Categorización de publicaciones de productos realizadas en Mercado Libre

Mentoría

Eduardo Barseghian - Benjamín Ocampo - Maximiliano Tejerina

Diplomatura en Ciencia de datos, Aprendizaje Automático y Aplicaciones
FaMAF

July 5, 2021

Contenidos

Table of Contents

En 2019, Mercado Libre impulsó un challenge donde desafió a predecir la categoría más adecuada de un producto dado el título de una publicación; una problemática que invita a explorar técnicas fundamentalmente de Procesamiento del Lenguaje Natural (NLP).
Para ello hizo público un dataset que contiene publicaciones con sus títulos y categorías a la que pertenecen. Este dataset fue filtrado para quedarnos con las categorías con más de 30000 publicaciones.
Hay cerca de 650.000 publicaciones, todas en español o portugués. Varias ya han sido revisadas y calificadas como RELIABLES. En caso contrario , se las etiqueta como UNRELIABLE.

Encabezado del DataFrame

```
df.shape
```

```
(646760, 4)
```

```
df.head(20)
```

	title	label_quality	language	category
0	Galoneira Semi Industrial	unreliable	portuguese	SEWING_MACHINES
1	Máquina De Coser Brother Industrial	unreliable	spanish	SEWING_MACHINES
2	Teclado Casio Wk-240 76 Teclas Profissional St...	unreliable	portuguese	MUSICAL_KEYBOARDS
3	Heladera Gafa 380 Impecable Urgente	unreliable	spanish	REFRIGERATORS
4	Butaca 6 Cuotas Sin Interes!! Para Auto Bebes...	unreliable	spanish	BABY_CAR_SEATS
5	Reloj De Pared - Varios Modelos	unreliable	spanish	WALL_CLOCKS
6	Teclado Sintetizador Moxf8 Preto Yamaha	unreliable	portuguese	MUSICAL_KEYBOARDS
7	Coche Travel System Graco Modes Trinidad C/ Hu...	unreliable	spanish	BABY_STROLLERS
8	Bermuda Les Mills Mujer Nuevo Exclusivo Import...	unreliable	spanish	SHORTS
9	Geladeira Eletrolux Rd30	unreliable	portuguese	REFRIGERATORS
10	Reloj Pared Vox Tronic Blanco Numeros 23cm Gar...	reliable	spanish	WALL_CLOCKS
11	Conjunto Mala De Viagem - P, M - Com Rodinha -...	reliable	portuguese	SUITCASES
12	Máquina De Costura - Singer Stylist 7258 Com V...	unreliable	portuguese	SEWING_MACHINES
13	Rottwalier Femea	unreliable	portuguese	PUREBRED_DOGS
14	Carrinho De Bebê Tutti Baby Semi Novo	unreliable	portuguese	BABY_STROLLERS
15	Teclado Yamaha Psr E-363 - Preto Com Fonte Biv...	unreliable	portuguese	MUSICAL_KEYBOARDS
16	Relógio Parede Vintage Cerveja Decoração Área ...	unreliable	portuguese	WALL_CLOCKS
17	Carrinho Bebe Gemeos	reliable	portuguese	BABY_STROLLERS
18	Patines Nena Tipo Soy Luna . Envíos Solo En La...	unreliable	spanish	ROLLER_SKATES
19	Mala De Viagem Bordo Com Rodinhas Cereja Primí...	reliable	portuguese	SUITCASES

20 categorías

SEWING_MACHINES
REFRIGERATORS
WALL_CLOCKS
SHORTS
PUREBRED_DOGS
COFFEE_MAKERS
MATTRESSES
MEMORY_CARDS
MOTORCYCLE_JACKETS
RANGES
MUSICAL_KEYBOARDS
BABY_CAR_SEATS
BABY_STROLLERS
SUITCASES
ROLLER_SKATES
WINES
PANTS
ELECTRIC_DRILLS
HAIR_CLIPPERS
KITCHEN_SINKS

