Clasificación de textos basada en redes neuronales

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Curso 2020-2021

Autor: Mario Campos Mocholí
Tutores: Encarnación Segarra Soriano
Lluís Felip Hurtado Oliver
Emilio Sanchis Arnal

Índice

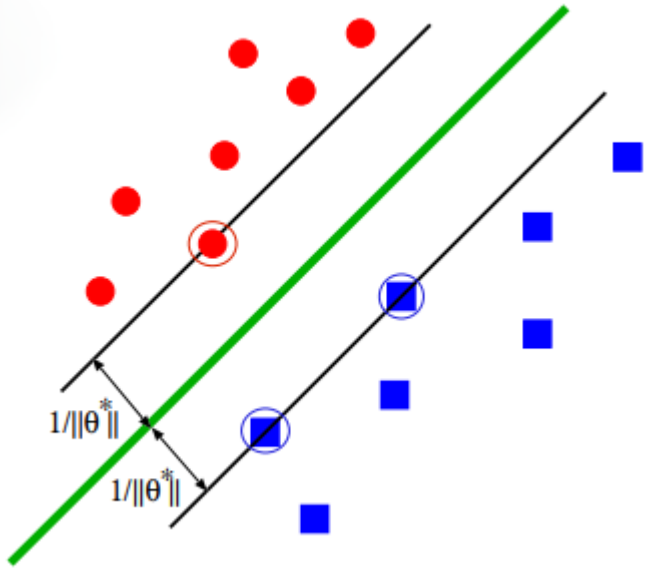
Índice

Motivación y objetivos

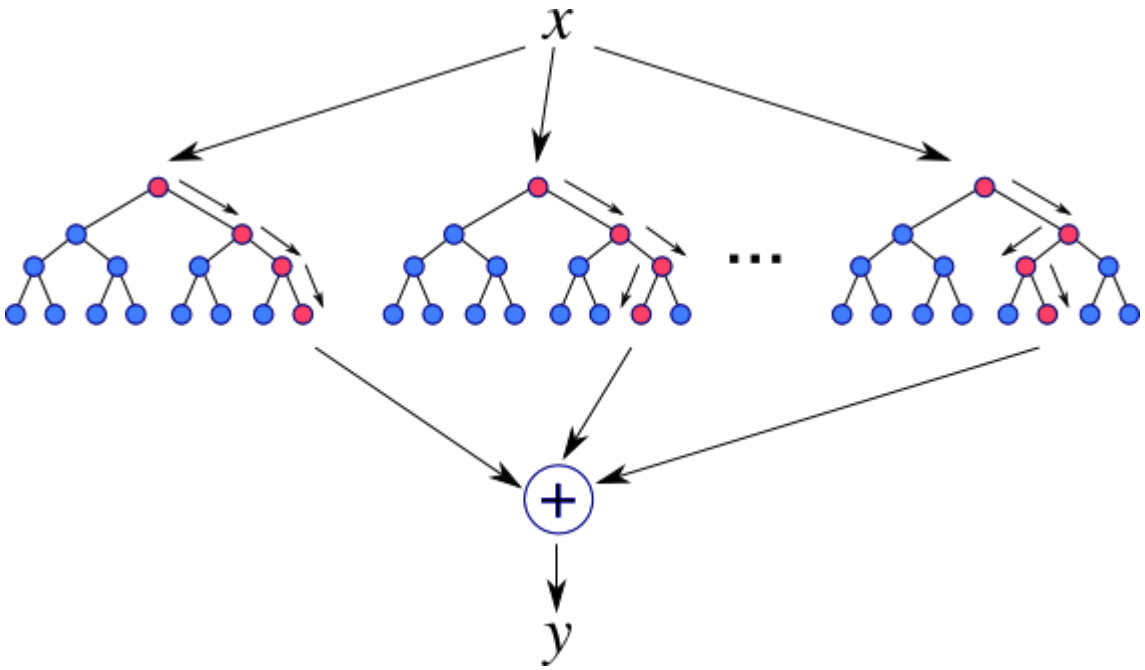
- Determinar si es posible realizar herramienta de apoyo a la catalogación para la CVMC
- Elaboración de un *corpus* único.
- Aplicar técnicas de NLP para textos en catalán.
- Representar el conjunto de noticias mediante *embeddings*.
- Evaluar modelos de clasificación automática sobre el conjunto de noticias.

