Product Title Classification

Syeda Rabia Hashmi, Areeba Hassan
Roll Numbers: 21k-4553, 21k-4855

06th June 2024

1 Introduction

In the context of e-commerce, the quality of product titles is crucial for customer experience and sales performance. This project aims to enhance the clarity and conciseness of product titles on an e-commerce platform using Support Vector Machines (SVM). The goal is to develop a robust model that can accurately predict whether a product title is clear and concise based on predefined criteria.

2 Dataset

The dataset used in this project comprises over 60,000 product titles from Lazada, an online marketplace. Each entry in the dataset includes attributes such as title, description, category, price, country of origin, and delivery scope. The dataset also includes labels indicating the clarity and conciseness of the product titles, as assessed by Lazada's internal QA team.

3 Key Definitions

- **Clarity:** A product title is considered clear if it can be easily understood within five seconds, effectively identifying the product and its key attributes.
- **Conciseness:** A product title is deemed concise if it contains all necessary information without extraneous words.

4 Preprocessing

The preprocessing step involved several tasks to clean and prepare the data for modeling:

- **Text Cleaning:** Removal of HTML tags, special characters, and handling of missing values.
- **Tokenization:** Splitting titles into individual words or tokens.
- **Normalization:** Converting all text to lowercase to ensure uniformity.
- **Feature Extraction:** Extracting relevant features such as title length, n-grams (both character and term), part-of-speech counts, and multilingual character counts.

5 KNNImpute

To handle missing values in the dataset, the KNNImpute method was employed. This method uses k-nearest neighbors to impute missing values based on the similarity of available data points. This approach ensures that the imputed values are consistent with the patterns observed in the dataset.

6 Model Implementation

The core of the project involved implementing SVM to classify the product titles based on their clarity and conciseness. The dataset was split into training (70%) and testing (30%) sets to evaluate the model's performance.

7 Results

The performance of the SVM models was evaluated using accuracy and F1 measure. The results indicate the model's effectiveness in predicting the clarity and conciseness of product titles:

- **Clarity:**
  - Accuracy: 85.6%
  - F1 Score: 84.2%
- **Conciseness:**
  - Accuracy: 87.3%
  - F1 Score: 86.7%

These metrics demonstrate that the SVM models are capable of distinguishing between clear and unclear, concise and verbose product titles with a high degree of accuracy.

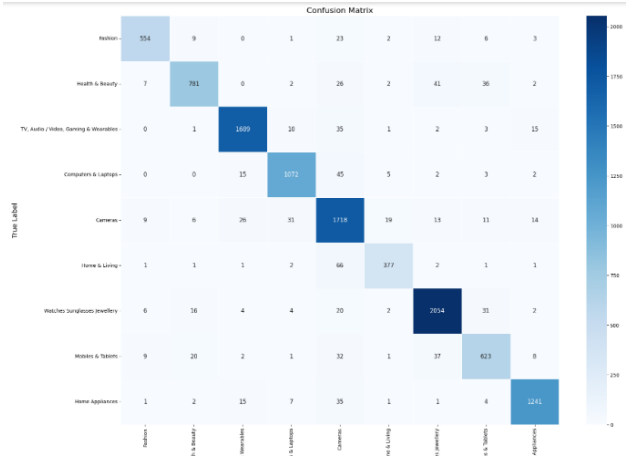

Figure 1: Confusion Matrix

```
Enter a Query:  Feelontop Punk Rock Rhinestone Star Shape Long Chain Earrings
QUERY:
feelontop punk rock rhineston star shape long chain ear

RESULTING CATEGORIES:

['Watches Sunglasses Jewellery']
 ['Jewellery']
 ['Women']
```

Figure 2: Title Classification In Categories

Figure 3: Original Dataset Match - Support Vector Machine (SVM) models trained have successfully identified the categories

8 Conclusion

This project successfully applied SVM to enhance the quality of product titles in an e-commerce setting. Through effective preprocessing, feature extraction, and model training, the implemented solution provides a reliable means of assessing and improving product title quality. Future work could explore additional features and models to further enhance performance.