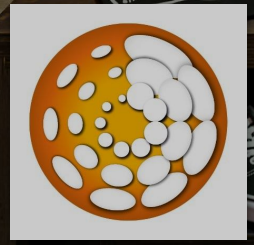
RECUPERACIÓN Y MINERÍA DE TEXTO

Clasificación de Productos

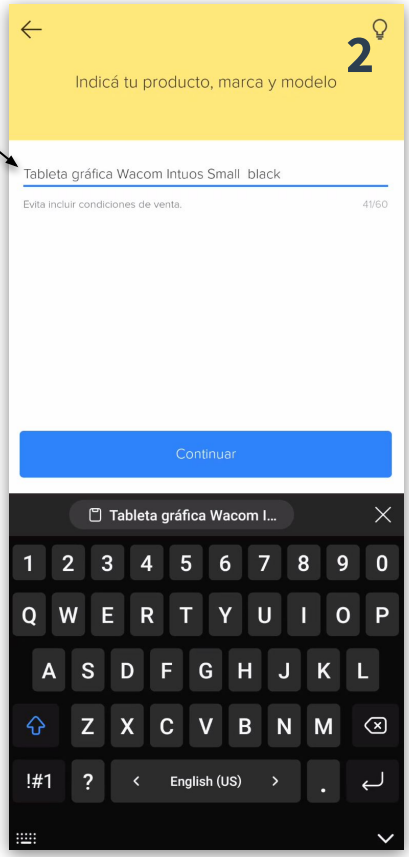


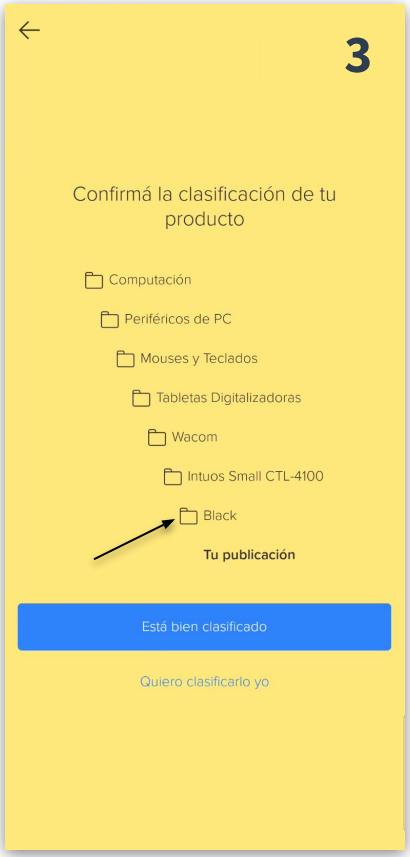
Adrián Marino
Gabriel Figueiro

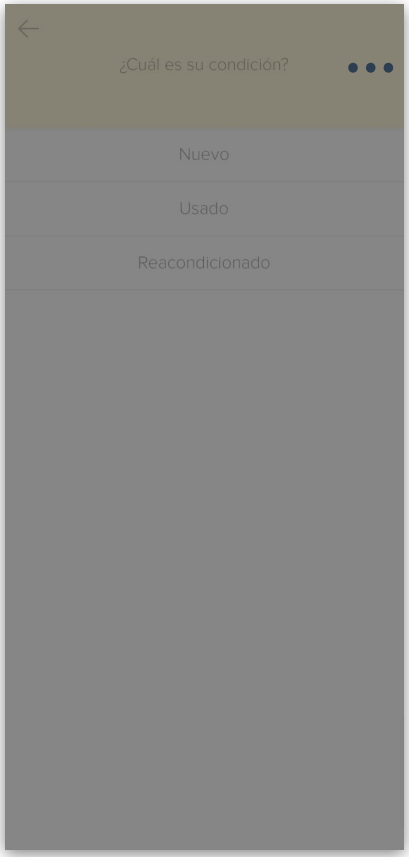
Pregunta

- **Caso de Uso**
  - El usuario realiza el alta de un nuevo producto en un e-commerce para su posterior venta.
- **Pregunta**
  - ¿Es posible sugerir la categoría del producto?
- **Contexto**
  - Existe un árbol de categorías.
  - Cada nodo de hoja contiene productos.

