Categorización de Publicaciones de Mercado Libre

Integrantes:

- Benjamín Ocampo
- Eduardo Barseghian
- Maximiliano Tejerina

Recap del problema

- Dataset de 650000 publicaciones etiquetadas en 20 posibles categorías.
- Publicaciones pueden ser en ***español*** o ***portugués***.
- 15% de las publicaciones calificadas como ***reliable***. El resto fueron marcadas como ***unreliable***

title	label_quality	language	category
Galoneira semi industrial	unreliable	portuguese	SEWING_MACHINES
Maquina De Cocer Brother Industrial	unreliable	spanish	SEWING_MACHINES
Heladera Gafa 380 Impecable Urgente	unreliable	spanish	MUSICAL_KEYBOARDS
Maquina de Cortar el Pelo. Starex	reliable	spanish	HAIR_CLIPPERS
...	...	...	...

Preprocesamiento

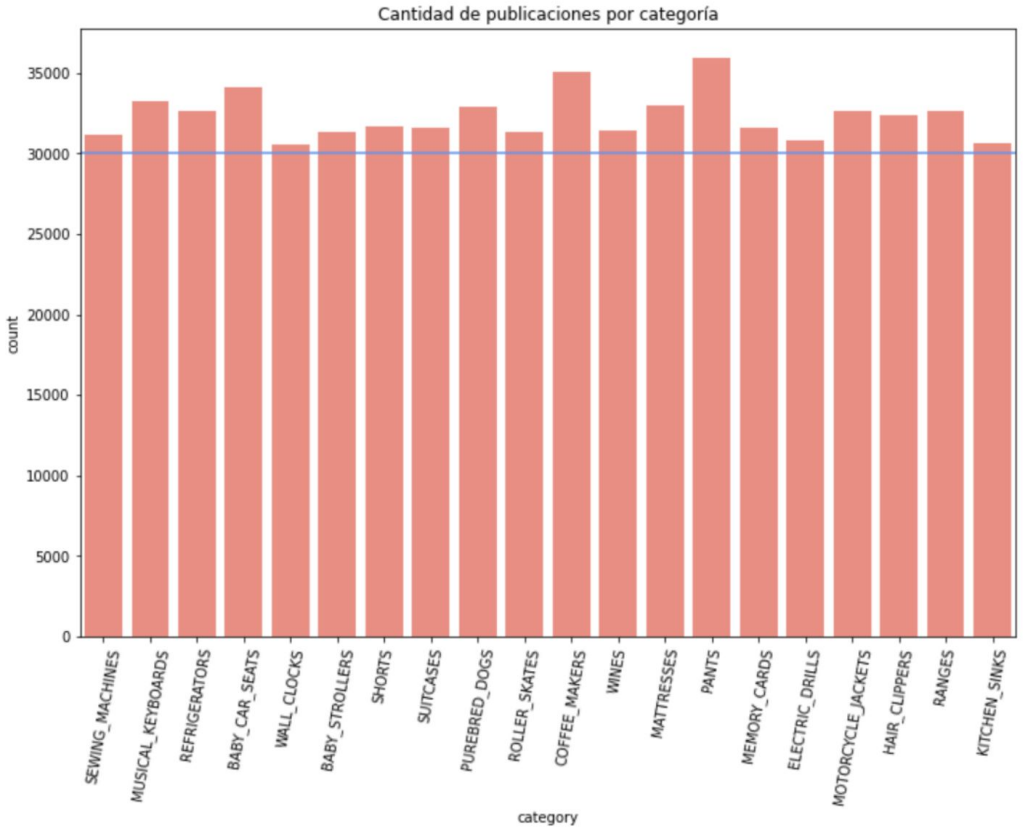
- Limpieza de la columna ***title***.
- Codificación de la columna ***category***.
- Almacenamiento del conjunto de datos en: "https://www.famaf.unc.edu.ar/~nocampo043/ML_2019_challenge_dataset_preprocessed.csv"

cleaned_title	encoded_category
galoneira semi industrial	15
maquina de cocer brother industrial	15
heladera gafa 380 impebable urgente	9
maquina de cortar el pelo starex	4
...	...