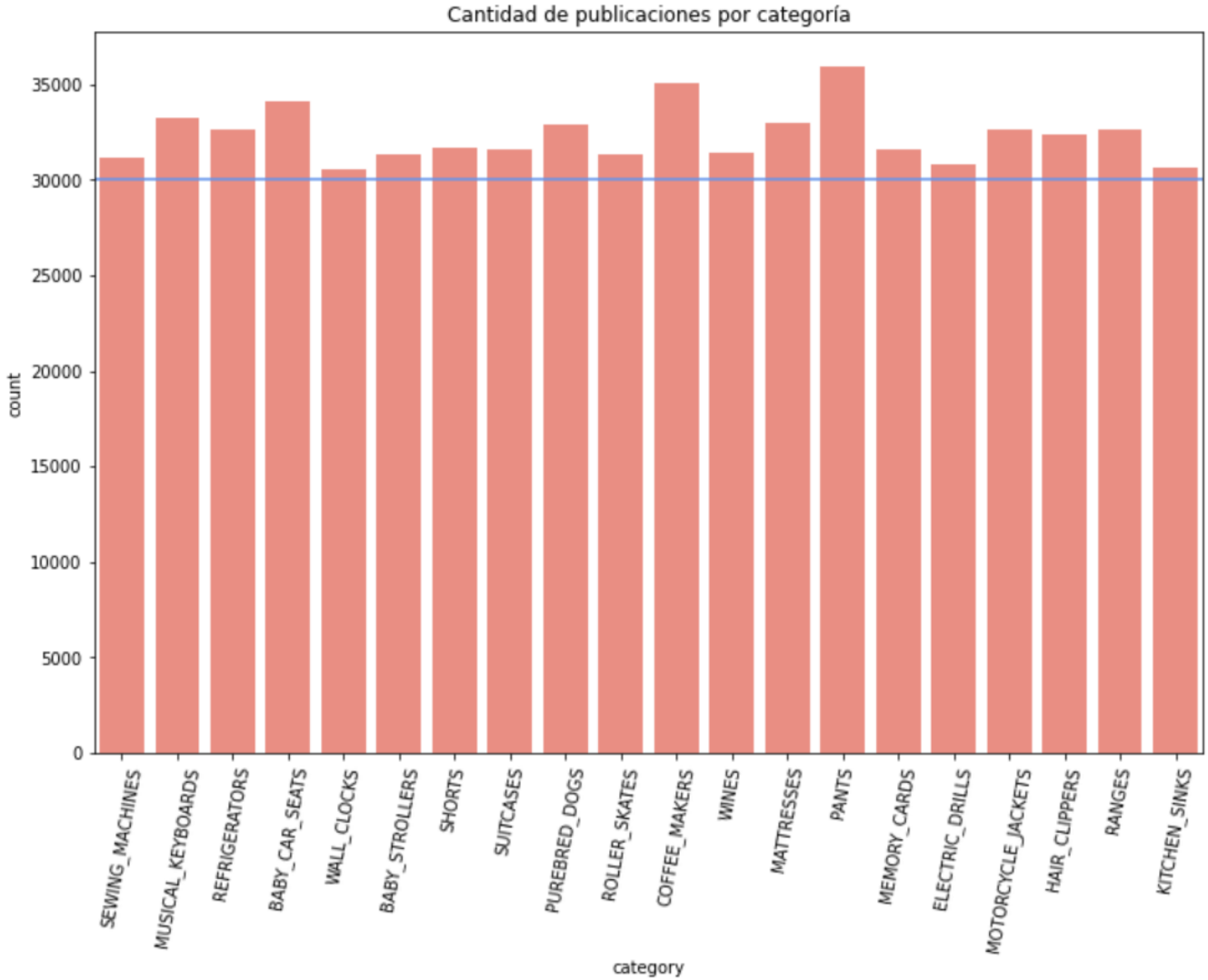

Figure: Todas las categorías tienen entre 30000 y 36000 publicaciones.