Caso de uso: Mercado Libre

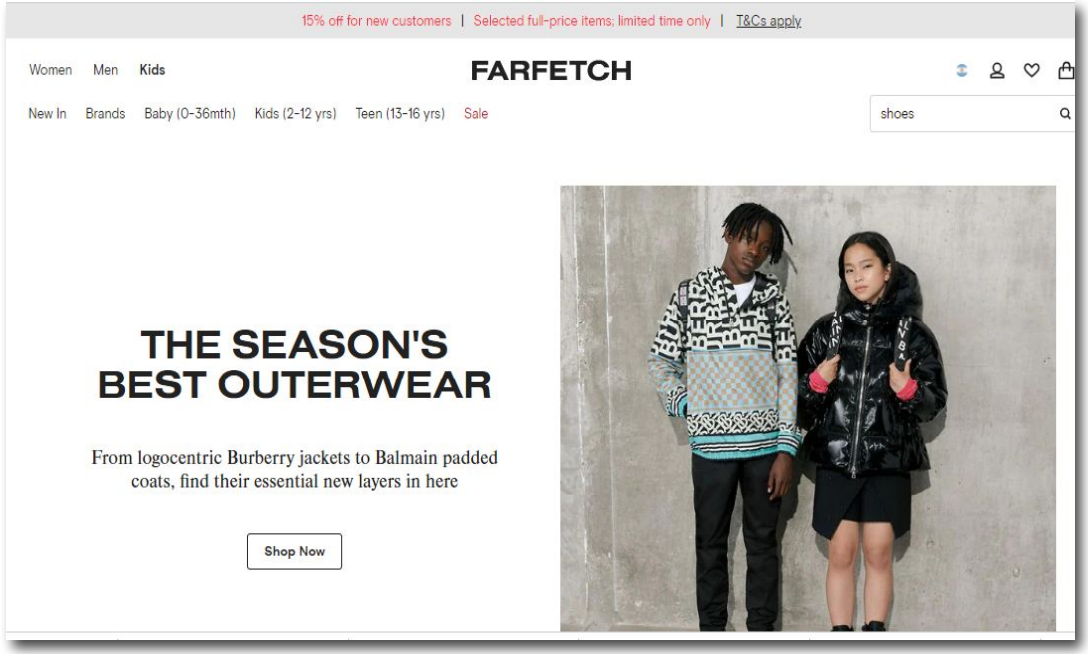






Dataset: Farfetch

FARFETCH



Farfetch is a British-Portuguese online luxury fashion retail platform that sells products from over 700 boutiques and brands from around the world. The company was founded in 2007 by the Portuguese entrepreneur José Neves with its headquarters in London and main branches in Lisbon and Porto.
Wikipedia

Dataset disponible en:
https://eval.ai/web/challenges/challenge-page/1721/overview

Dataset: Farfetch

FARFETCH

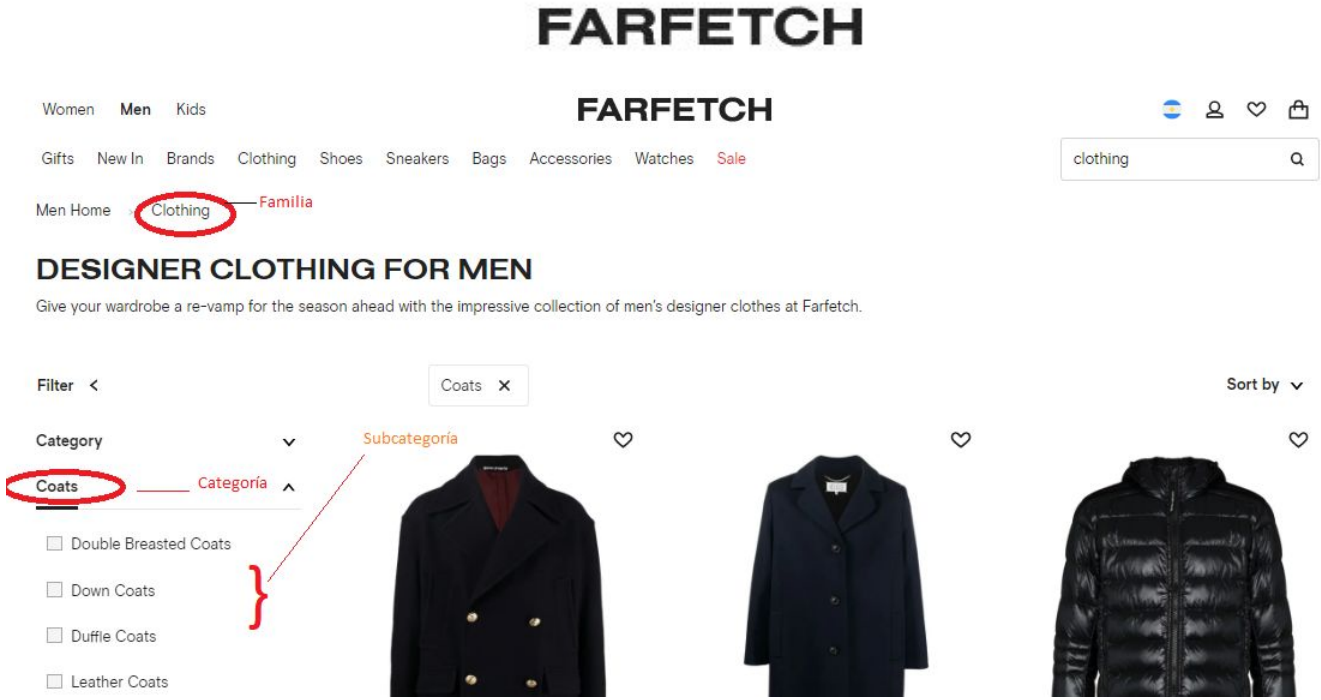
Features

- product.id
- **product.gender**
- product.main_colour
- product.second_color
- **product.brand**
- **product.materials**
- **product.description**
- product.attributes
- **product.highlights**

