

MathCo

Report On

“Product Classification Using Text Data”

Submitted By

Harsh Adhikari

Contents

|  |
| --- |
|  **Introduction** |
|  **Data Overview** |
|  **Model Training** |
|  **Performance Metrics** |
|  **Confusion Matrix** |
|  **Difficult-to-Classify Cases** |
|  **Insights and Recommendations** |
|  **Conclusion** |

**Introduction**

This report analyzes the performance of a machine learning model designed to classify fashion products into their respective subcategories based on product descriptions. Utilizing a dataset of 5,000 observations, the model employs Logistic Regression alongside text pre-processing techniques, including normalization and lemmatization, to enhance accuracy.

The primary objective is to evaluate the model's performance through various metrics, such as accuracy, precision, recall, and F1-score. Additionally, the report identifies difficult-to-classify cases, providing insights into potential improvements and recommendations for optimizing the model for better classification outcomes in the dynamic fashion industry.

**Data Overview**

The dataset utilized for this project comprises 5000 observations of various fashion products. Each observation includes several key features that provide detailed insights . The primary features of the dataset are as follows:

* **Category:** This feature represents the broad classification of the product, such as "Accessories," "Suits," or "Footwear."
* **SubCategory:** A more specific categorization within the main category, such as "Bags," "Socks," "Tuxedos," or "Dresses."
* **ProductName:** This field contains the name of the product, often including brand names and key characteristics.
* **Description:** A detailed narrative that outlines the product's features, design elements, materials, and usage context.
* **Brand Tone:** It includes values such as "casual," "formal," "luxury," and "playful," which can influence consumer perception and preference.
* **Keywords:** Key phrases extracted from the product description that summarize essential aspects or features of the product.
* **Seasonality:** This feature indicates whether a product is part of a specific seasonal collection, such as "Winter," "Summer," "Spring," or "Autumn."

**Model Training**

The model training process involved splitting the dataset into training and testing subsets, allocating **80% for training and 20% for testing**. Text pre-processing techniques were applied to the product descriptions, including lowercasing, punctuation removal, stopword elimination, and lemmatization to enhance the quality of the input data.

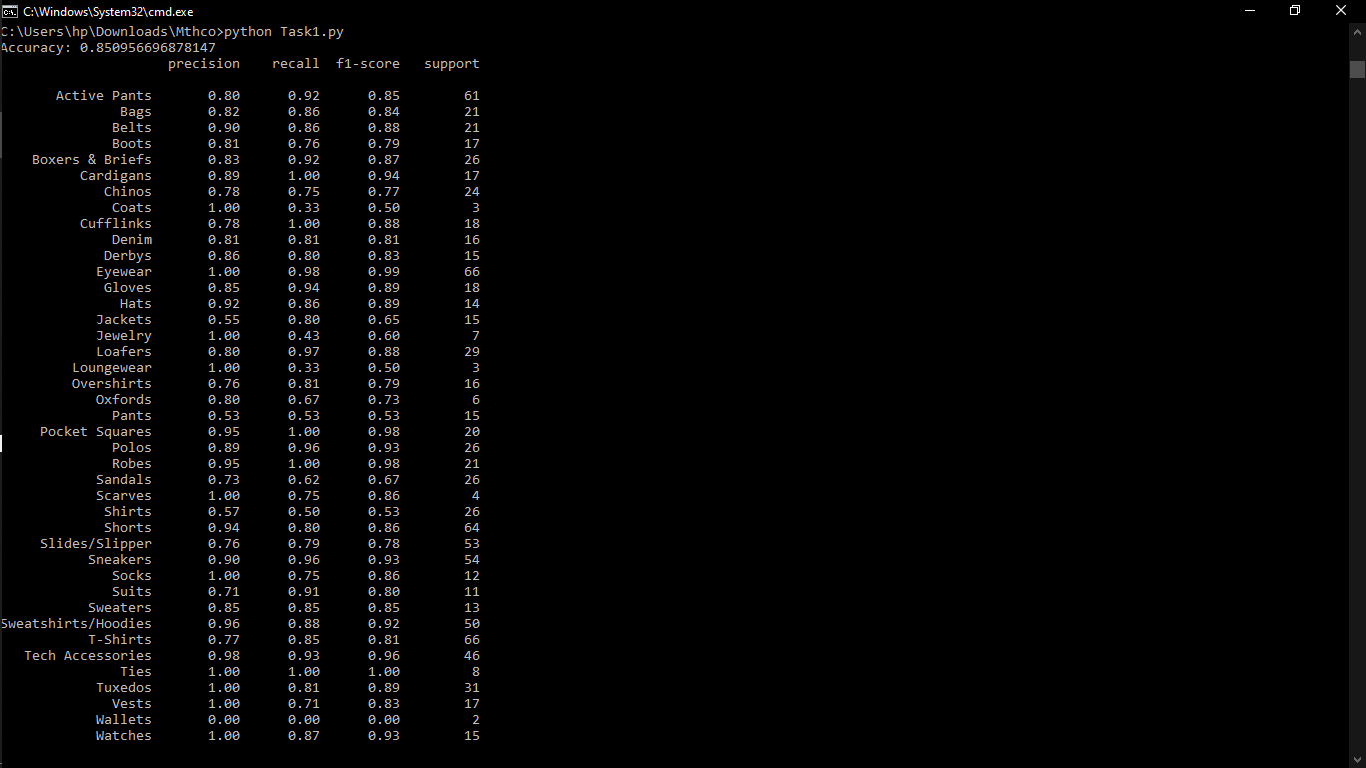
The Logistic Regression algorithm was employed to classify the products into their respective subcategories based on the cleaned descriptions.

