** exhibited more overlapping clusters, visually confirming the greater complexity and class imbalance that challenge the model at this level.

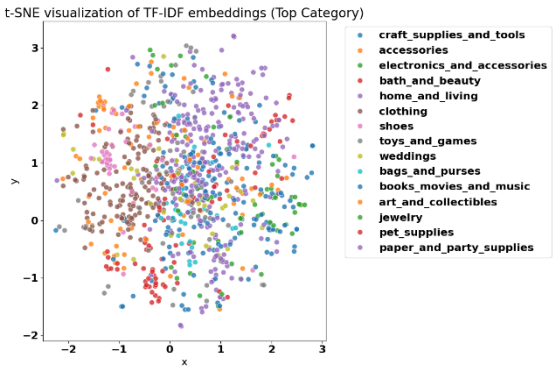

Figure 3: t-SNE visualization of TF-IDF embeddings (Top Categories)

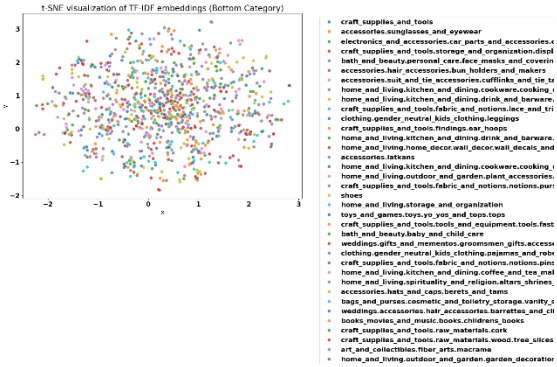

Figure 4: t-SNE visualization of TF-IDF embeddings (Bottom Categories)

VI. SUBMISSION AND FORMAT

Predictions for the test dataset were generated for both classification tasks and formatted according to the assignment requirements. The final output file included the product_id along with the corresponding predictions for top_category_id and bottom_category_id. This output was saved in the Parquet file format, ensuring both compact storage and efficient data retrieval.

VII. CONCLUSION AND FUTURE WORK

The research investigated traditional machine learning methods for dealing with complex hierarchical product classification tasks on Etsy. Text-based product listing features used in the TF-IDF and Logistic Regression pipeline delivered both computational efficiency and interpretability as a solution framework. Predictive success at top-category levels showed that basic models effectively perform when analyzing structured text data.

The classification issues at the bottom categories stem from the severe class imbalance along with the extreme taxonomical fineness. The large difference in performance between top and bottom levels underscores the necessity to use advanced models which can effectively address the intricate product taxonomy system. The baseline serves as an operational starting point to establish a reference standard that enables performance evaluation with advanced methods.

Upcoming research will emphasize the capability of the model to perform generalized classifications of different bottom-level product categories. The addition of transformer-based systems would enhance contextual embedding quality and applying hierarchical categorization techniques would strengthen structural category relationships. Specialized loss functions and data augmentation approaches will serve as both main priorities for handling class imbalance. The model's overall effectiveness can be enhanced by adding structured fields together with advanced error evaluation strategies.

Future Work:

- Incorporating transformer-based models like DistilBERT or RoBERTa could enhance performance, particularly for bottom-level categories by leveraging contextualized embeddings.
- Employing hierarchical classification frameworks that utilize predictions from top_category_id to inform bottom_category_id could improve accuracy.
- Addressing class imbalance using techniques like focal loss, SMOTE, or data augmentation strategies may help to better classify underrepresented categories.
- Exploring multimodal approaches by integrating structured fields (e.g., material, occasion) alongside textual features can further enrich the feature space.
- Conducting error analysis and visualizing confusion matrices to refine the classification boundaries and improve performance.

VIII. REFERENCES

[1] R. Peeters, C. Bizer, and G. Glavaš, "Intermediate training of BERT for product matching," arXiv preprint arXiv:2004.13654, 2020.

[2] J. Tracz et al., "BERT-based similarity learning for product matching," arXiv preprint arXiv:2010.03585, 2020.

[3] T. Zhang, F. Damerau, and D. Johnson, "Text categorization using support vector machine and

term frequency-inverse document frequency," in Proc. Int. Conf. on Machine Learning (ICML), 2003.