Predictive Analytics for Product Clustering in an E-Commerce Dataset

Abstract

Online marketplaces often aggregate product listings from multiple merchants, leading to duplicated or inconsistently described products. This project applies predictive analytics to an e-commerce product dataset in order to automatically assign product listings to their correct product clusters using textual information from product titles. Exploratory Data Analysis (EDA), business problem formulation, predictive task definition, and the CRISP-DM methodology were employed. A supervised multi-class text classification model using TF-IDF features and Multinomial Naive Bayes was implemented in Python. Results show that, despite severe class imbalance and a large number of clusters, product titles contain meaningful information for automated product matching. The project demonstrates the feasibility of machine learning for improving catalog quality in e-commerce systems.

I. Dataset Description and Exploratory Data Analysis

Dataset Overview

The dataset used in this project is derived from the LSApp repository and contains product listings collected from multiple merchants. Each row represents a single product listing.

The dataset consists of **35,311 rows** and **7 columns**, with no missing values and no duplicate rows.

Attribute Description

Column	Description
Product ID	Unique identifier for each product entry
Product Title	Product name/title containing brand and specifications
Merchant ID	Identifier for the seller
Cluster ID	Identifier grouping equivalent products
Cluster Label	Canonical product name
Category ID	Numerical category identifier
Category Label	Descriptive category name

Basic Statistics

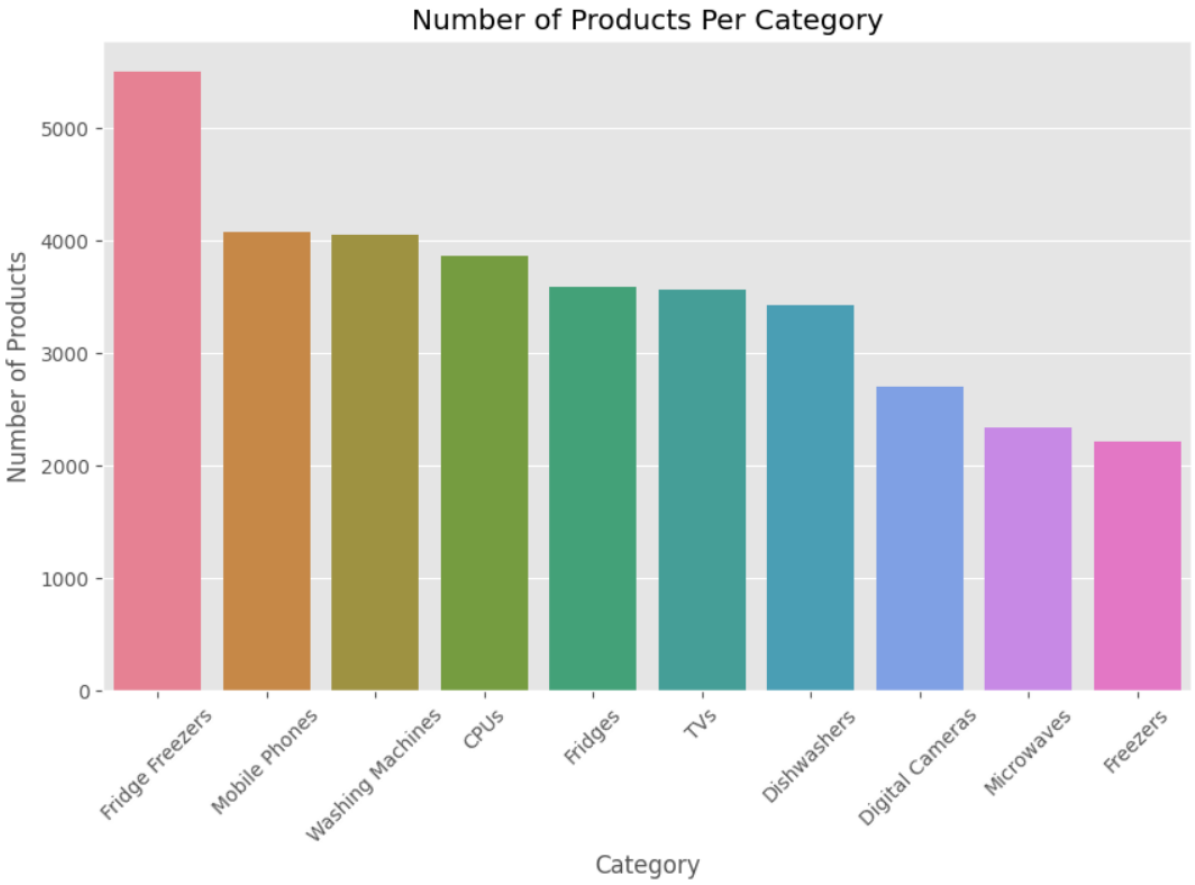
- Number of rows: 35,311
- Number of columns: 7
- Missing values: None
- Duplicate rows: None

These statistics indicate a clean dataset suitable for machine learning.