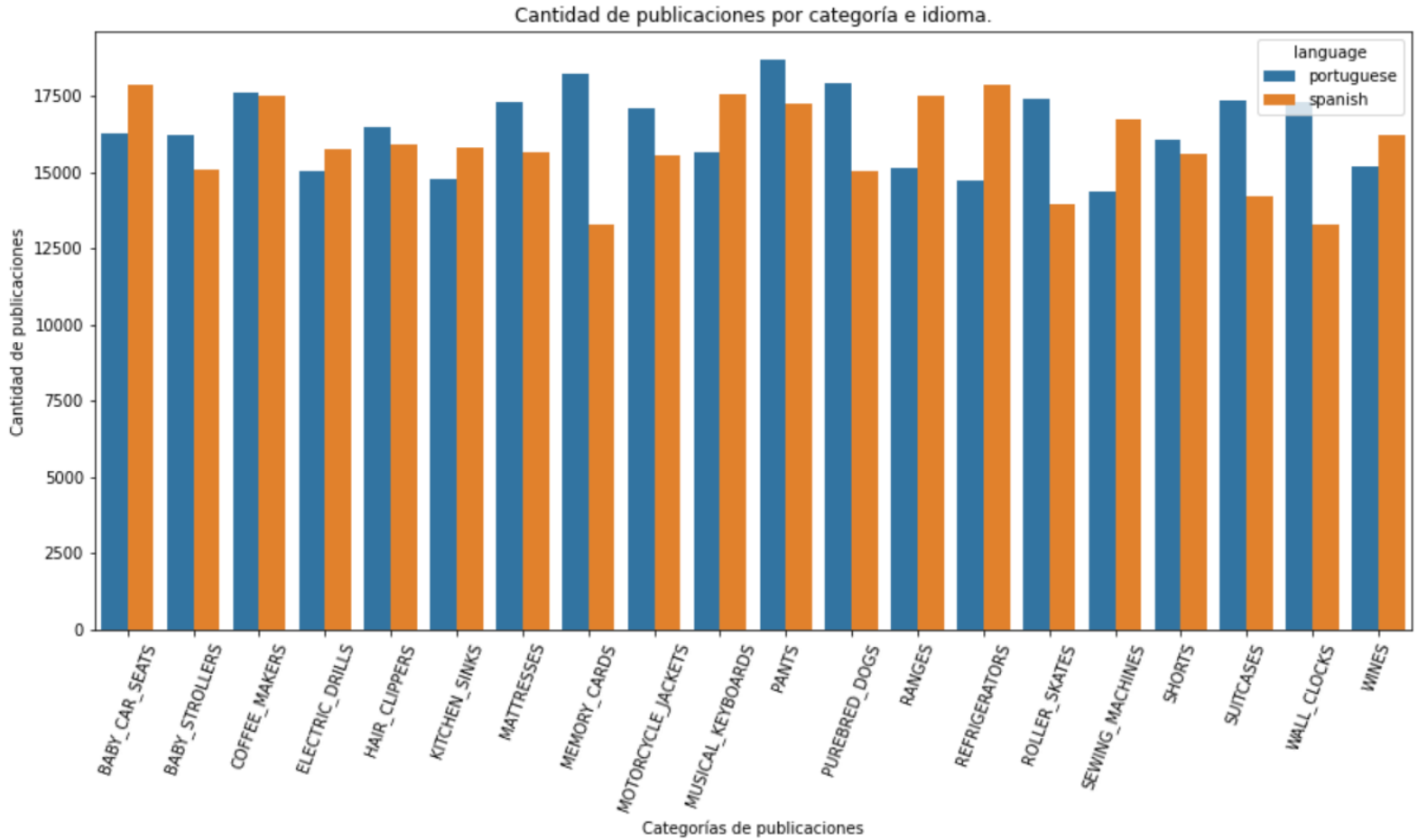

Figure: Para todas las categorías, cada idioma aparece al menos en el 42% de los casos. Del total de 646760 publicaciones, 317768 (49,13%) son en español y 328992 (50,86%) son en portugués.

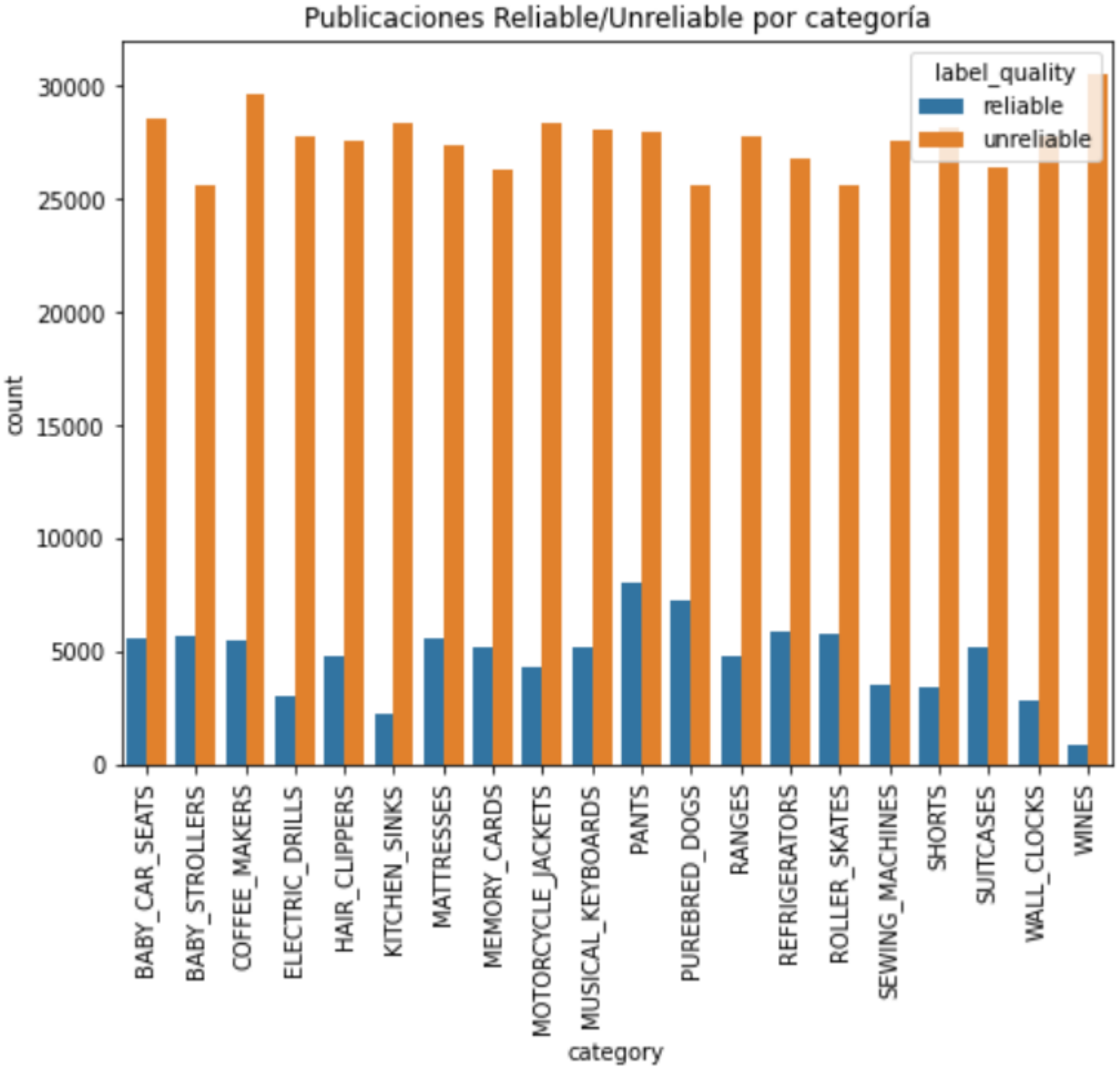

Figure: Para ninguna categoría la proporción de reliables es mayor a 22,3%.