After training, the model's performance was evaluated using various metrics such as accuracy, precision, recall, and F1-score, alongside cross-validation to ensure robustness and generalizability in its classification capabilities.

**Performance Metrics**

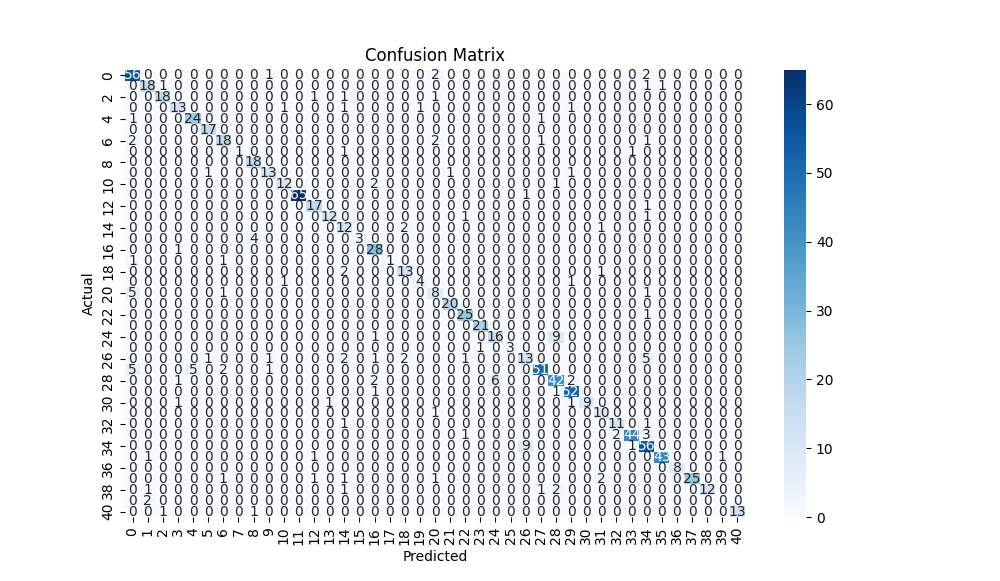
To evaluate the effectiveness of the machine learning model, several performance metrics were utilized, providing insights into its classification capabilities. Key metrics included:

* **Accuracy:** This measures the proportion of correctly predicted instances out of the total instances in the test set. For instance, an accuracy of 85% indicates that 85% of the predictions were correct.
* **Precision:** This metric assesses the accuracy of the positive predictions. It is calculated as the ratio of true positive predictions to the total predicted positives, indicating how many of the predicted subcategories were correct.
* **Recall:** Also known as sensitivity, recall measures the model's ability to identify all relevant instances. It is the ratio of true positive predictions to the actual positives, showing how effectively the model captures all instances of a specific subcategory.
* **F1-Score:** This metric is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when dealing with imbalanced classes, as it emphasizes both false positives and false negatives.
* **Classification Report:** A detailed report that summarizes the precision, recall, and F1-score for each subcategory, allowing for a nuanced understanding of the model's performance across different classes.