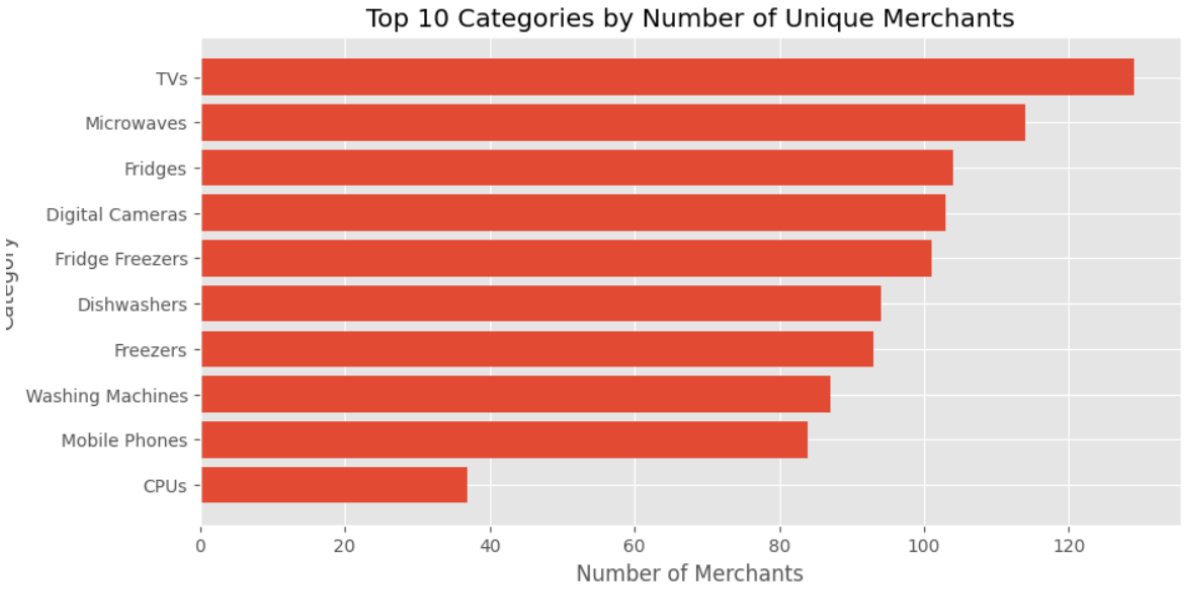
Exploratory Visualizations

Several visualizations were produced to better understand the data:

1. **Category Distribution**
   This chart shows that electronics and home appliances dominate the dataset, with categories such as Fridges-Freezers, Mobile Phones, TVs, and Washing Machines appearing most frequently.



2. **Merchants per Category**
   This visualization reveals that certain categories attract many merchants, indicating high competition. TVs, Microwaves and major appliances show the highest merchant participation.



3. **Products per Cluster**
   Analysis of cluster sizes shows that some clusters contain many listings, confirming that identical products are offered by multiple merchants. At the same time, many clusters are small, highlighting strong class imbalance.

4. **Word Cloud of Product Titles**
   The word cloud highlights common terms such as brand names and specifications. The wide variation in wording demonstrates inconsistent naming conventions across merchants.

EDA Summary

Exploratory analysis reveals:

- Strong category imbalance.
- Significant duplication of products across merchants.
- Highly variable product title formats.
- Severe imbalance in cluster sizes.

These findings motivate the use of machine learning to automatically group product listings and improve catalog consistency.

II. Business Problem Definition Using Issue Trees

Business Context

The dataset represents an e-commerce aggregation environment where multiple merchants list similar or identical products. Each merchant uses their own naming conventions, resulting in inconsistent product titles and duplicated catalog entries.

Such inconsistencies negatively affect:

- Search accuracy
- Price comparison
- Recommendation systems
- Customer trust

Business Problem Statement

How can duplicate or near-duplicate product listings across merchants be automatically identified and grouped in order to improve product catalog accuracy and enhance user experience?

Issue Tree Analysis

Main Problem:
Poor product catalog accuracy caused by duplicate and inconsistent listings.