Target

- Branch
  - family
  - category
  - subcategory

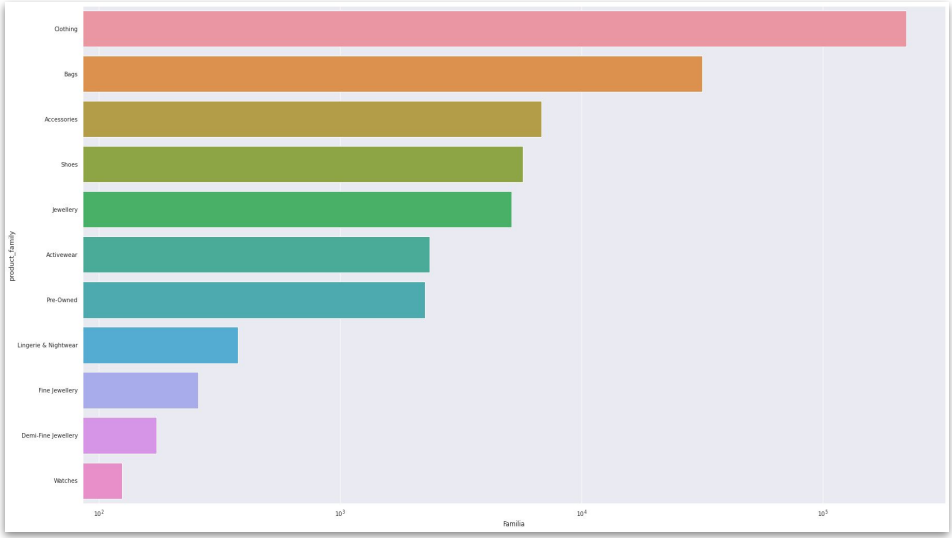
Plataforma


FARFETCH
FARFETCH
Women Men Kids
Gifts New In Brands Clothing Shoes Sneakers Bags Accessories Watches Sale
clothing
Men Home Clothing
Familia
DESIGNER CLOTHING FOR MEN
Give your wardrobe a re-vamp for the season ahead with the impressive collection of men's designer clothes at Farfetch.
Filter
Coats
Sort by
Category
Subcategoría
Coats
Categoría
Double Breasted Coats
Down Coats
Duffle Coats
Leather Coats