**Confusion Matrix**

The confusion matrix is a valuable tool for visualizing the performance of the classification model. In our implementation, it was generated using the confusion\_matrix function from Scikit-learn. This matrix compares the actual subcategory labels with the predicted labels, displaying true positives, false positives, true negatives, and false negatives. Each row of the matrix represents instances of the actual subcategory, while each column represents predicted subcategories. By plotting the confusion matrix using Seaborn, we could easily identify misclassifications and assess which subcategories the model struggled with most, thereby providing insights into areas for improvement.



**Difficult-to-Classify Cases**

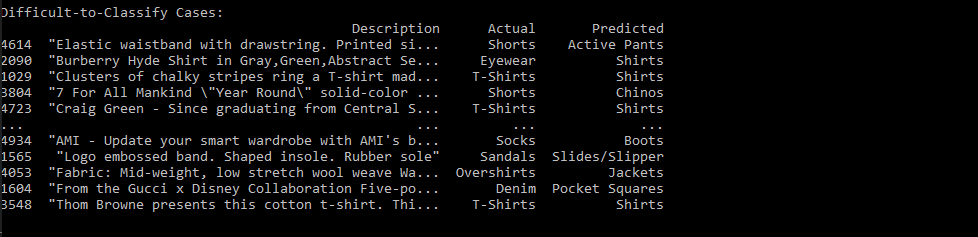
The analysis of difficult-to-classify cases reveals several insights into the challenges faced by the classification model. The following points summarize the key findings from the identified misclassifications:

1. **Ambiguity in Descriptions:** Many product descriptions contain vague or overlapping features that lead to confusion. For instance, the description of "Elastic waistband with drawstring. Printed stripes on soft cotton" was misclassified as "Active Pants" instead of "Shorts." This suggests that certain keywords or phrases are not sufficiently distinctive to accurately inform the model's predictions.
2. **Similar Product Attributes:** Products within similar categories often share common attributes, which can lead to misclassifications. The case of "Clusters of chalky stripes ring a T-shirt made from cotton" was incorrectly labeled as "Shirts" instead of "T-Shirts." This indicates that the model may struggle to differentiate between closely related subcategories, particularly when the descriptions do not clearly highlight unique features.
3. **Brand and Style Influence:** Descriptions that reference specific brands or styles may not provide enough contexts for accurate classification. For example, "Burberry Hyde Shirt in Gray, Green, and Abstract Series" was predicted as "Shirts" when it was actually classified as "Eyewear." This misclassification could arise from the model focusing more on the descriptive elements rather than recognizing the brand's typical category.

**Difficult-to-Classify Products**

The following products were identified as difficult to classify, highlighting the need for further refinement in model training:

* **Shorts:** Misclassified as "Active Pants" when described as having an elastic waistband and printed stripes.
* **Eyewear:** Misclassified from a description referencing a specific brand, "Burberry Hyde Shirt," which may not clearly indicate its category.
* **T-Shirts:** Misclassified as "Shirts" in several instances, indicating overlapping features between the two subcategories.
* **Socks:** Misclassified as "Boots" based on a description that likely contained attributes common to both types.
* **Over shirts:** Misclassified as "Jackets," suggesting that the model had difficulty distinguishing between these two related categories.



**Conclusion**

The Logistic Regression model demonstrates a satisfactory performance in classifying fashion products based on descriptions. However, there remains potential for improvement, especially regarding cases that are difficult to classify. The model's and dataset’s robustness and reliability can be enhanced.