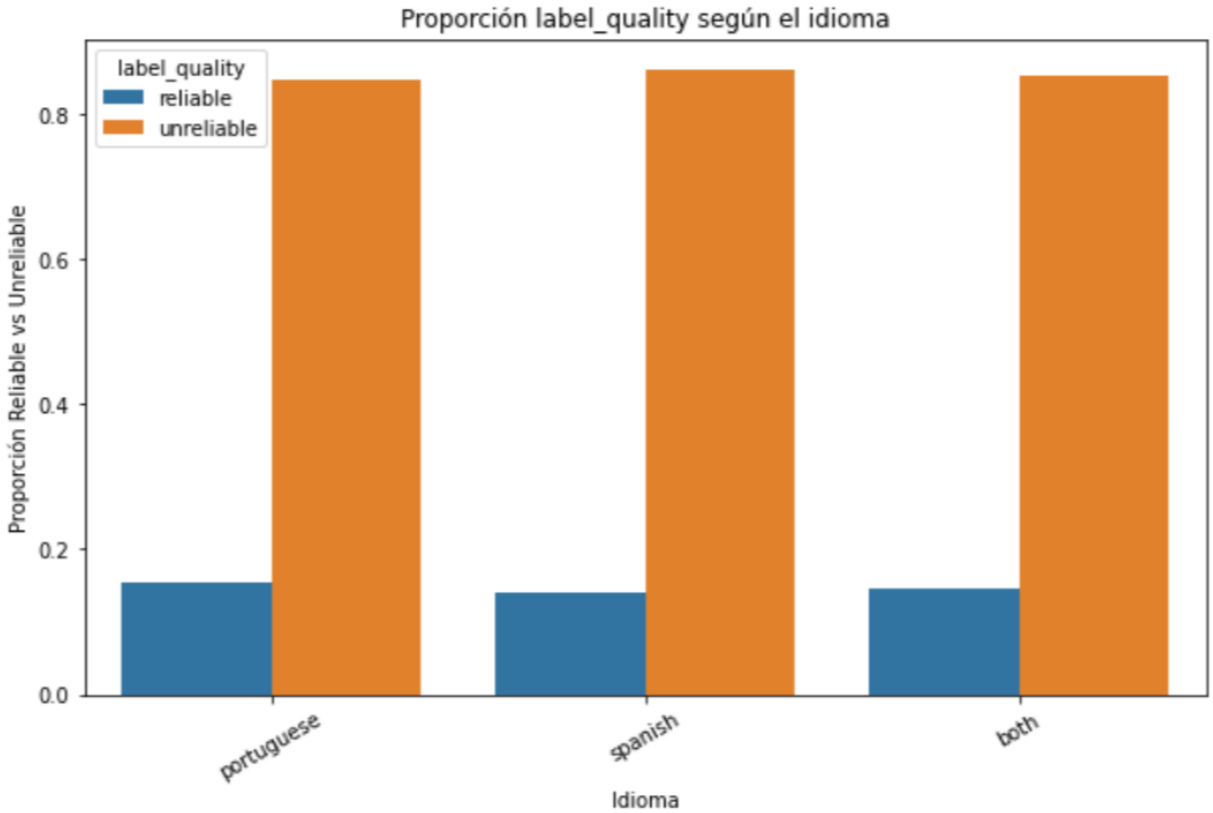

Figure: Label_quality es independiente del idioma, siempre el porcentaje de Reliables ronda el 15%.

Table of Contents

Se contabilizaron (en ambos idiomas):

- palabras
- **stopwords**: preposiciones, artículos, conjunciones, etc...
- caracteres especiales
- dígitos

	count_words	count_stopwords	count_digits	count_special_chars	title	category
0	3	0	0	0	Galoneira Semi Industrial	SEWING_MACHINES
1	5	1	0	0	Máquina De Coser Brother Industrial	SEWING_MACHINES
2	7	0	5	1	Teclado Casio Wk-240 76 Teclas Profissional St...	MUSICAL_KEYBOARDS
3	5	0	3	0	Heladera Gafa 380 Impecable Urgente	REFRIGERATORS
4	10	3	3	2	Butaca 6 Cuotas Sin Interes!! Para Auto Bebes...	BABY_CAR_SEATS
...	...	...	...	...	...	...
646755	5	0	1	0	Thank You Malbec X 6	WINES
646756	8	1	0	0	Cachorros Jack Rusell Terrier Pelo Corto Ultim...	PUREBRED_DOGS
646757	9	0	8	0	Colchão Box Casal Castor Vitagel Euro One Face...	MATTRESSES
646758	6	3	0	2	Maquina De Cortar El Pelo. Starex.	HAIR_CLIPPERS
646759	7	0	0	1	Trimmer Detailer Wahl + Kit Tijeras Stylecut	HAIR_CLIPPERS