Balanced Accuracy

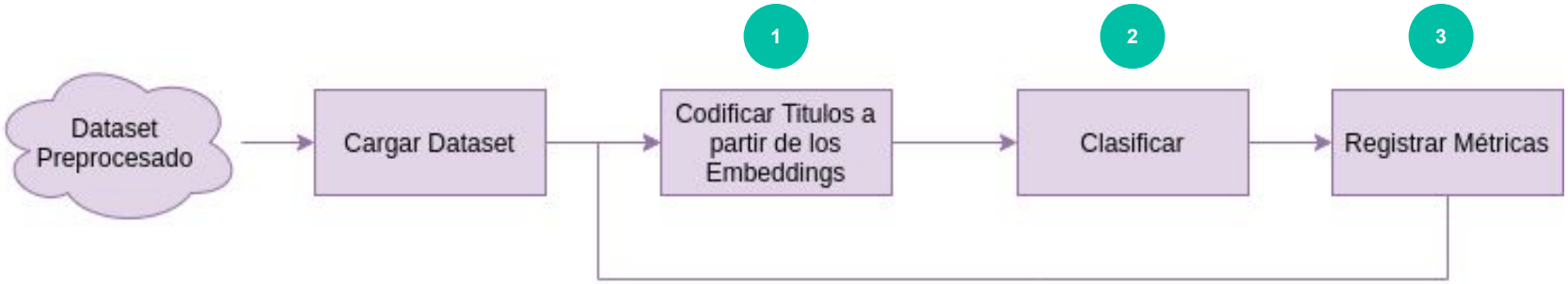
Estamos ante un problema de clasificación multiclase. Se trabajó con la métrica de Balanced Accuracy pues era la métrica del MeLi Challenge 2019.

Sin embargo para nuestro dataset filtrado pudo haber bastado accuracy pues estaban balanceadas las categorías.



Pipeline

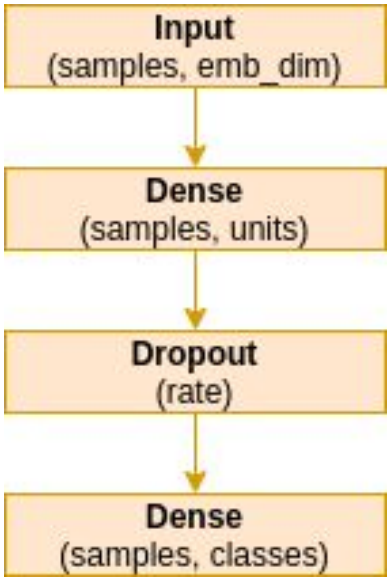
1. Embeddings
2. Clasificación
3. Registro de Resultados



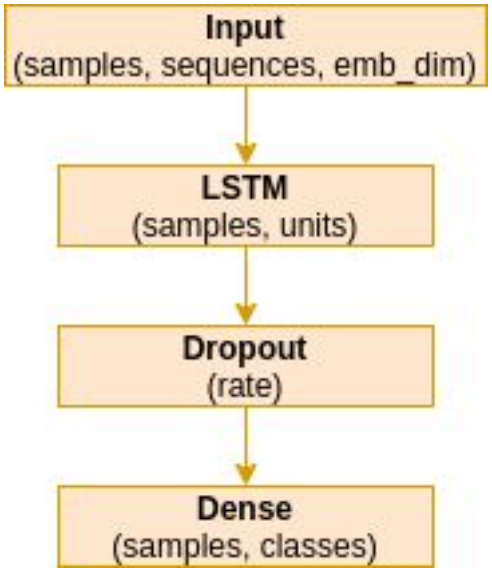
Modelos utilizados

- Regresión Logística ***(Baseline)***
- Red neuronal Feed Forward
- Red neuronal LSTM

Feed Forward

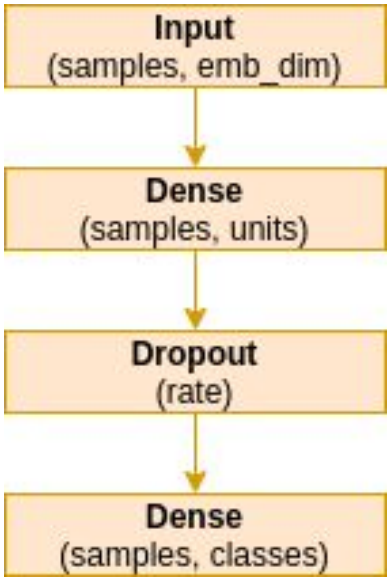


LSTM



Red Feed Forward

Por cuestiones de cómputo, fue corrida con un solo conjunto de (hiper)parámetros. Se le aplicó una 5-fold **Cross validation**.