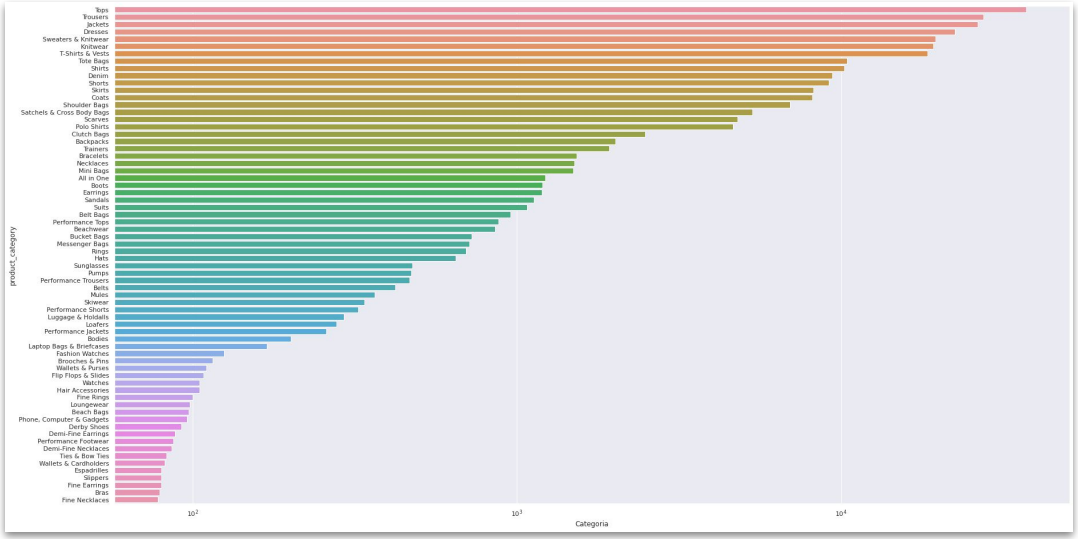
Dataset: Farfetch

FARFETCH

Familias de productos



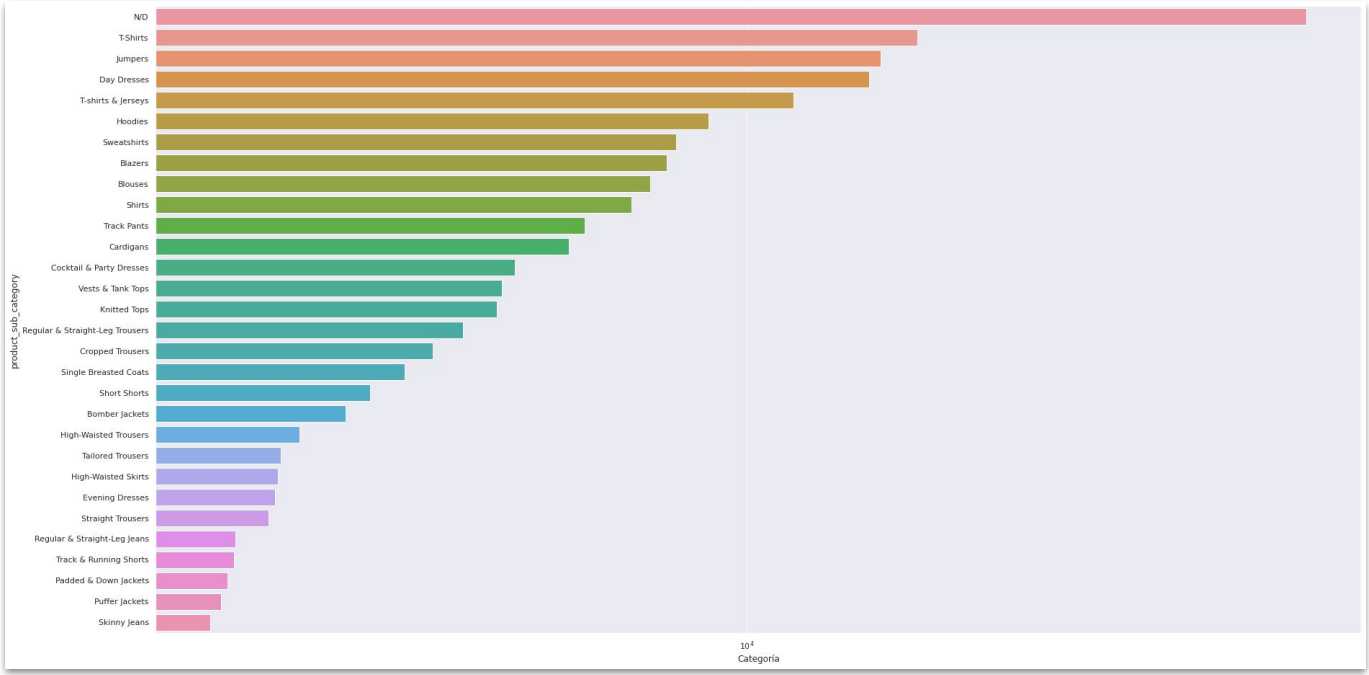
Categorías de productos



Dataset: Farfetch

FARFETCH

Subcategorías de productos



Dataset: Farfetch

FARFETCH

Unigramas

count	28933.000000
mean	293.636782
std	4373.587628
min	1.000000
25%	1.000000
50%	2.000000
75%	6.000000
max	225175.000000

Bigramas

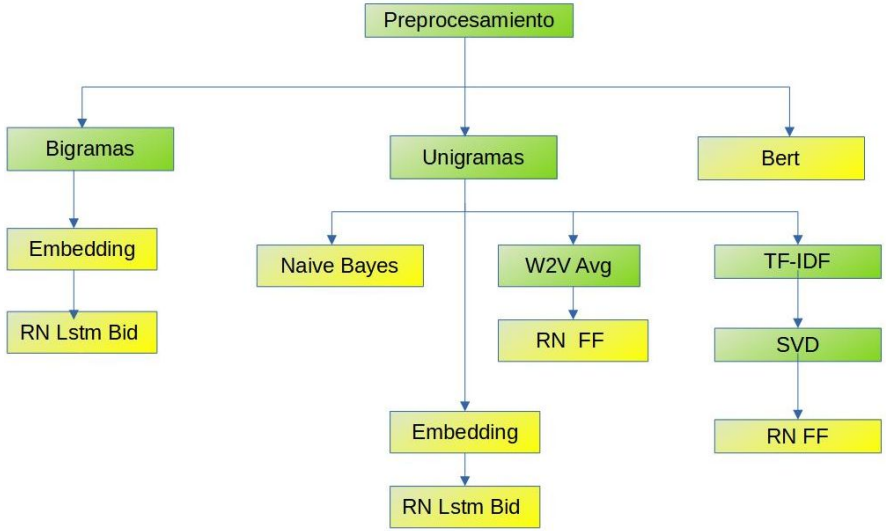
count	252992.000000
mean	31.002095
std	625.061755
min	1.000000
25%	1.000000
50%	1.000000
75%	3.000000
max	128671.000000

Dataset: Farfetch

FARFETCH

Nube de palabras por familia de productos

Modelos


Preprocesamiento
Bigramas
Unigramas
Bert
Embedding
Naive Bayes
W2V Avg
TF-IDF
RN Lstm Bid
RN FF
SVD
Embedding
RN FF
RN Lstm Bid

Preprocesamiento

Concatenación de variables description + highlights + gender

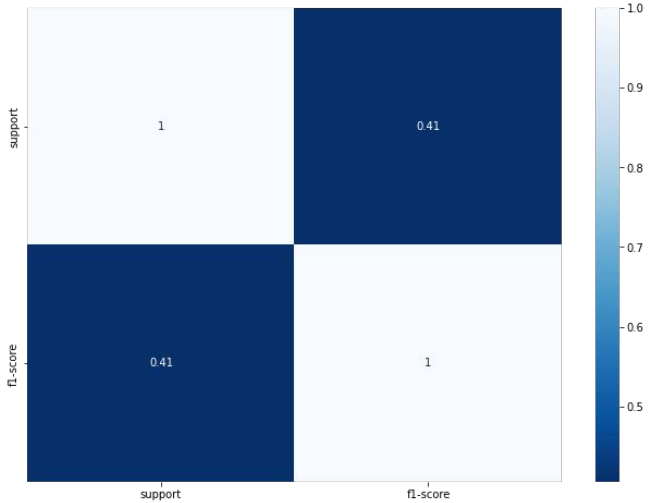
Eliminación de:

- Caracteres especiales
- Stop words en inglés
- Números
- Signos de puntuación

Modelos

- **Naive Bayes**
  - Features: short description, gender, highlights
  - Tokenizado con unigramas
  - Frecuencia mínima de tokens: 10