Índice

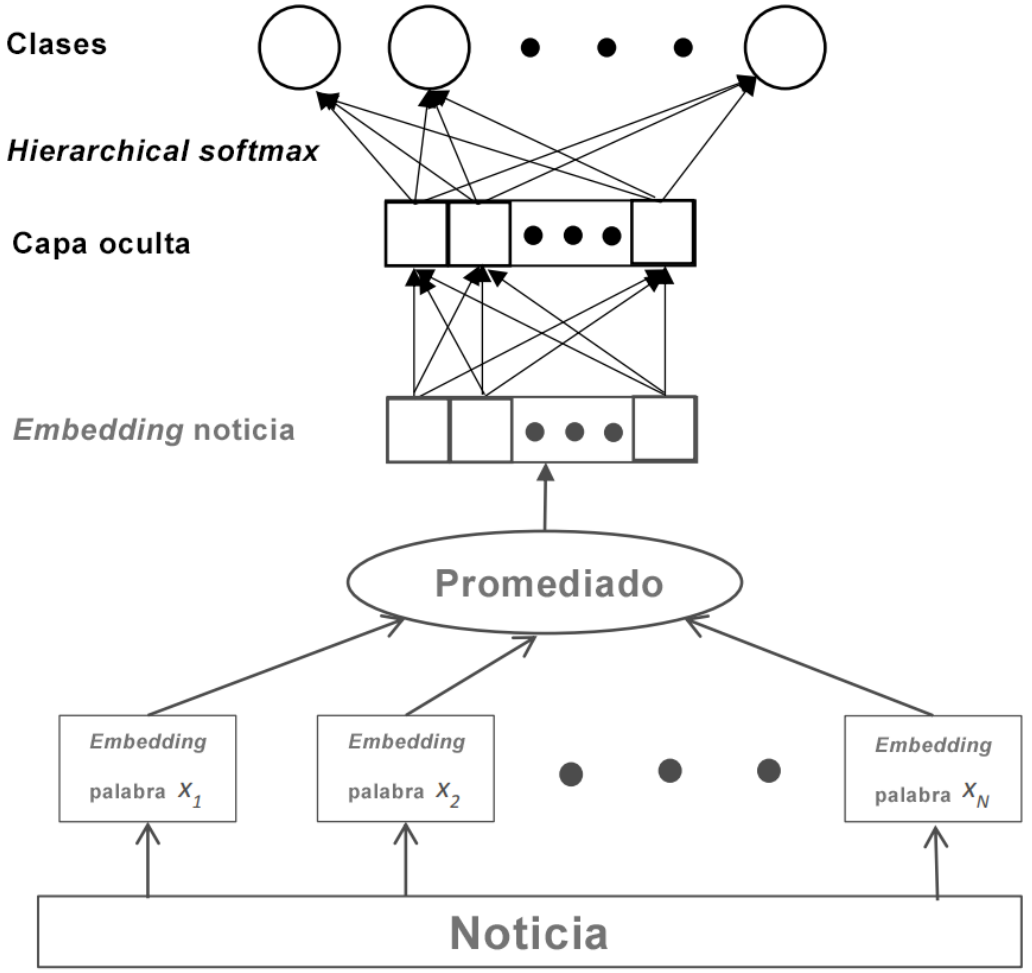
- Naive Bayes
- Support Vector Machine



- Random Forests



fastText


Clases
Hierarchical softmax
Capa oculta
Embedding noticia
Promediado
Embedding palabra x_1
Embedding palabra x_2
Embedding palabra x_N
Noticia

Índice

Frecuencia del término:

$$\mathrm{tf}(t,d) = \begin{cases} 1 + log(f(t,d)) & \text{si} \quad f(t,d) > 0 \\ 0 & \text{si} \quad f(t,d) = 0 \end{cases}$$

Frecuencia inversa del documento:

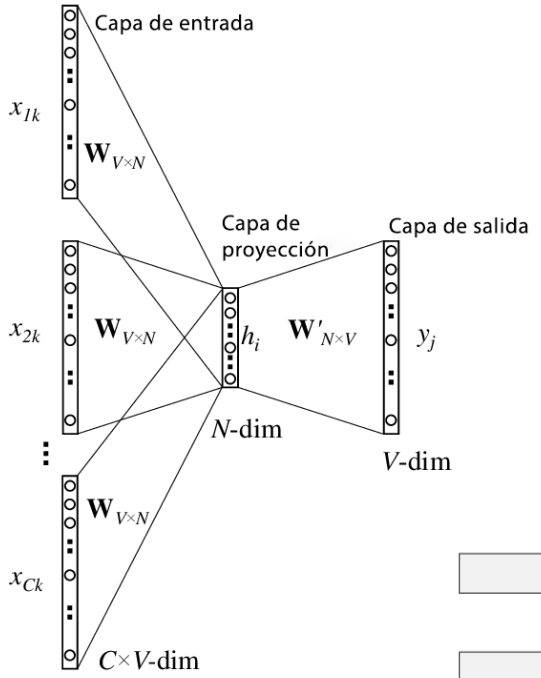
$$\mathrm{idf}(t,D) = log(\frac{|D|}{\mathrm{df}(t)})$$

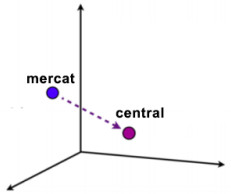
TF-IDF:

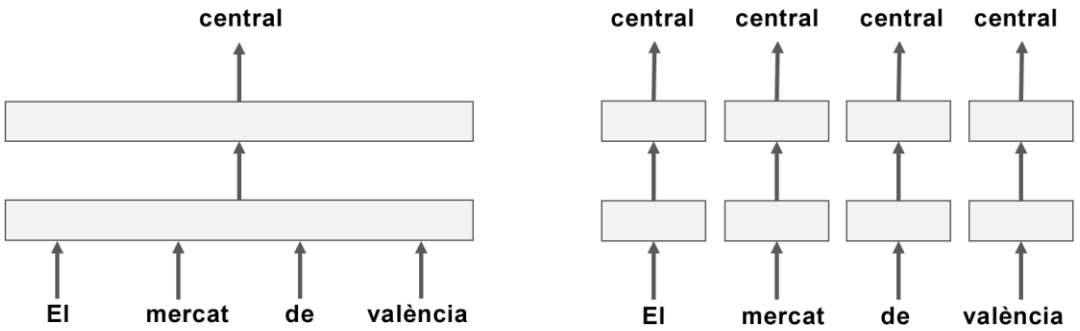
$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

fastText

CBOW

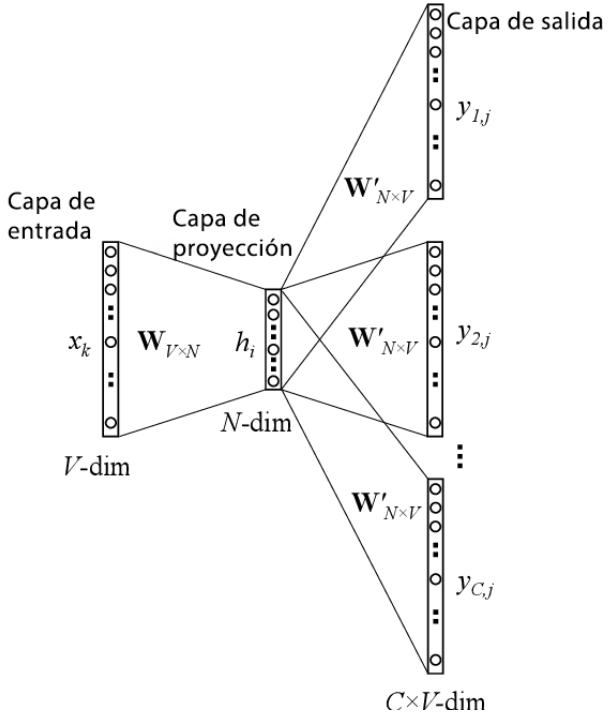






el mercat {central} de València

Skip-gram



Índice

- Conjunto de noticias de la CVMC.
- Datos desde 1999 a 2013 y desde 2018 a la actualidad.
- 21 ficheros Excel con 17 columnas, solo 2 útiles:
  - **Descripción**: contiene la noticia en sí, así como información adicional y ruido.
  - **Clasificación**: la clasificación interna de la *Corporació* de cada noticia.
- Necesario centralizar y procesar las noticias.

- *Corpus* en minúscula.
- Filtros de apóstrofes y *pronoms febles.*
- Filtro de *stopwords.*
- Filtros de ruido.

"((Entradeta))

La Marina de València es convertix des d'esta vesprada en el plató d''À Punt Directe', el magazín d'esta televisió. (*) Joan López i Sònia Fernàndez conduiran el programa, que dedicarà una atenció especial als protagonistes de les festes i dels festivals arreu del territori. Des del moll de ponent, i durant dos hores diàries, els presentadors dirigiran un equip de reporters que informaran de tota l'actualitat estiuenca a partir de les 5 i mitja de la vesprada.

((Durada, cua i peu text))

((Historia - OFF))"

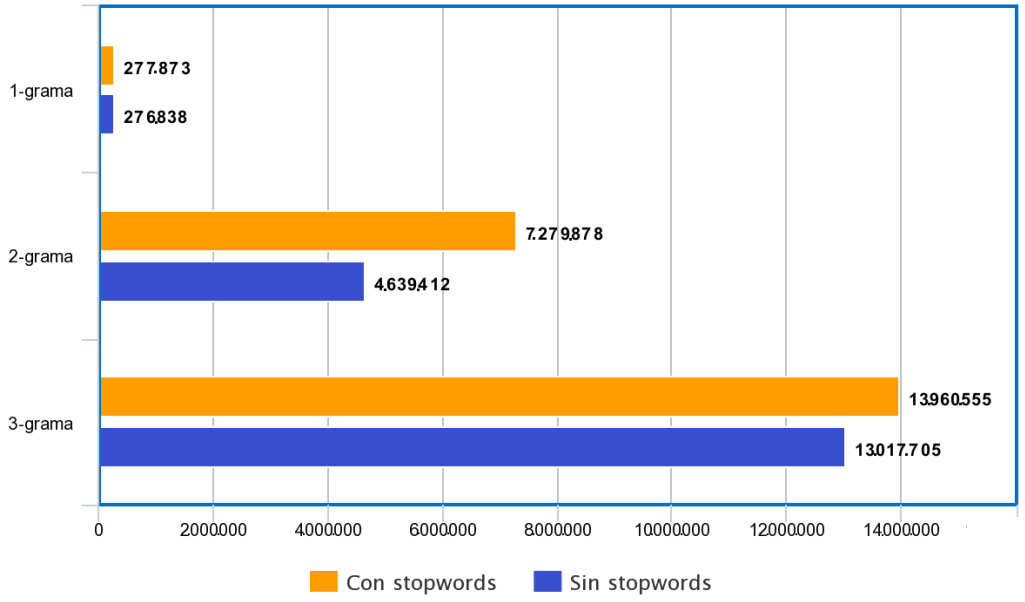
"la marina de València es convertix des esta vesprada en el plató à punt directe el magazín esta televisió joan lópez i sònia fernàndez conduiran el programa que dedicarà una atenció especial als protagonistes de les festes i dels festivals arreu del territorio des del moll de ponent i durant dos hores diàries els presentadors dirigiran un equip de reporters que informaran de tota actualitat estiuenca a partir de les i mitja de la vesprada"