samples = **20000** elementos
emb_dim = 100
batch_size = 100

units = 256 ; activación **ReLu**

dropout = 0.4

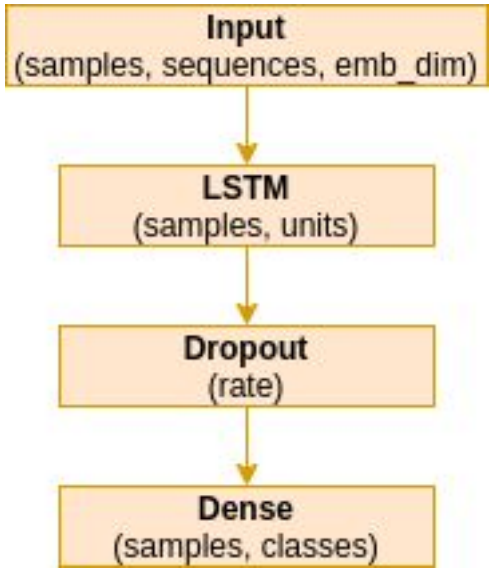
classes = 20 ; activación **SoftMax**

La función de pérdida elegida fue la **sparse_categorical_crossentropy**, con optimizador Adam (learning rate = 0.1)

Se corrieron **100 epochs**.

Red LSTM

En esta red, se entrenó sin CV y Hyper Tuning. **sentences** es la longitud del título más largo del conjunto de samples.



samples = **20000** elementos
emb_dim = 100

units = 256

dropout = 0.4

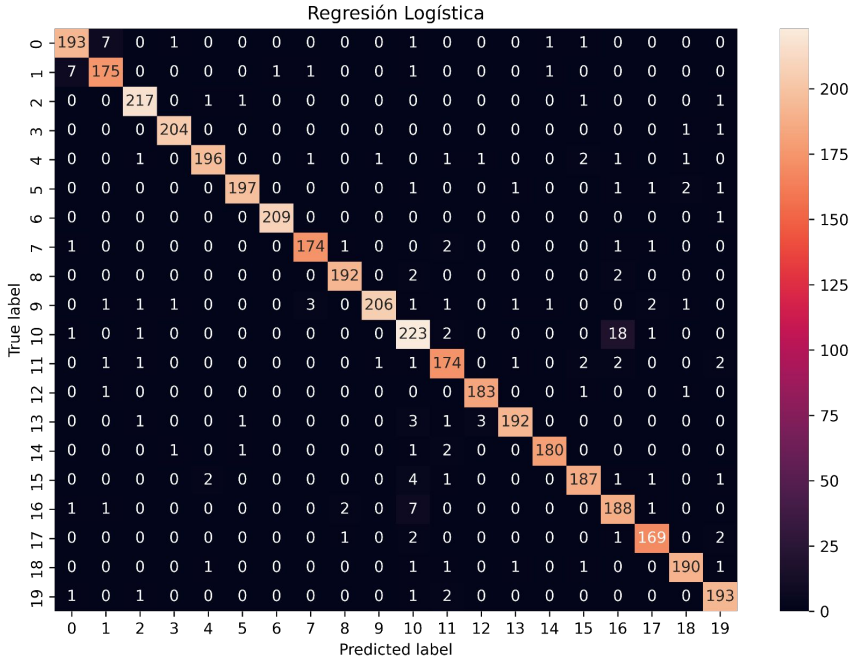
classes = 20 ; activación **SoftMax**

El **batch_size** elegido fue **100.**

La función de pérdida elegida fue la **sparse_categorical_crossentropy**, con optimizador Adam (learning rate = 0.1)