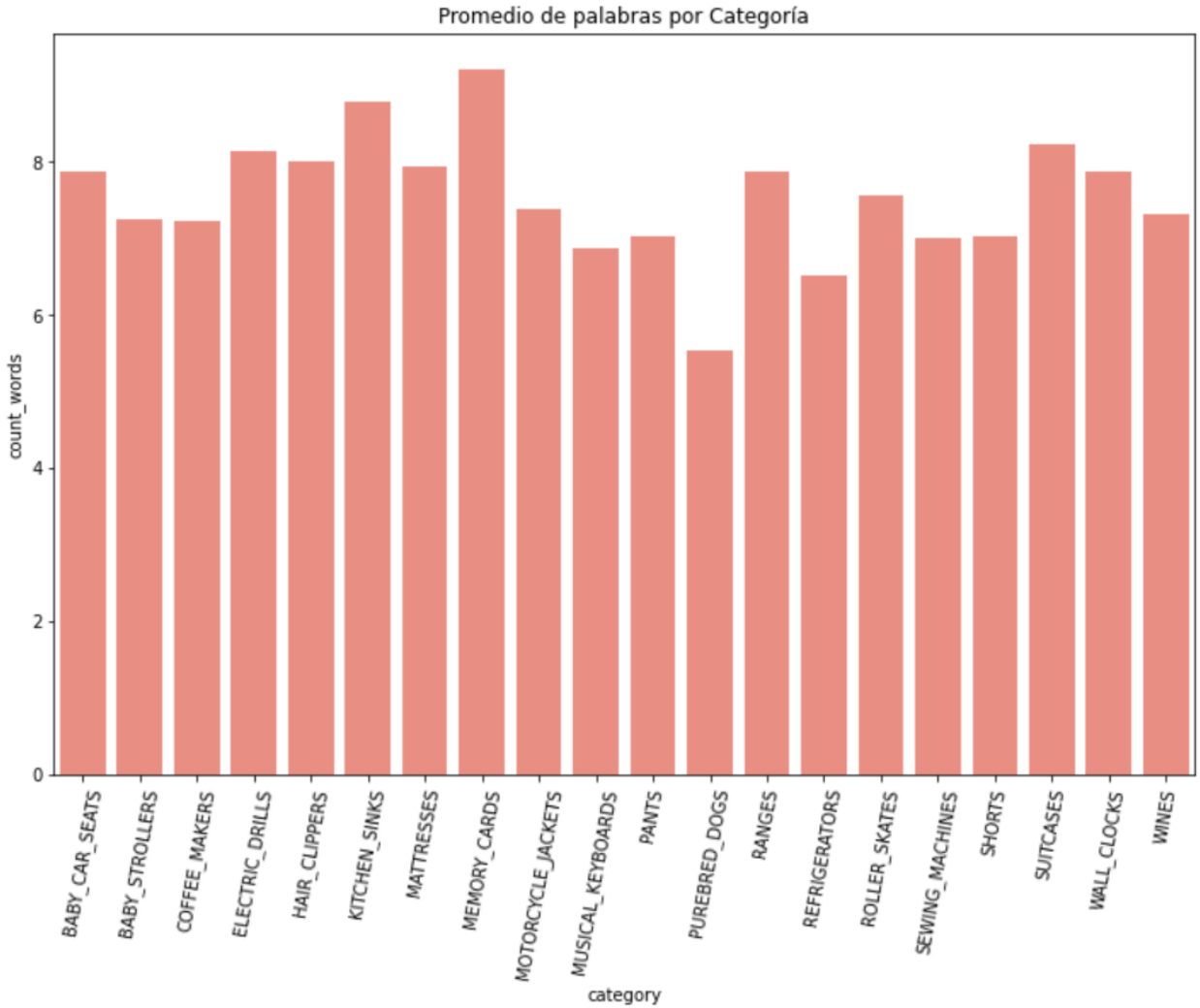

Figure: La cantidad promedio de palabras por título agrupada por categoría no oscila demasiado

.

category	count_words	count_stopwords	count_digits	count_special_chars
BABY_CAR_SEATS	7.879694	1.148084	1.781313	0.742646
BABY_STROLLERS	7.251651	0.838580	0.717252	0.690173
COFFEE_MAKERS	7.231284	0.532817	2.746154	0.716642
ELECTRIC_DRILLS	8.143543	0.584263	6.648864	1.312784
HAIR_CLIPPERS	8.017144	1.166378	2.135117	0.679291
KITCHEN_SINKS	8.778619	0.750449	5.718851	0.965628
MATTRESSES	7.944611	0.374587	5.576425	0.932751
MEMORY_CARDS	9.218318	0.481688	4.512958	0.870200
MOTORCYCLE_JACKETS	7.395156	0.298605	1.079657	0.802698
MUSICAL_KEYBOARDS	6.869213	0.422491	3.322557	1.051532
PANTS	7.024546	0.460123	1.342618	0.546187
PUREBRED_DOGS	5.543610	0.733540	0.334305	0.588192
RANGES	7.878726	0.541859	2.526512	0.622086
REFRIGERATORS	6.527103	0.559277	3.182320	0.643144
ROLLER_SKATES	7.568837	0.559721	2.683561	0.984444
SEWING_MACHINES	7.019050	0.938514	2.281313	0.688008
SHORTS	7.025217	0.498627	1.332870	0.551144
SUITCASES	8.226282	0.824478	2.398923	0.790374
WALL_CLOCKS	7.878007	1.219379	1.292222	0.616536
WINES	7.320902	0.438326	2.563075	0.864136

Figure: La cantidad promedio por categoría de palabras, stopwords, dígitos y caracteres especiales.