Root Causes

Merchant Data Issues

- Unstructured product titles
- Inconsistent brand naming
- Missing specifications

Clustering Algorithm Limitations

- Inaccurate similarity thresholds
- Lack of token normalization
- Limited text processing

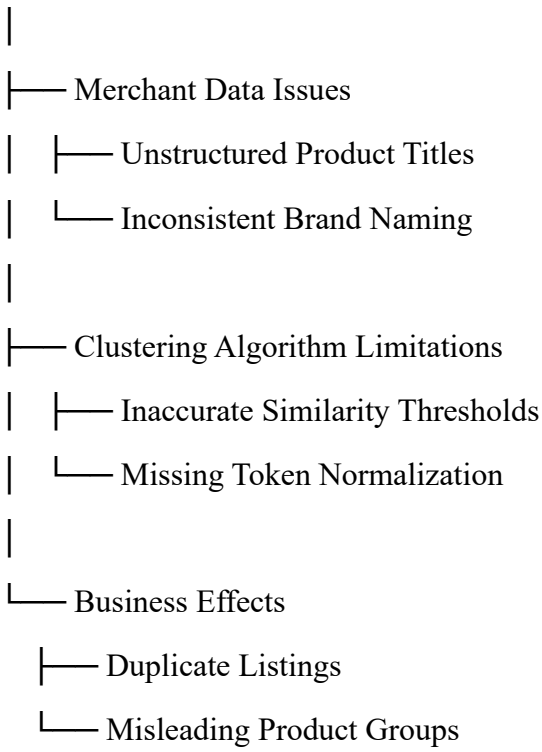
Business Effects

- Duplicate product displays
- Misleading price comparisons

- Reduced customer trust
- Lower conversion rates

Textual representation of the issue tree:

Improve Catalog Accuracy



├── Merchant Data Issues

│ ├── Unstructured Product Titles

│ └── Inconsistent Brand Naming

├── Clustering Algorithm Limitations

│ ├── Inaccurate Similarity Thresholds

│ └── Missing Token Normalization

└── Business Effects

├── Duplicate Listings

└── Misleading Product Groups

Business Objective

Develop a machine learning model that automatically assigns new product listings to the correct product cluster based on the product title.

III. Predictive Task Formulation

Objective

Predict the correct **Cluster ID** for a new product listing using its **Product Title**.

Inputs and Target

- Input feature: Product Title (text)
- Target variable: Cluster ID

Type of Task

- Supervised learning
- Multi-class text classification

Goal

When a merchant submits a new listing, the system predicts the most appropriate existing cluster, enabling automatic duplicate detection and improved catalog organization.

IV. Data Mining Methodology (CRISP-DM)

The CRISP-DM framework was used to guide the project.

1. Business Understanding

Define the need to reduce duplicate listings and improve catalog quality.

2. Data Understanding

Explore product categories, cluster distributions, merchant participation, and product title structure using EDA.

3. Data Preparation

- Convert text to lowercase
- Remove stopwords
- Apply TF-IDF vectorization
- Split data into training (80%) and testing (20%)

4. Modeling

A Multinomial Naive Bayes classifier was selected as a baseline model due to its efficiency and suitability for text data.

5. Evaluation

Performance was evaluated using accuracy and weighted F1-score.

6. Deployment

In a real system, the trained model would be integrated into the merchant upload pipeline to automatically assign clusters to new listings.

V. Implementation and Predictive Analysis

Tools

Python was used with the following libraries:

- pandas
- scikit-learn
- matplotlib
- wordcloud

Modeling Pipeline

1. Load dataset from TSV file
2. Select Product Title as input and Cluster ID as target
3. Split data into training and testing sets
4. Apply TF-IDF vectorization
5. Train Multinomial Naive Bayes classifier
6. Evaluate on test data

Evaluation Results

- **Accuracy:** 11.28%
- **Weighted F1-score:** 9.25%

Interpretation

Although these values appear low, the task involves classification across a very large number of clusters, many of which contain only a few samples. Under such conditions, random assignment would yield near-zero accuracy. Therefore, the results demonstrate that product titles contain meaningful information for predicting product clusters.

The weighted F1-score reflects severe class imbalance, where many clusters have very limited representation. Despite these challenges, the baseline model confirms the feasibility of automated product clustering using textual features.

Performance could be improved by:

- Using Support Vector Machines or transformer-based embeddings
- Removing extremely rare clusters
- Applying hierarchical classification
- Incorporating additional features such as category or brand