# Clasificación de opiniones de hoteles mediante Perceptrón y Redes Neuronales

Álvaro Hernández

Escuela de Ingeniería de Bilbao / UPV-EHU (December 5, 2021)

## Objetivos

- Dada una opinión en forma de texto, predecir la puntuación numérica asociada a la opinión, con valores del 1 al 5.
- Preguntas a investigar:
  Pregunta 1 ¿Qué representación funciona mejor?
  1. TF-IDF
  2. Word Embeddings

  Pregunta 2 ¿Qué clasificador funciona mejor?
  1. Baseline: Perceptrón simple
  2. Redes neuronales

## Tarea y Data

**Tarea**

- Dado como input (texto) una opinión, predecir la puntuación que le corresponde.
- Fuente de los datos: [Kaggle, 2020]
- Cada instancia está representada por la opinión (Review) y la puntuación otorgada (Rating).

## Representaciones del texto

- **Preproceso:** Los pasos que se han dado en el preproceso han sido los siguentes:
  - Eliminación de Stopwords
  - Eliminación marcas de puntuación
  - Corrección letras y números
  - Lemanización
  - Corrección de palabras rotas
  - Eliminación (por segunda vez) de Stopwords

| Preproceso | Docs | Esp atrib | Nº Clases |
|---|---|---|---|
| Raw | 20491 | 54000 | 5 |
| TF-IDF (freq = 50) | 20491 | 3300 | 5 |
| Word Embedding | 20491 | 300 | 5 |

Table 1:Data: descripción cunatitativa

- **Representation.** Se han utilizado dos representaciones distintas. Para TF-IDF, se ha reducido el espacio vectorial tras haber conseguido el mejor resultado F-Score con 2300 atributos. Con Word Embedding se ha mantenido el espacio de 300, por ser el más común.

## Clasificador 1: Perceptrón simple

Ya que después se utilizará la red neuronal, como aproximación se ha utilizado el perceptrón, que es la base de las redes neuronales.

1. Se ha utilizado la librería **Scikit-Learn**, de modelos lineales.

## Clasificador 2: Redes neuronales

Son clasificadores complejos y que requieren de alta capacidad de computo.

- En esta ocasión también se ha utilizado la librería Scikit-Learn, en particular las librería **neural_network** y **RandomizedSearchCV**.
- Se ha realizado una hiperparametrización (**learning rate, número de iteraciones, tamaño de la(s) capa(s) oculta(s)**).

## Experimentación para dar respuesta a las *Preguntas Objetivo*. Primer acercamiento

El objetivo era conseguir la combinación de clasificador y representación vectorial que diese mayor F-Score En todos los casos se han hecho dos tipos de evaluciones. Primeramente, se han obtenido las matrices de confusión de una evaluación **Hold-Out (70-30)**.

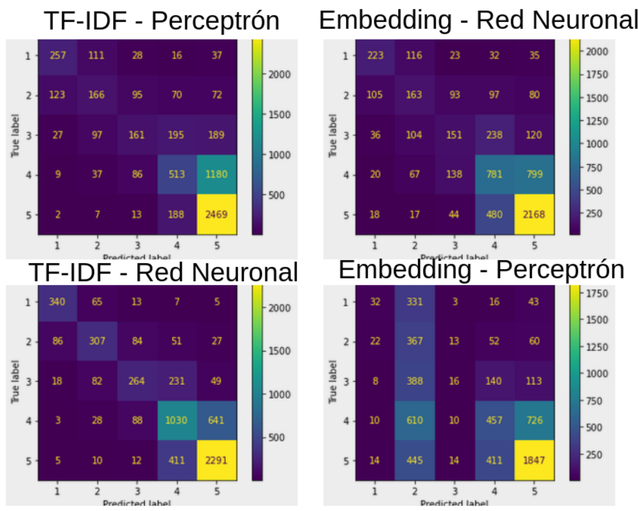


Figure 1:Matriz de confusión de todas las combinaciones

Así se puede saber cuantos aciertos tiene cada combinación clasificador-representación para el conjunto de testeo.

| Combinación | Accuracy (Test - Hold-Out) |
|---|---|
| Per - TF-IDF | 0.58 |
| Per - Embedding | 0.49 |
| RN - TF-IDF | 0.69 |
| RN - Embedding | 0.57 |

## Experimentación final

Por otro lado, para la toma de decisión final se ha realizado una evaluación cruzada K-Fold. El criterio para decidir la mejor combinación será el mayor valor F-Score.

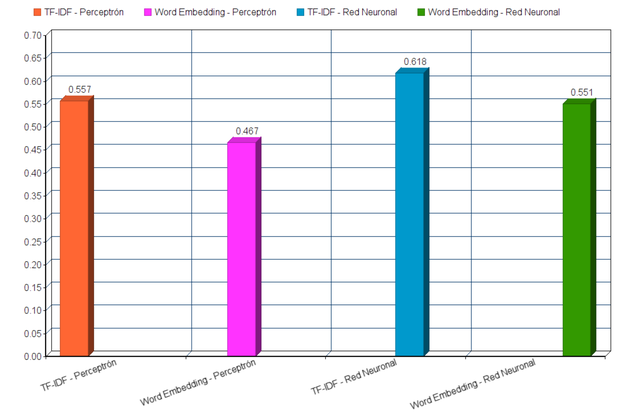


Figure 2:F-Score de las combinaciones

**Resultados**

- En la aproximación y en los resultados finales TF-IDF ha sacado mejores resultados.
- A su vez, el clasificador de la red neuronal ha funcionado mejor que el perceptrón simple

## Conclusiones

- **Fortalezas**: Se ha logrado hacer una buena parametrización para todos los clasificadores.
- **Debilidades**: Un dataset con 5 clases no muy bien diferenciadas es un dataset complicado para clasificación.
- **Trabajos futuros**: Para la próxima ocasión se intentará trabajar con un dataset en el que las clases estén mejor diferenciadas para ver la fortaleza de las redes neuronales.

## Bibliography

[Kaggle, 2020] Kaggle, Larxel, k. g. (2020). Trip advisor hotel reviews.