Clasificacion de Texto Sobre la Corte Suprema

Gorka Cidoncha Marquiegui

UPV/EHU (December 25, 2022)

Objetivos

- Objetivo: Dados lo hechos de la Corte Suprema (input: Texto plano) predecir el ámbito judicial con el mayor F-Score posible.
- Preguntas de Investigación:
  - RQ1 Cual es la mejor Representación?
    1. TF-IDF
    2. Topic Modeling
  - RQ2 Cual es el mejor Clasificador?
    1. Decision Tree
    2. SGDClassifier

Representacion del Texto [RQ1]

- **Pre-Procesado**

Pre-Procesado	Docs	Vocab	Classes
Borrar Nulls Stop-words y Minúsculas Lematizar	3161	25333	14

Table 1:Descripción cuantitativa de los Datos

- **Representación**

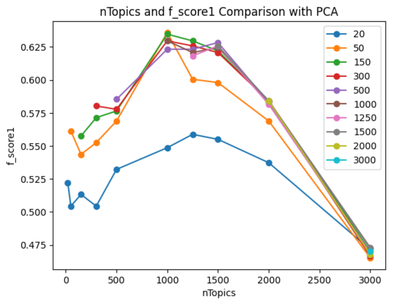

Figure 1:Gráfico de relación nTopics PCA.

Score	TF-IDF	Topic-Modeling
F-Score	0.549	0.628
Weighted F-Score	0.45	0.61

Table 2:Puntuación máxima de las diferentes representaciones

Se puede observar como el uso del Topic Modeling da mejores resultados sobre los 1000 Tópicos con 150 de PCA.

Clasificador No supervisado y Aleatorio

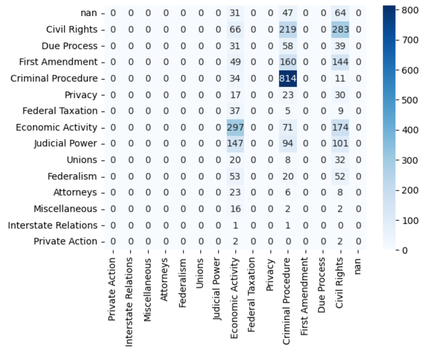

Figure 2:Matriz de Confusión de el Clustering Jerárquico

Score	No Supervisado	Aleatorio
F-Score	0.422	0.07
Weighted F-Score	0.274	0.09

El clasificador no supervisado nos da buenos resultados comparándolos con el clasificador aleatorio.

Clasificador Arboles de Decisión

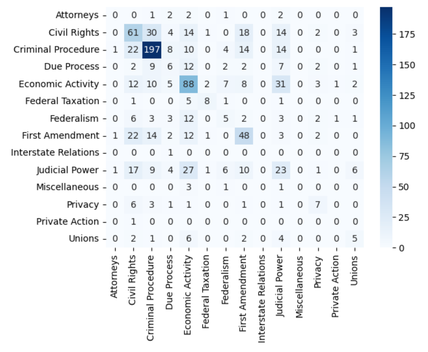

Figure 3:Matriz de Confusión de el Clasificador por Arboles de Decisión

- F-Score: 0.48
- Weighted-F-Score: 0.48
- Los resultados son notablemente mejores sobre todo para la puntuación equilibrada.

Clasificador SDGC

Para encontrar los mejores parámetros para el clasificador se hace un barrido de estos.

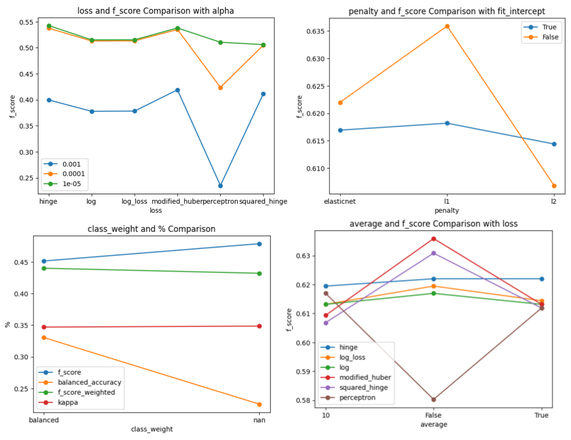

Figure 4:Barrido de parámetros con F-Score

- Gracias a los resultados obtenidos con el barrido usamos los siguientes parámetros:

Mejores Parámetros (Modelo1):

- Loss: Huber-Hinge
- Alpha: 0.001
- Penalty: l1
- Fit-Intercept: False
- Learning Rate : Adaptative
- Eta0 : 1.0

Modelo	F-score	F-Wighted	BalancedAcc
Modelo1	0.612	0.58	0.361
Modelo2	0.503	0.52	0.45

Table 3:Puntuación de los modelos obtenidos

- Gracias a los resultados obtenidos de sacan dos modelo.
- El primero y que se ha tomado con el mejor tiene un f-score muy alto, pero no esta equilibrado.
- El segundo modelo da un f-score mas bajo pero al estar mas equilibrado se ha dado por bueno.

Extra SGDC

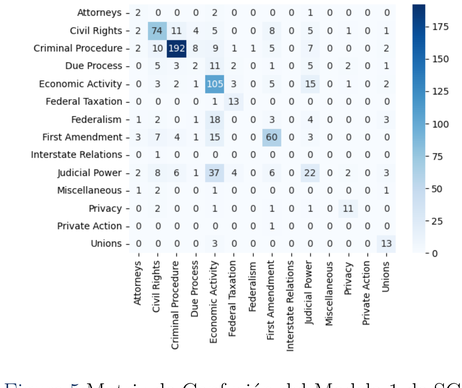

Figure 5:Matriz de Confusión del Modelo 1 de SGDC

Discusión

- Como era de esperar los Clasificadores supervisados han dado mejores resultados.
- La representación por Topic Modeling ha funcionado mejor que TF-IDF debido al gran numero de palabras de cada documento.

Conclusiones

- Los clasificadores supervisados suelen funcionar mejor en este tipo de problemas.
- La representación es igual de importante que el clasificador elegido.
- Hay que saber interpretar el barrido de Datos de forma adecuada.
- Al ser tan pocos documentos los resultados varían aun con los mismos parámetros, por lo que hay que contrastar los resultados.
- Se a tomado el f-score para la puntuación pero no significa que sea el único valor importante, dependiendo de lo que se desee predecir y como otros valores como la precision o el recall pueden cobrar mas importancia.

En este póster se representa los apartados mas importantes. Para mas resultdos y el codigo: Pyhon Notebook