F1-score macro: 0.53

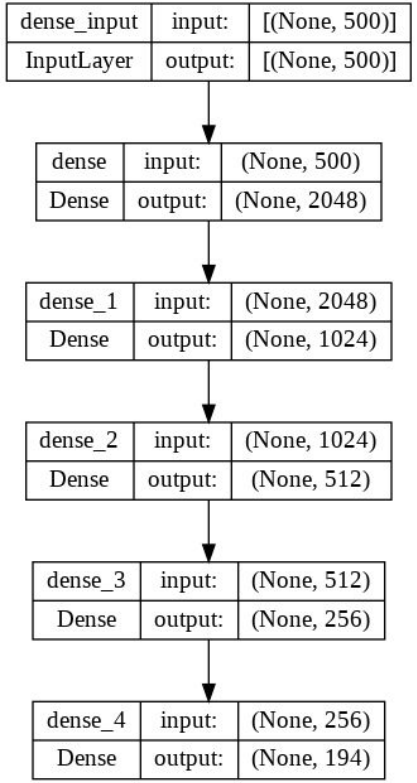


Modelos: Clasificador RN Feed Forward

- **Red Neuronal Feed Foward TF-IDF SVD**
  - Features: short description, gender, highlights
  - Tokenizado con unigramas
  - Frecuencia mínima del token: 10
  - TF-IDF
  - SVD de 500 componentes

Modelos: Clasificador RN Feed Forward

- **Red Neuronal Feed Foward TFIDF SVD**



F1-score: 0.64

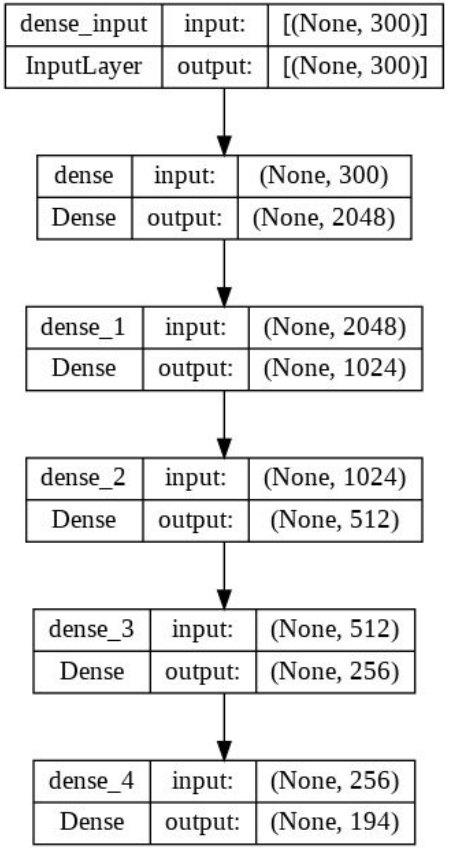
Modelos: Clasificador RN Feed Forward

- **word2Vect Averange**
  - Features: short description, gender, highlights
  - Tokenizado con unigramas
  - Frecuencia mínima del token: 20
  - vocabulario de 2.800 tokens
  - word2vect: ventana de 5 palabras y 30 iteraciones 300 dimensiones
  - Promediamos los embeddings para obtener un sentence embedding

Modelos: Clasificador RN Feed Forward

- **Red Neuronal Feed Foward**

F1-score: 0.61

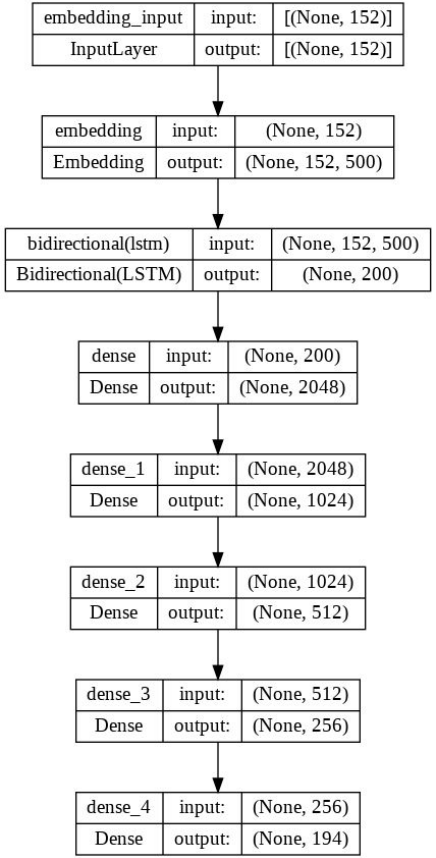


Modelos: Clasificador RNN LSTM Bidireccional Tokenizada con Unigramas

- **Red Neuronal LSTM Bidireccional con unigramas**
  - Features: short description, gender, highlights
  - Tokenizado con unigramas
  - Frecuencia mínima del token: 20
  - vocabulario de 2.800 tokens
  - Largo máximo de descripción de 152 tokens