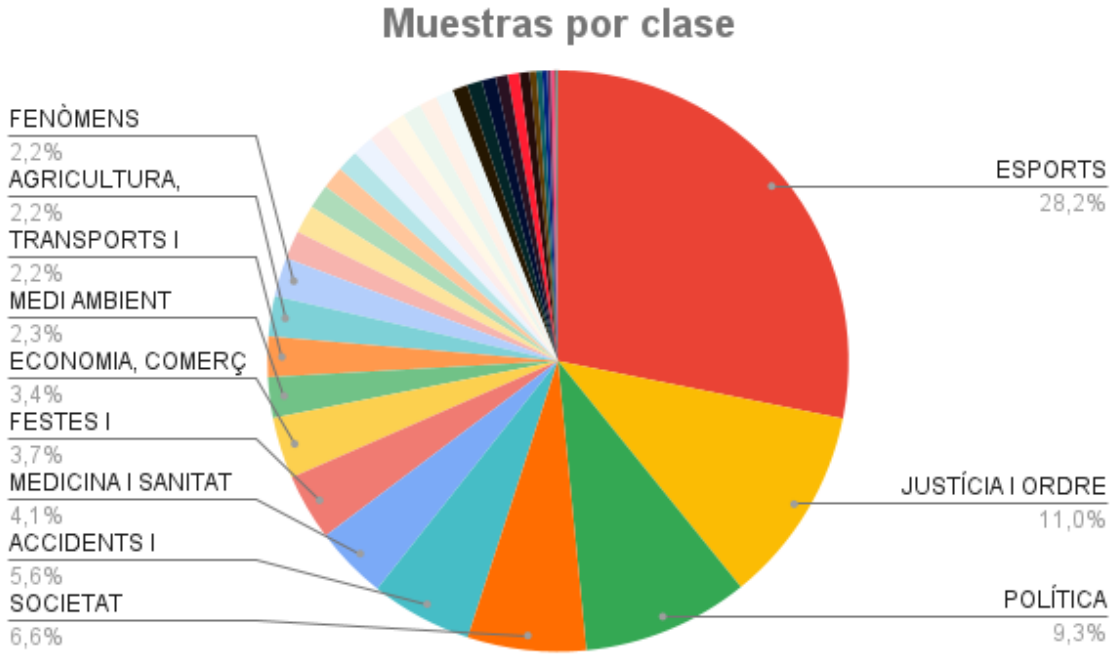
- Modelo de n-gramas.

Ej.: *"Les populars mascletaes de fogueres Alacant"*

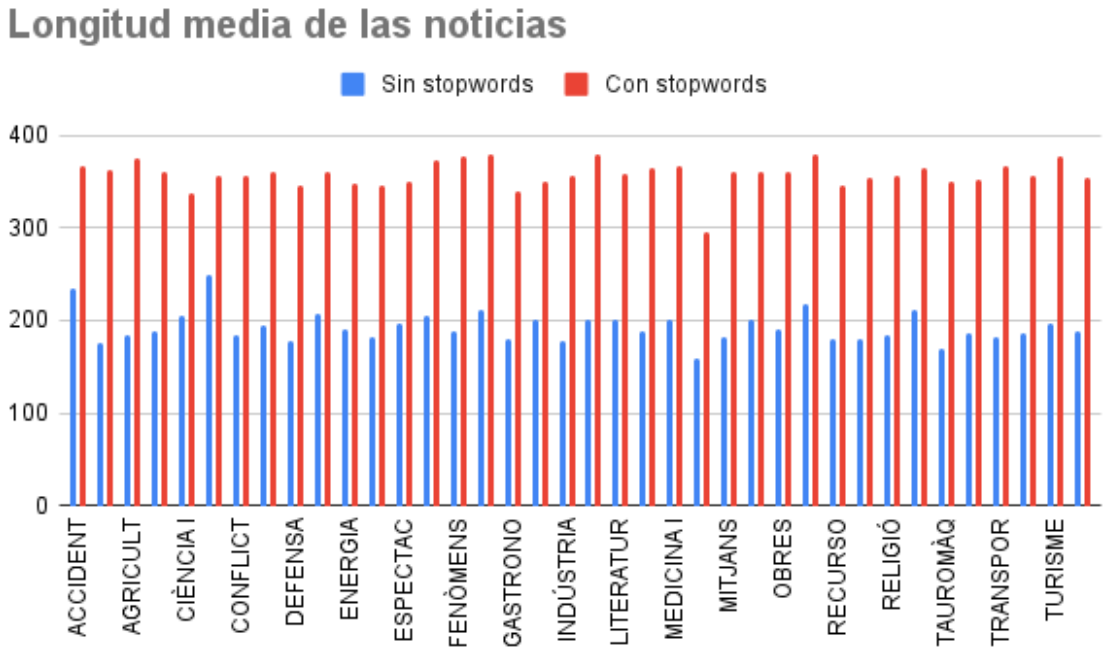
- **1-grama:** $, *Les, populars, mascletaes, de, fogueres, alacant*, $
- **2-grama:** *($ Les), (Les populars), (populars mascletaes), (mascletaes de), (de fogueres), (fogueres alacant), (alacant* $)
- **3-grama:** ($ *Les populars), (Les populars mascletaes), (populars mascletaes de), (mascletaes de fogueres), (de fogueres alacant), (fogueres alacant* $)