Buscamos las palabras más frecuentes por categoría. Se removieron las stopwords, caracteres numéricos y especiales. Se llevaron todas las palabras a minúsculas. Luego para cada categoría se obtiene la frecuencia de todas sus palabras para finalmente quedarse con las mejores 10.
No se separó por idioma, llevando así a dar lugar a palabras frecuentes que tengan misma traducción.

	category	top10_word_freq
0	BABY_CAR_SEATS	[(auto, 13055), (kg, 11740), (butaca, 10273), ...
1	BABY_STROLLERS	[(carrinho, 14981), (bebê, 10015), (cochecito,...
2	COFFEE_MAKERS	[(cafetera, 15022), (cafeteira, 14190), (v, 73...
3	ELECTRIC_DRILLS	[(w, 16248), (taladro, 14368), (furadeira, 142...
4	HAIR_CLIPPERS	[(maquina, 9336), (wahl, 7920), (máquina, 6494...
5	KITCHEN_SINKS	[(bacha, 12567), (johnson, 10927), (cuba, 1052...
6	MATTRESSES	[(colchão, 15933), (x, 12718), (colchon, 8968)...
7	MEMORY_CARDS	[(gb, 27759), (sd, 16749), (micro, 16742), (ca...
8	MOTORCYCLE_JACKETS	[(jaqueta, 16633), (campera, 13618), (moto, 76...
9	MUSICAL_KEYBOARDS	[(teclado, 23205), (casio, 8659), (yamaha, 800...

	category	top10_word_freq
10	PANTS	[(calça, 14796), (jeans, 10027), (pantalon, 51...
11	PUREBRED_DOGS	[(filhotes, 5674), (macho, 4225), (filhote, 37...
12	RANGES	[(cocina, 16894), (fogão, 12891), (bocas, 1122...
13	REFRIGERATORS	[(heladera, 17012), (geladeira, 7163), (frost,...
14	ROLLER_SKATES	[(patins, 15878), (roller, 7160), (patines, 57...
15	SEWING_MACHINES	[(maquina, 12090), (máquina, 9834), (coser, 89...
16	SHORTS	[(short, 14075), (bermuda, 11323), (kit, 5461)...
17	SUITCASES	[(mala, 13864), (valija, 11176), (viagem, 1034...
18	WALL_CLOCKS	[(relógio, 13860), (parede, 13644), (reloj, 12...
19	WINES	[(vinho, 12119), (ml, 9620), (malbec, 8716), (...