Modelos: Clasificador RNN LSTM Bidireccional Tokenizada con Unigramas

F1-score: 0.66

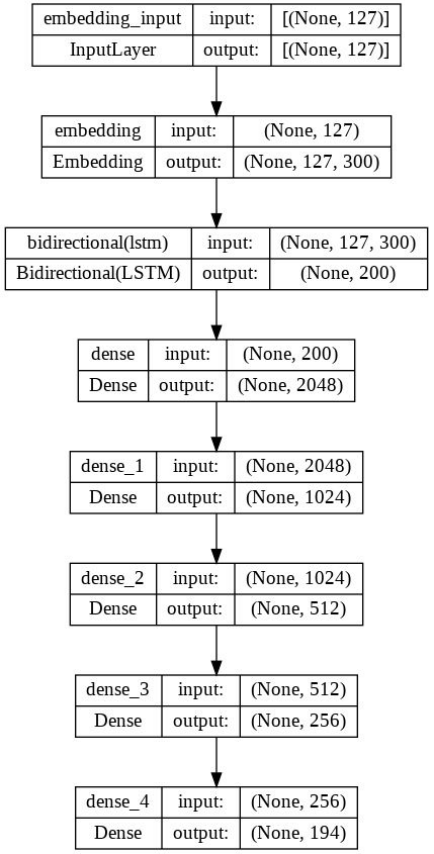


Modelos: Clasificador RNN LSTM Bidireccional Tokenizada con Bigramas

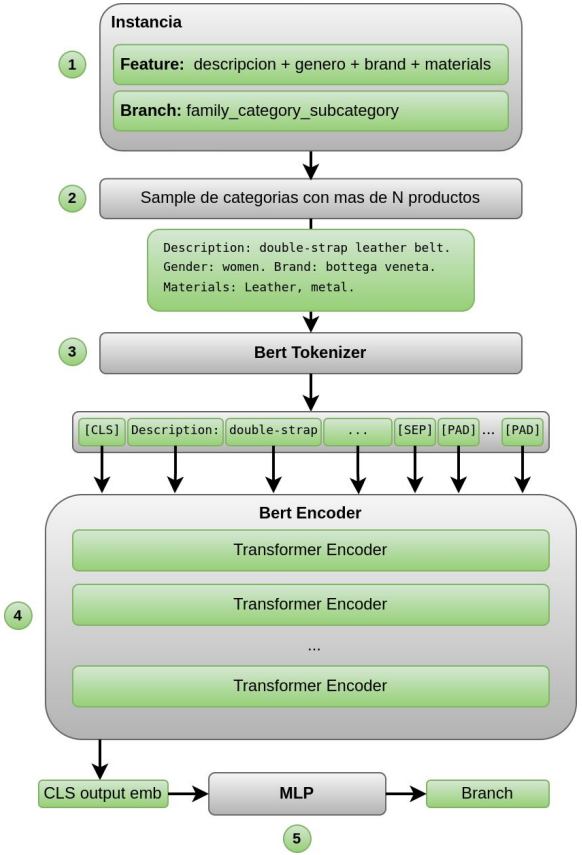
- **Red Neuronal LSTM Bidireccional**
  - Features: short description, gender, highlights
  - Tokenizado con bigramas
  - Frecuencia mínima del token: 20
  - vocabulario de 8.541 tokens
  - Largo máximo de descripción de 152 tokens

Modelos: Clasificador RNN LSTM Bidireccional Tokenizada con bigramas

F1-score: 0.64



Modelos: Clasificador BERT



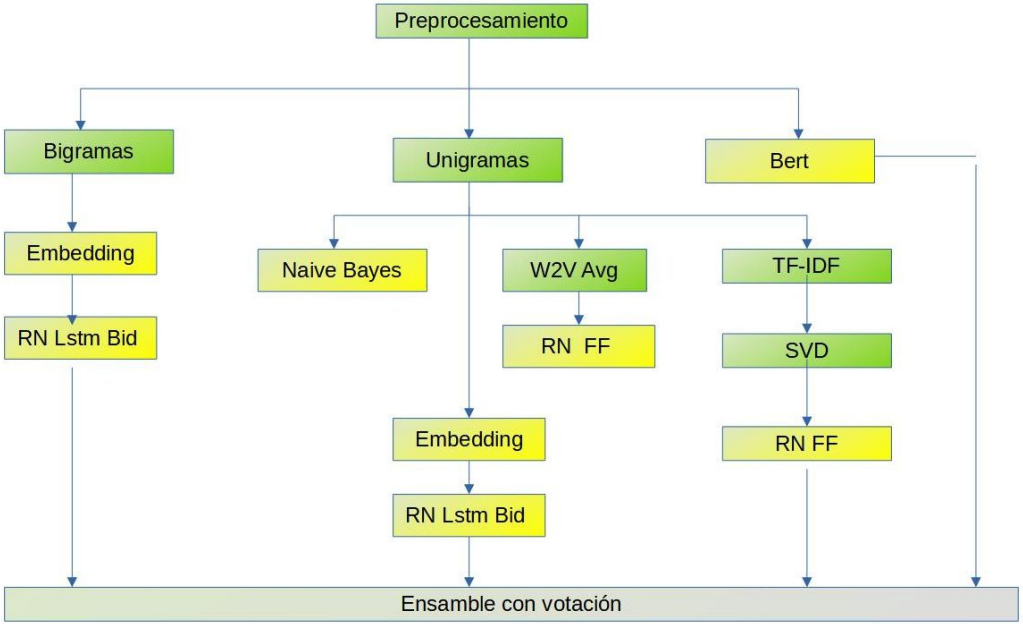
1. **Variables**
   - Feature: Es la variable input del modelo.
     - Es una concatenación de variables.
     - Género y branch aumentaron el f1-score en un 10% y 1%.
   - Branch: Es la clase a predecir.
2. **Sampling**: Para balancear el dataset sampleamos las categorías por encima de 1129 productos (Decil 0.7).
3. **Tokenizer**
   - Generamos las secuencias con Bert Tokenizer.
   - El tokenizer wrapea los tokens con CLS y SEP y completa con token PAD al máximo de secuencia definido (Max len: 39).
   - Tamaño máximo de la secuencia de 512.
4. **Encoder**: Probamos un Encoder BERT pre-entrenado con los siguientes pesos :
   - bert-base-cased (*)
   - bert-base-uncased (*)
   - distilbert-base-uncased
5. **MLP**
   - Tomamos el embedding de la salida correspondiente al token CLS.
   - Embedding de 768 dimensiones (en el caso del modelo de estos modelos).
   - Este embedding es la entrada de una red MPL (Capa densa).
   - La salida de la MLP pasa por una Softmax con igual número de salidas que clases(branch) en nuestro dataset.