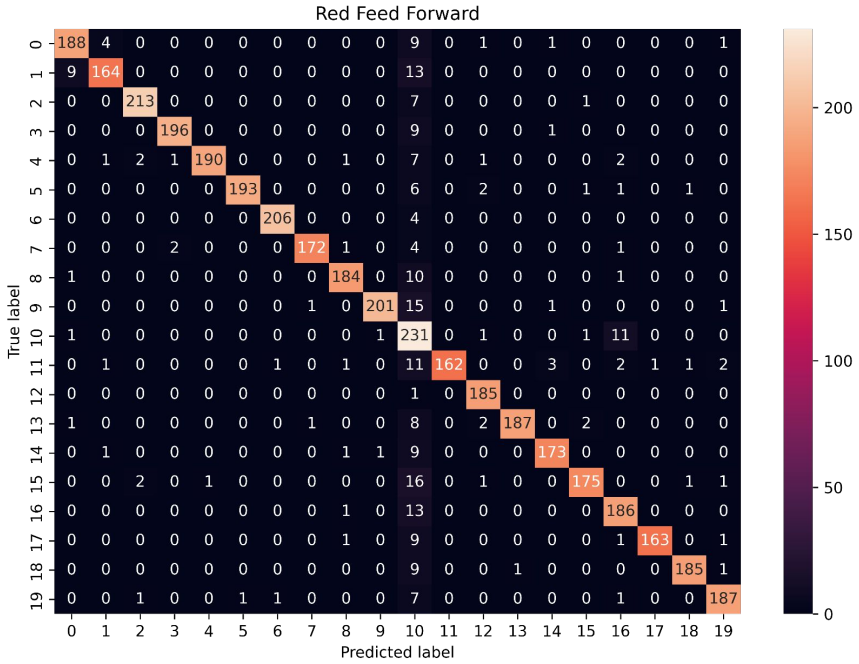
Se corrieron **100 epochs**.

Regresión Logística



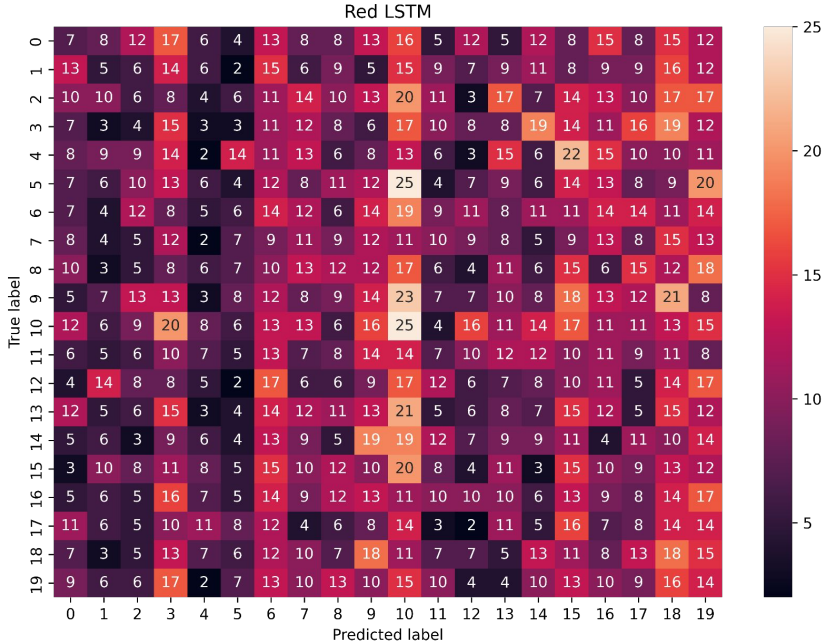
fold 0	0.958
fold 1	0.9628
fold 2	0.9558
fold 3	0.9531
fold 4	0.9610
blc_acc	0.9349
blc_acc_rel	0.9530
blc_acc_unrel	0.9304

Red Feed Forward



fold 0	0.9647
fold 1	0.9633
fold 2	0.9609
fold 3	0.9575
fold 4	0.9631
blc_acc	0.9611
blc_acc_rel	0.9650
blc_acc_unrel	0.9598

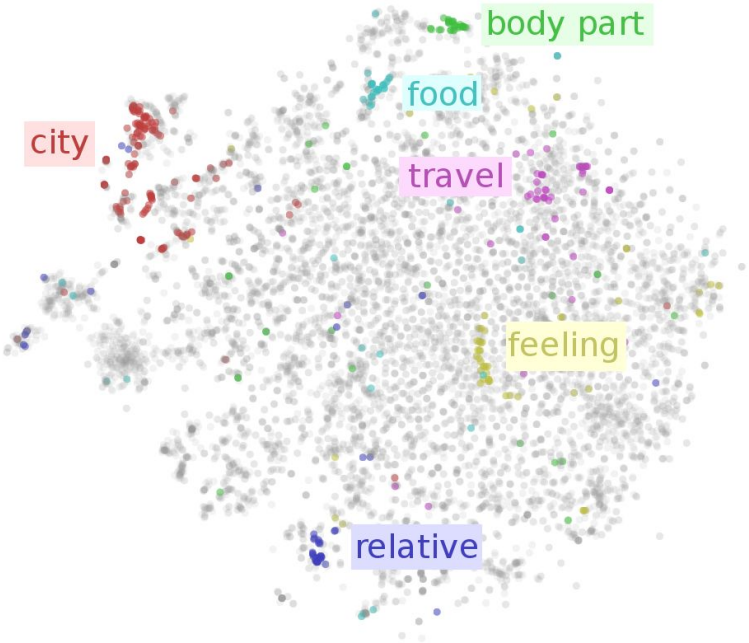
Red LSTM



blc_acc	0.05165
blc_acc_rel	0.05165
blc_acc_unrel	0.05254

No Supervisado

- Clustering utilizando K means.
- Embedding sobre el título.
- Encontrar el valor óptimo de clusters.
- Visualización.