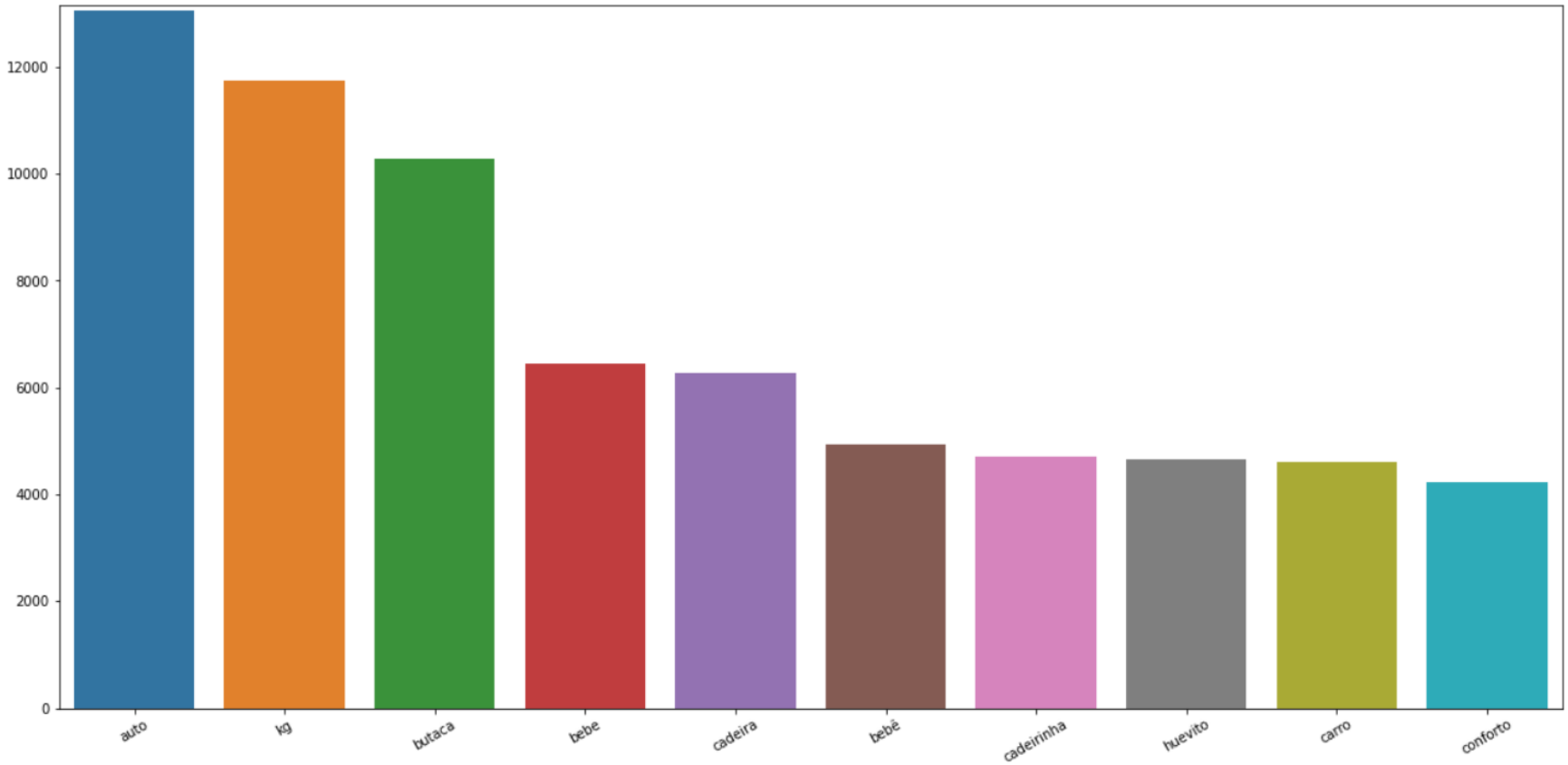

Figure: Top 10 palabras frecuentes en la categoría BABY_CAR_SEATS

.

Table of Contents

Se procedió a realizar un preprocesamiento de los títulos donde se reemplazan:

- Mayúsculas por minúsculas.
- Caracteres que no tienen una codificación ascii.
- Números y símbolos.
- Contracciones de palabras por su expresión completa.
- **Stopwords** (por idioma)

	title	cleaned_title
0	Galoneira Semi Industrial	galoneira semi industrial
1	Máquina De Coser Brother Industrial	maquina coser brother industrial
2	Teclado Casio Wk-240 76 Teclas Profissional St...	teclado casio wk teclas profissional standard
3	Heladera Gafa 380 Impecable Urgente	heladera gafa impecable urgente
4	Butaca 6 Cuotas Sin Interes!! Para Auto Bebes...	butaca cuotas interes auto bebesit kg
...	...	...
646755	Thank You Malbec X 6	thank you malbec x
646756	Cachorros Jack Rusell Terrier Pelo Corto Ultim...	cachorros jack rusell terrier pelo corto ultim...
646757	Colchão Box Casal Castor Vitagel Euro One Face...	colchao box casal castor vitagel euro one face xx
646758	Maquina De Cortar El Pelo. Starex.	maquina cortar pelo starex
646759	Trimmer Detailer Wahl + Kit Tijeras Stylecut	trimmer detailer wahl kit tijeras stylecut

646760 rows × 2 columns

Figure: Columnas título y título procesado.

.

Tokenización y Secuencias

Limpiados los títulos, se confeccionó un vocabulario con sus palabras o **tokens** mediante la clase **Tokenizer** de Keras vía *fit_on_texts*.

```
word_tokenizer.word_counts
```

```
OrderedDict([('galoneira', 1067),
             ('semi', 3729),
             ('industrial', 12259),
             ('maquina', 40530),
             ('coser', 8954),
             ('brother', 1417),
             ('teclado', 23218),
             ('casio', 8711),
             ('wk', 626),
             ('teclas', 6866),
             ('profissional', 4604),
             ('standard', 497),
             ('heladera', 17026),
             ('gafa', 1723),
             ('impecable', 3462),
```

```
word_tokenizer.word_index
```

```
 'polido': 891,
 'hero': 892,
 'nuevos': 893,
 'cybex': 894,
 'shaver': 895,
 'chaqueta': 896,
 'rasgada': 897,
 'sleep': 898,
 'local': 899,
 'cubas': 900,
 'cabelos': 901,
 'palermo': 902,
 'cochesito': 903,
 'exclusive': 904,
 'vestir': 905,
```

word_counts: frecuencias de las palabras del vocabulario. **word_index**: nos muestra cada palabra con su índice. A cada palabra se le asigna un índice según su frecuencia; a mayor frecuencia menor índice. El token *maquina* tiene la mayor frecuencia y así índice 1.

Se codificaron los títulos asignándole un vector correspondiente a los índices de cada una de sus palabras(*texts_to_sequences*). Por ejemplo, al título *galoneira semi industrial* se le asigna el vector [580, 188, 40].

```
(
    word_tokenizer.word_index["galoneira"],
    word_tokenizer.word_index["semi"],
    word_tokenizer.word_index["industrial"]
)
```

```
(580, 188, 40)
```

Se rellenan los títulos con ceros para que tengan todos igual longitud.

```
encoded_titles = sequence.pad_sequences(encoded_titles, padding='post')
encoded_titles[:5]
```

```
array([[ 580,  188,   40,    0,    0,    0,    0,    0,    0,    0,    0,
            0,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   1,   59,  474,   40,    0,    0,    0,    0,    0,    0,    0,
            0,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   8,   62,  926,   90,  148, 1131,    0,    0,    0,    0,    0,
            0,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [  13,  401,  206, 1629,    0,    0,    0,    0,    0,    0,    0,
            0,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [  51,  140,  723,   32,  517,   28,    0,    0,    0,    0,    0,
            0,    0,    0,    0,    0,    0,    0,    0,    0,    0]])
```