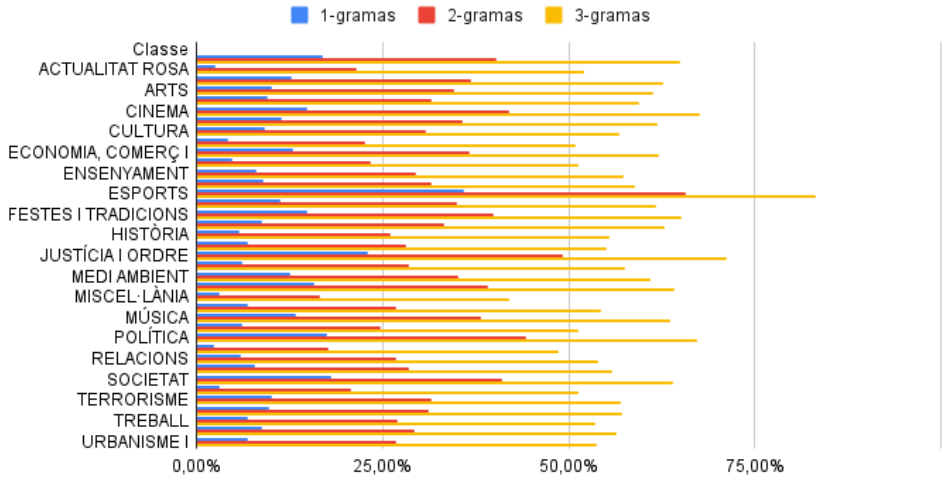
- Dos corpus resultantes, uno con *stopwords* y otro sin *stopwords*.
- Alrededor de un 60,80% de noticias útiles.
- Un total de 38 clases únicas.

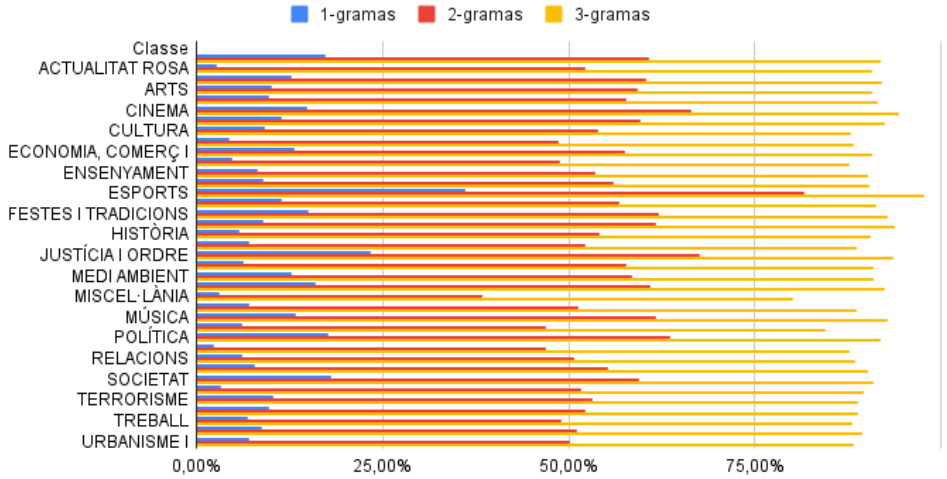




- La longitud de las noticias se reduce sin *stopwords*.
- Mayor número de n-gramas únicos sin *stopwords*.