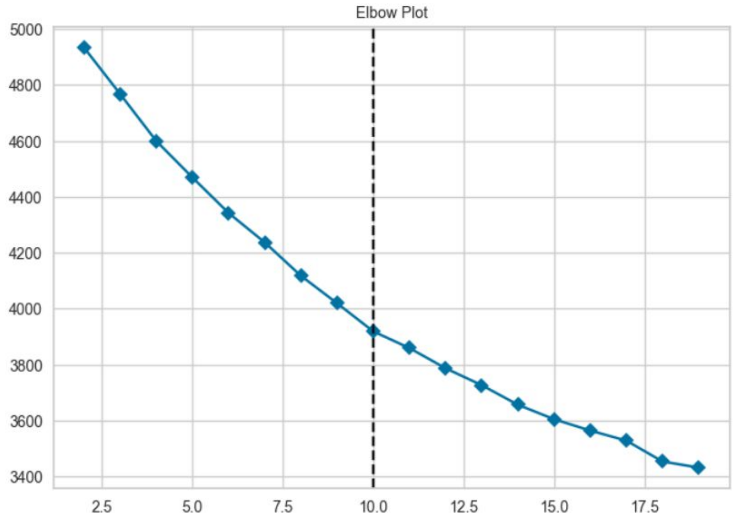
Clustering de títulos

- Se usó TSNE para las visualizaciones, y FastText para la codificación.
- Fasttext promedia las representaciones a nivel de subpalabras y caracteres.
- Palabras que están en distinto idioma, por ejemplo, la palabra ***barbero*** se ubica cerca de sus semejantes en español y portugués.

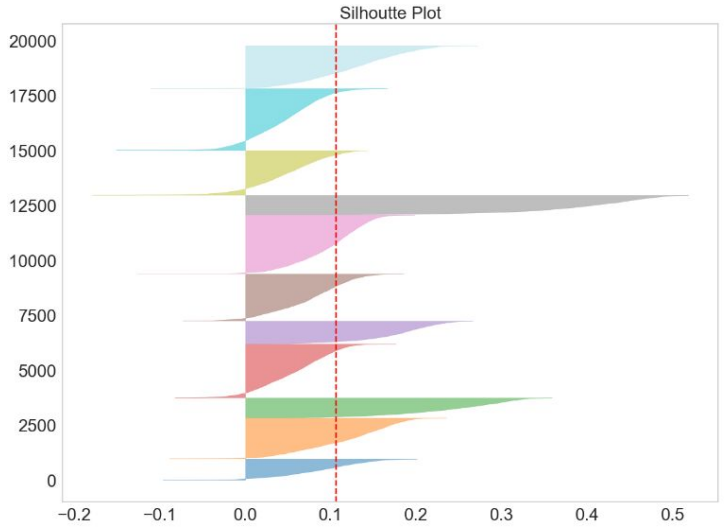
```
[(0.6704239845275879, 'barber'),
 (0.582217276096344, 'barberia'),
 (0.41741418838500977, 'lijadora'),
 (0.40764909982681274, 'cortadora'),
 (0.40148794651031494, 'hdk'),
 (0.3844088912010193, 'cortadoras'),
 (0.34548690915107727, 'gamma'),
 (0.3447607457637787, 'shaver'),
 (0.34402868151664734, 'clip'),
 (0.33779293298721313, 'imetec')]
```

Método del Codo

- Nos proporciona el número óptimo de clusters
- No logramos obtener un codo pronunciado a pesar que el algoritmo indique que el óptimo es con 11 clusters.



KMeans y Coeficiente de Silhouette



Trabajo a Futuro

- Repetir los procesos de Aprendizaje Supervisado y No Supervisado con todo el conjunto de datos.
- Implementar pipelines para ser ejecutados en computadores dedicadas.
- Probar estos modelos participando en la competencia.
- Probar otros parámetros en FastText para mejorar la clusterización.

Gracias!