Codificación de etiquetas: *Label encoding*

De igual manera que para los títulos, se necesitó codificar sus categorías asignadas. Sin embargo, no fue por medio de secuencias de números sino más bien a través de un índice por cada etiqueta, usando una instancia de la clase `LabelEncoder` de la librería `scikit-learn`.

```
le = LabelEncoder()
encoded_labels = le.fit_transform(df["category"])
encoded_labels
```

```
array([15, 15,  9, ...,  6,  4,  4])
```

```
le.classes_
```

```
array(['BABY_CAR_SEATS', 'BABY_STROLLERS', 'COFFEE_MAKERS',
       'ELECTRIC_DRILLS', 'HAIR_CLIPPERS', 'KITCHEN_SINKS', 'MATTRESSES',
       'MEMORY_CARDS', 'MOTORCYCLE_JACKETS', 'MUSICAL_KEYBOARDS', 'PANTS',
       'PUREBRED_DOGS', 'RANGES', 'REFRIGERATORS', 'ROLLER_SKATES',
       'SEWING_MACHINES', 'SHORTS', 'SUITCASES', 'WALL_CLOCKS', 'WINES'],
      dtype=object)
```

Table of Contents

Queremos representar las palabras con vectores que mantengan su semántica; que aquellas con un significado similar tengan una representación vectorial similar.
Un ***word embedding** es una matriz de parámetros entrenables donde a cada palabra del vocabulario se le asigna un vector representante. Se obtuvo esta matriz con 2 métodos distintos:

- Entrenando los parámetros por medio de los datos disponibles (**Custom**).
- Utilizando representaciones disponibles en la web (**Pretrained**).

Custom

Para instanciar un embedding del vocabulario, se necesita

- longitud del título más largo
- dimensión de los vectores representación.
- cantidad total de palabras distintas en los títulos
- número de dimensiones para representar las palabras fuera del vocabulario
- tamaño del vocabulario (suma de las 2 anteriores)

```
embedding_layer = Embedding(len(word_tokenizer.word_index)+1,
                            64,
                            input_length=encoded_titles.shape[1]
                           )
embedding_layer(tf.constant([1, 2]))
```

```
<tf.Tensor: shape=(2, 64), dtype=float32, numpy=
array([[-3.72067690e-02,  2.35418938e-02,  3.17604467e-03,
        -1.58705562e-03,  1.04783401e-02, -6.85185194e-03,
        -1.10574588e-02, -2.70983707e-02, -2.74406746e-03,
        -2.83305179e-02, -4.43014391e-02, -4.74758074e-03,
         3.40985693e-02, -1.29303560e-02,  3.66448052e-02,
        -1.66088119e-02, -4.94696274e-02,  1.54838823e-02,
         7.46102259e-03,  2.92506926e-02, -2.57844087e-02,
         3.24547924e-02,  4.47853468e-02, -2.96919942e-02,
         3.88157107e-02, -4.24836874e-02, -1.44682750e-02,
         3.88641618e-02,  3.66107561e-02,  3.18094604e-02,
        -7.78632239e-03,  4.58587147e-02, -4.54711430e-02,
        -1.79902464e-03, -3.51481922e-02,  8.54960829e-03,
        -4.30772677e-02,  3.89744304e-02,  2.00498216e-02,
         1.00603476e-02,  4.05639410e-03, -3.30186374e-02,
         4.87604178e-02, -1.93628203e-02, -2.20450759e-02,
        -4.04535048e-02, -9.17077065e-04,  4.17115353e-02,
         4.78942730e-02, -1.35492459e-02,  2.64997222e-02,
         2.23558210e-02,  4.24204953e-02, -3.46097574e-02,
         3.87491472e-02,  7.43114948e-03,  3.39420773e-02,
         4.35891263e-02, -4.01272774e-02, -7.46065378e-03,
         4.15328853e-02, -1.20298378e-02,  2.39137746e-02,
        -2.20017321e-02],
       [ 7.23880529e-03,  6.56664371e-04,  9.99033451e-03,
        -4.67067361e-02, -5.70065901e-03, -3.92155163e-02,
```

Word2Vec

A continuación se entrenan los parámetros obtenidos con **Word2vec** que consiste en generar *skip-grams* a través de una lista de oraciones. Estos *skip-grams* son pares de palabras (target, context) donde a target se le asocia context que puede o no encontrarse en su contexto (*skip-grams* positivo o negativo). Se busca generar para cada palabra un *skip-grams* positivo y **n** negativos, y usarlos para entrenar un *embeddings* personalizado.
Una vez hecho el entrenamiento con **Word2vec** se hicieron disponibles los *embeddings* en los archivos **metadata.tsv** y **vectors.tsv**.

Pretrained

Dado que en el conjunto de datos se encuentran títulos en español y portugués, se utilizaron word embeddings preentrados con el modelo **FastText**.