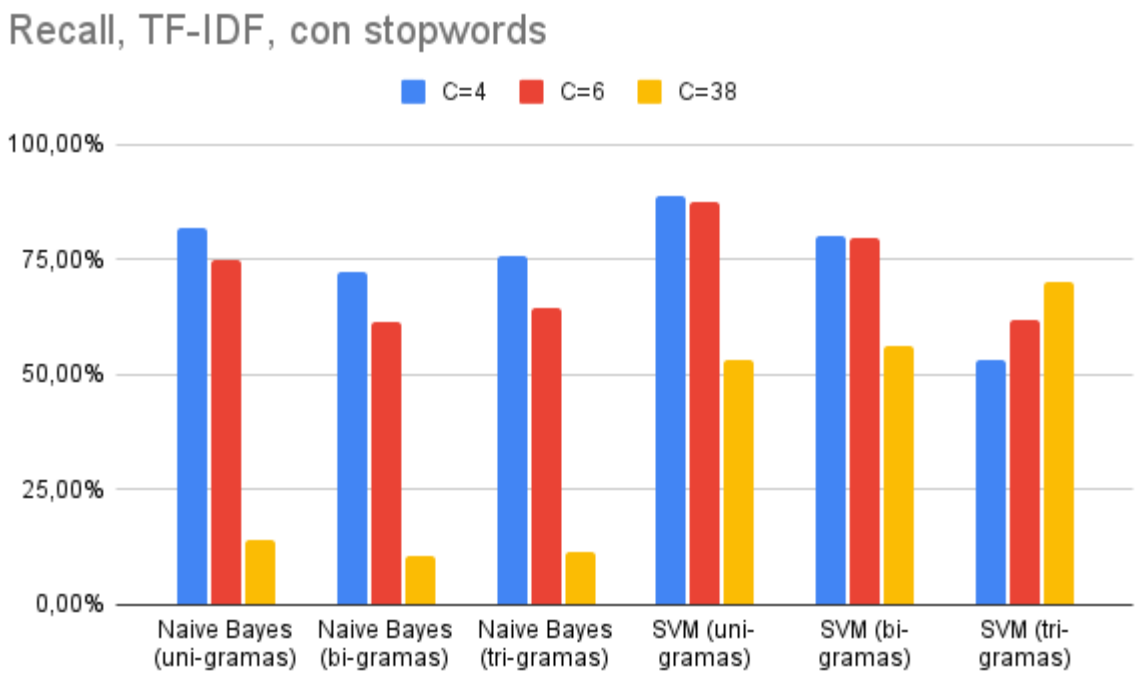


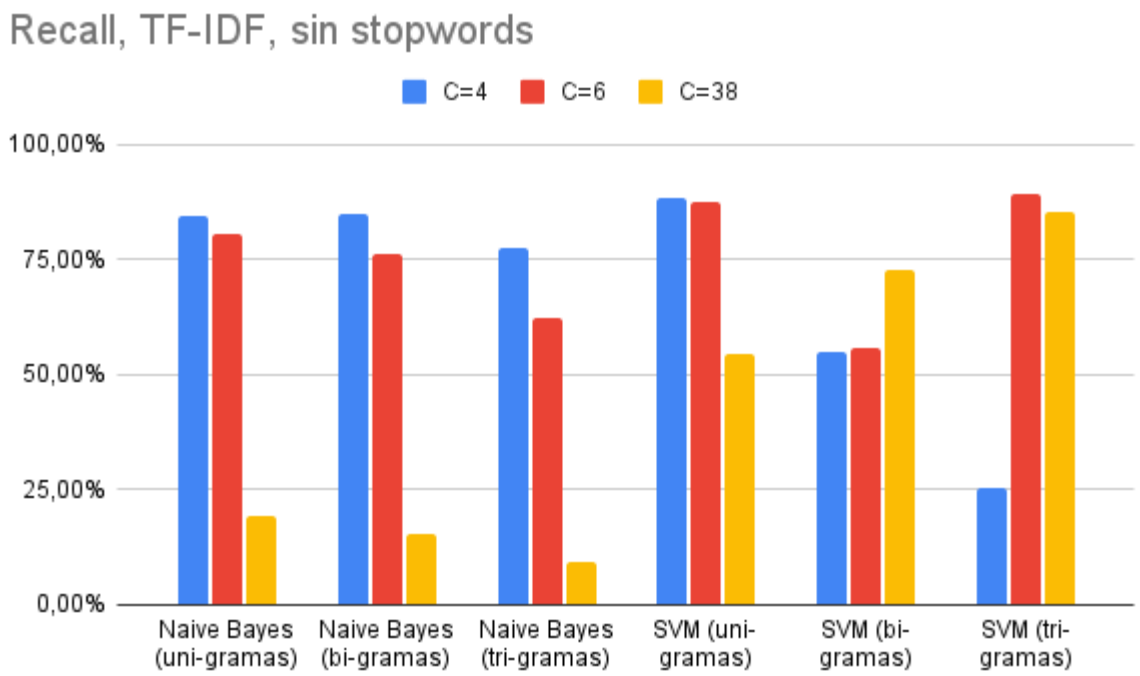


Índice

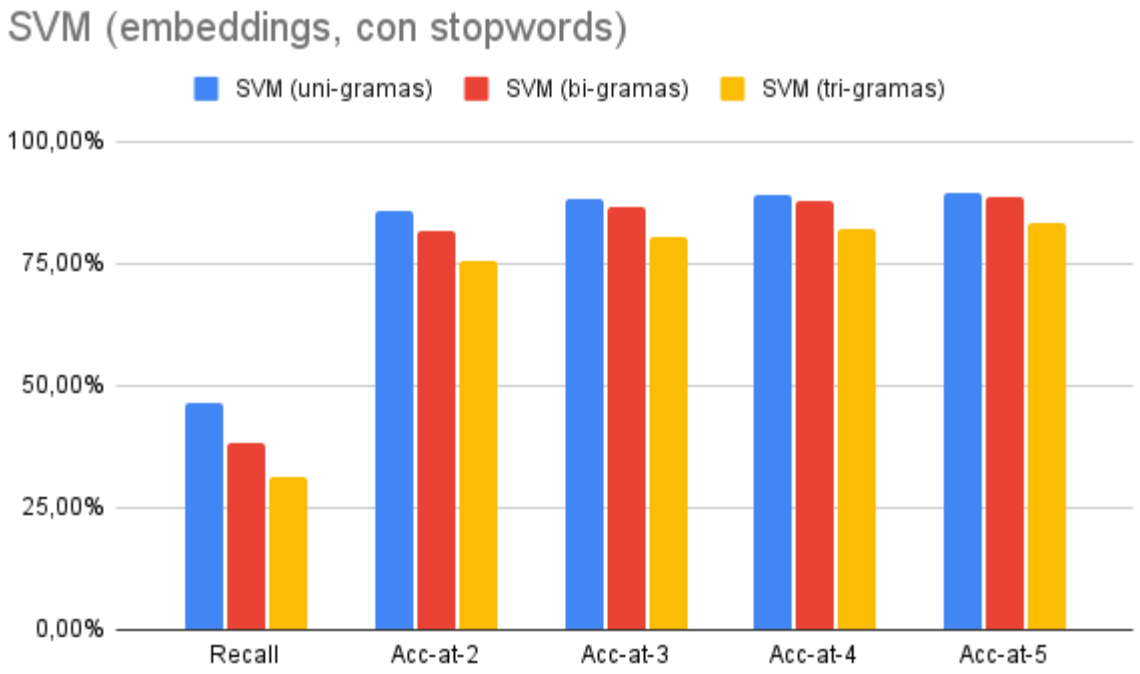
- Partición 70% para entrenamiento, 30% para test.
- Partición *stratified*, mismo porcentaje de cada clase en cada conjunto.
- Mismas particiones, independientemente de la representación utilizada y el uso o no de *stopwords*.
- Dos métricas utilizadas:
  - *Recall* macro
  - *Accuracy-at-k*