Modelos: Clasificador BERT

Métricas (avg)

Por Nivel	Precision	Recall	F1-Score	Accuracy
Familia	0.85	0.88	**0.86**	0.99
Categoria	0.84	0.86	**0.84**	0.93
Subcategoria	0.69	0.71	**0.70**	0.81

Modelos: Ensamble por votación



F1-score: 0.70

Resultados Generales

F0.5-score

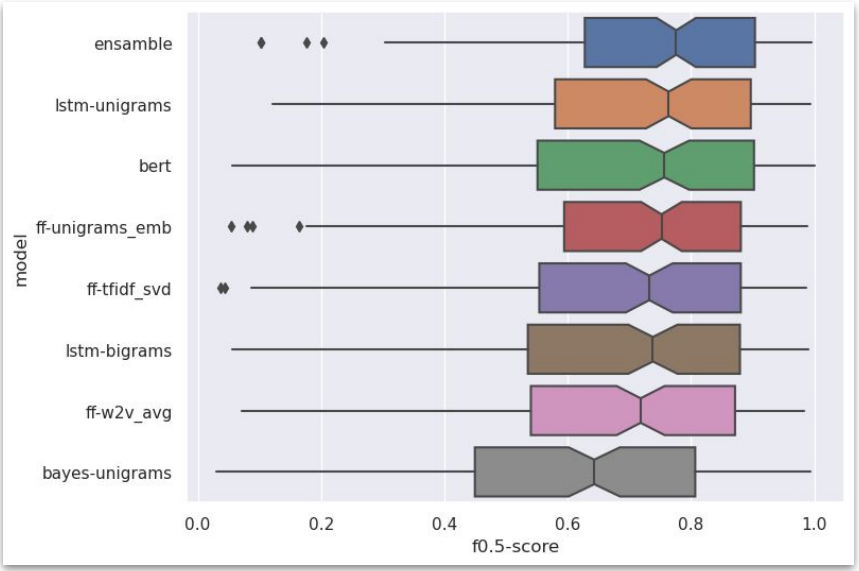
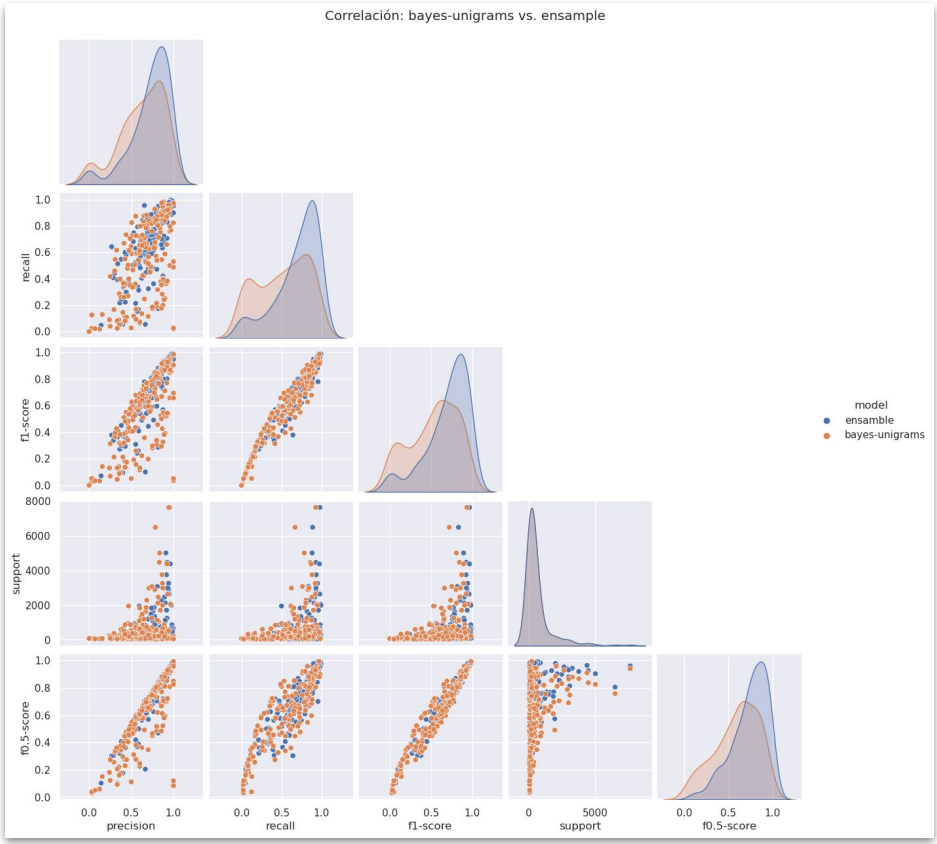


