Comparación de Modelos para la Clasificación de Artículos de Comercio Electrónico Basada en el Texto de la Descripción

Luis Felipe Rangel

Febrero 2025

1. Introducción

En el presente reporte se analizan las siguientes ingenierías de características:

1. **Vectores de recuento:** Count Vector por su nombre en Inglés, es una notación matricial del conjunto de datos en la que cada fila representa un documento del texto, cada columna representa un término del texto y cada celda representa el recuento de frecuencia de un término particular en un documento particular.

2. **Vectores TF-IDF:** La puntuación TF-IDF representa la importancia relativa de un término en el documento y en todo el conjunto de texto. La puntuación TF-IDF está compuesta por dos términos: el primero calcula la frecuencia normalizada del término, Term Frequency (TF) en Inglés, el segundo término es la frecuencia inversa del documento, Inverse Document Frequency (IDF) en Inglés, calculada como el logaritmo del número de documentos del conjunto de texto dividido por el número de documentos en los que aparece el término específico.

    - **TF-IDF a Nivel Palabra:** Matriz que representa las puntuaciones TF-IDF de cada término en diferentes documentos.
    - **TF-IDF a Nivel N-grama:** Los N-gramas son una combinación de N términos juntos.
    - **TF-IDF a Nivel Caracter:** Matriz que representa las puntuaciones TF-IDF de los N-gramas de nivel de caracter en el conjunto de texto.

Así mismo, se emplean los siguientes modelos de aprendizaje automático como clasificadores:

1. **Clasificador Bayesiano Ingenuo:** es una técnica de clasificación basada en el teorema de Bayes y que asume la independencia entre los predictores. Un clasificador Bayes ingenuo supone que la presencia de una característica

particular en una clase no está relacionada con la presencia de ninguna otra característica.

2. **Clasificador Lineal o Regresión Logística:** mide la relación entre la variable dependiente categórica y una o más variables independientes estimando probabilidades utilizando una función logística/sigmoidea.

3. **Máquina de Vectores de Soporte:** Support Vector Machine (SVM) en Inglés, es un algoritmo de aprendizaje automático supervisado que se puede utilizar tanto para problemas de clasificación como de regresión. El modelo extrae el mejor hiperplano o línea posible que separa las dos clases.

El conjunto de datos se ha extraído de una plataforma de comercio electrónico de India [1]. La primer columna contiene alguna de las siguientes 4 categorías: Household, Books, Electronics, Clothing & Accessories. La segunda columna contiene la descripción del producto.

2. Metodología

Primero se revisó la distribución de frecuencias de las categorías en el conjunto de datos para confirmar que no se encuentren tan desproporcionados [Figura 1].

Posteriormente, a la columna que contenía el texto de la descripción del producto se le realizó el siguiente preprocesamiento:

- **Tokenización:** Se dividieron las reseñas en unidades más pequeñas, llamadas tokens, para facilitar su procesamiento y análisis.
- **Eliminación de palabras irrelevantes:** Se eliminaron las palabras vacías (*stopwords*) en inglés, que no aportan significado relevante para el análisis, tales como "the", .ªnd", ïs", entre otras.
- **Minúsculas:** Se convirtió el contenido de cada token a minúsculas.
- **Puntuación:** Se removieron números y signos de puntuación para asegurar que cada token solo contuviera letras.
- **Desinencias:** Se eliminaron desinencias (Stemming en Inglés) para reducir las palabras a su forma raíz, usando el algoritmo PorterStemmer.

El conjunto de datos se dividió en un 80 % entrenamiento y un 20 % validación. Con estos datos se midió el desempeño de cada uno de los modelos para cada una de las ingenierías de características.

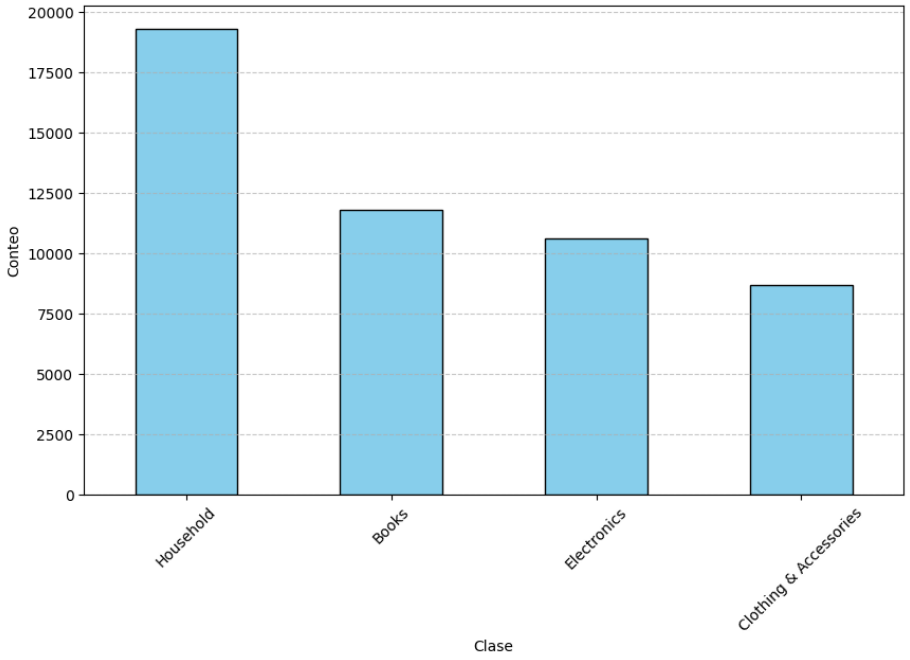

Figura 1: Distribución de frecuencia de las categorías.

3. Resultados

A continuación, se presenta una tabla de los resultados obtenidos para las descripciones de producto analizadas. La tabla muestra la precisión que se obtuvo para cada uno de los modelos aplicados con su respectiva ingeniería de características.

Ing. características / Modelo	**Naive Bayes**	**Reg. Logística**	**SVM**
Vectores de Recuento	95 %	98 %	96 %
TF-IDF a Nivel Palabra	95 %	97 %	98 %
TF-IDF a Nivel N-grama	85 %	90 %	91 %
TF-IDF a Nivel Caracter	92 %	95 %	97 %

Cuadro 1: Resultados de los modelos con diferentes técnicas de ingeniería de características.

El mejor resultado se obtiene con vectores TF-IDF a Nivel Palabra con el modelo de Máquina de Vectores de Soporte.

Referencias

[1] Gautam. (2019). E commerce text dataset (version - 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3355823