Tabla Comparativa

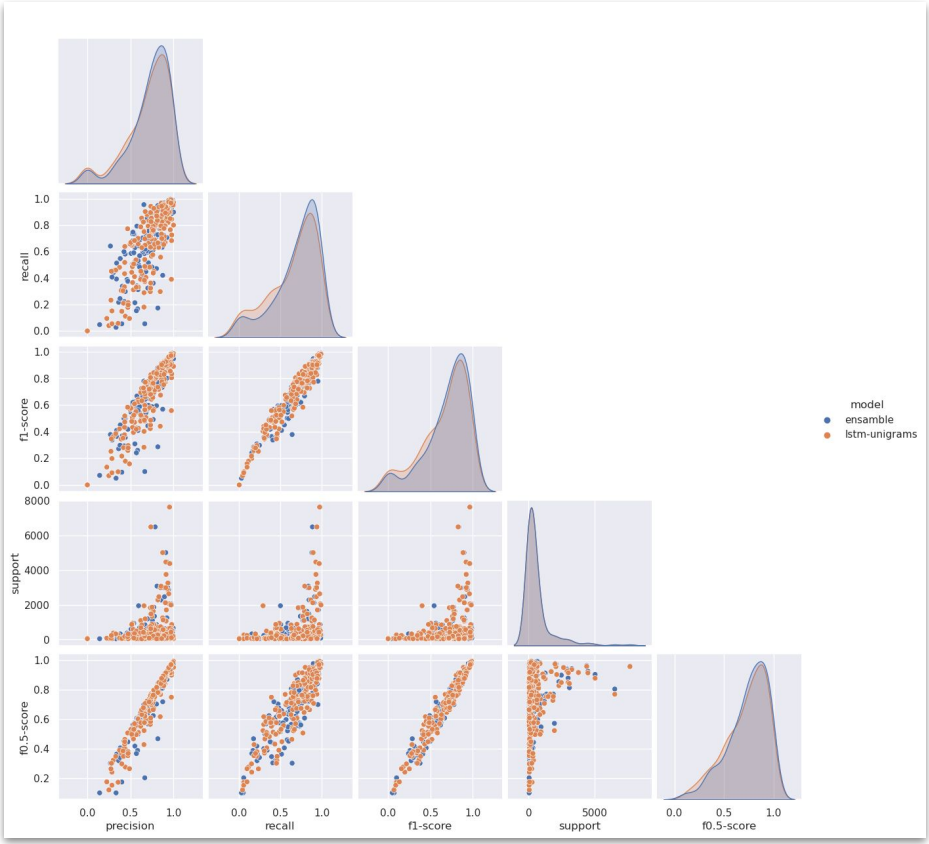
model	precision	recall	f1-score	f0.5-score
ensamble	0.73	0.70	0.70	0.74
lstm-unigrams	0.71	0.65	0.67	0.72
bert	0.69	0.73	0.70	0.71
ff-unigrams_emb	0.70	0.66	0.66	0.71
ff-tfidf_svd	0.67	0.63	0.64	0.70
lstm-bigrams	0.68	0.63	0.64	0.69
ff-w2v_avg	0.66	0.59	0.61	0.69
bayes-unigrams	0.63	0.51	0.53	0.60

Resultados



- **Distribución de F0.5-score por clase**
  - **Ensample**
    - Está sesgada a la derecha.
    - Mayor cantidad de clases con f-score alto.
    - Pocas clases hasta 0.3 aprox.
  - **Bayes**
    - Menos sesgada a la derecha.
    - Menor concentración de clases para valores **alto** de f-score.
    - Mayor concentración de clases para valores **bajos** de f-score.
- **Precision**
  - **Ensample**
    - Mayor número de clases con alta precision.
  - **Bayes**
    - Se distribuye en valores más bajos de precisión.
- **Recall**
  - **Bayes**
    - Tiende a capturar pocos ejemplos positivos por clase.
    - Mayor probabilidad en valores bajos de recall.

Resultados



- **F1-score**
  - **Lstm-unigrams**
    - Concentra más categorías en valores bajo de f0.5-score.
- **F0.5-score**
  - **Lstm-unigrams**
    - Concentra más categorías en valores bajo de f0.5-score.
  - **Ensample**
    - Aumento importante en precision, recall en menor medida.

- **En resumen**
  - Ambos modelos son muy similares en términos de precisión.
  - Bert gana en Recall.

model	precision	recall	f1-score	f0.5-score
ensamble	0.73	0.70	0.70	0.74
lstm-unigrams	0.71	0.65	0.67	0.72
bert	0.69	0.73	0.70	0.71

Proximos Pasos

- Analizar en detalle ejemplos mal clasificados.
- Probar con modelos bert-large-cased/uncased.
- Ensamble utilizando Probabilidades.
- Nuevos Features
  - main y secondary colors.
  - Imágenes de productos.
- Detectar productos mal asignados a una categoría.
- Fusionar ramas.
  - Categorías sinónimo.
  - Contienen el mismo tipo de productos.

¿Preguntas?