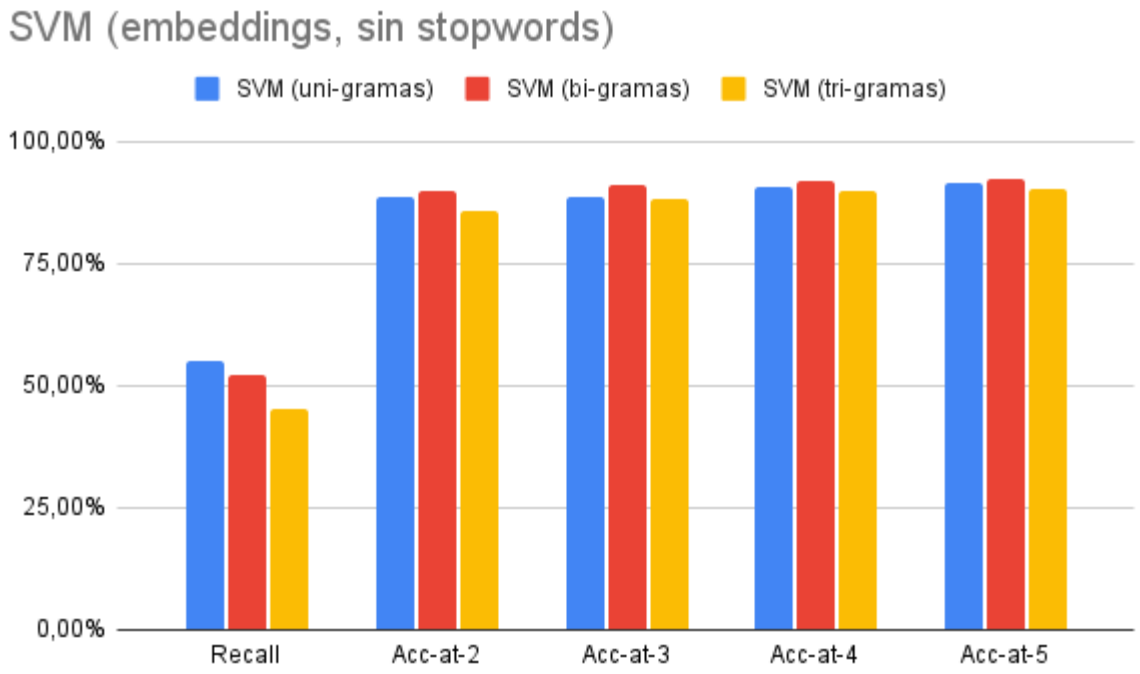
Resultados experimentales



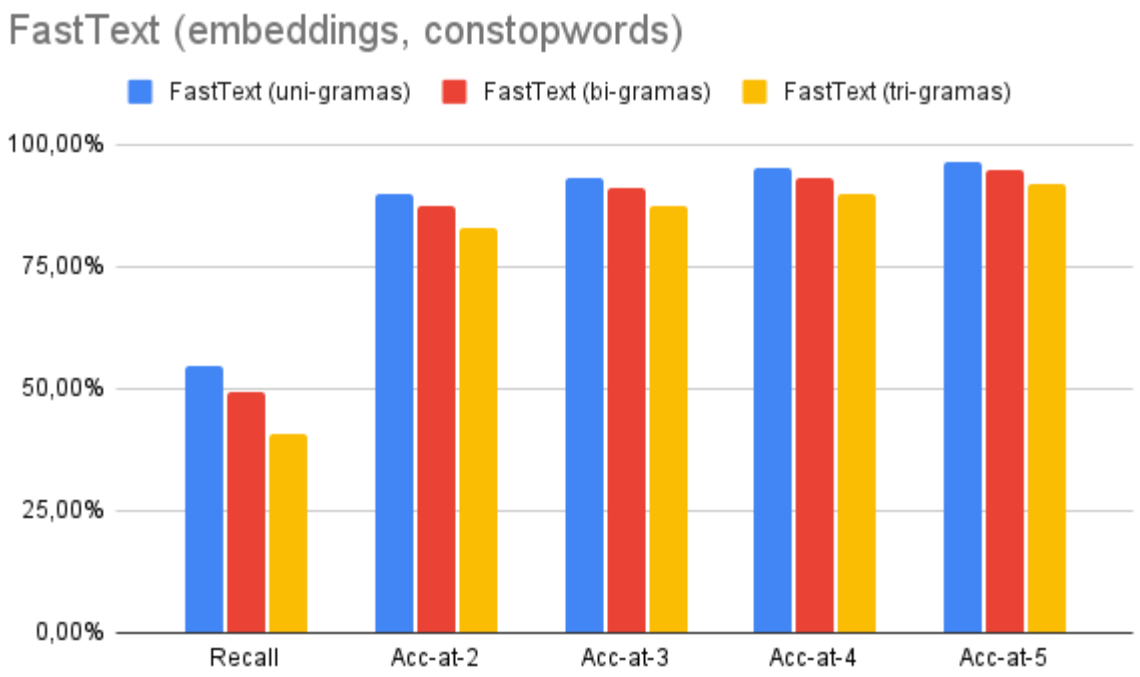


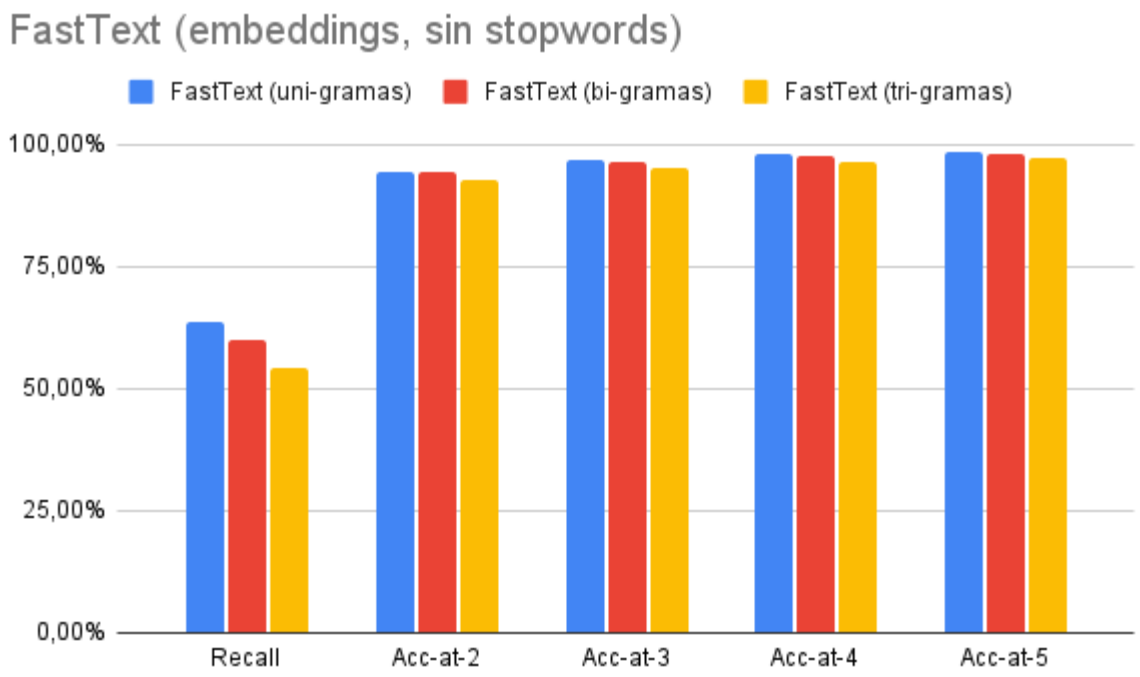
Resultados experimentales



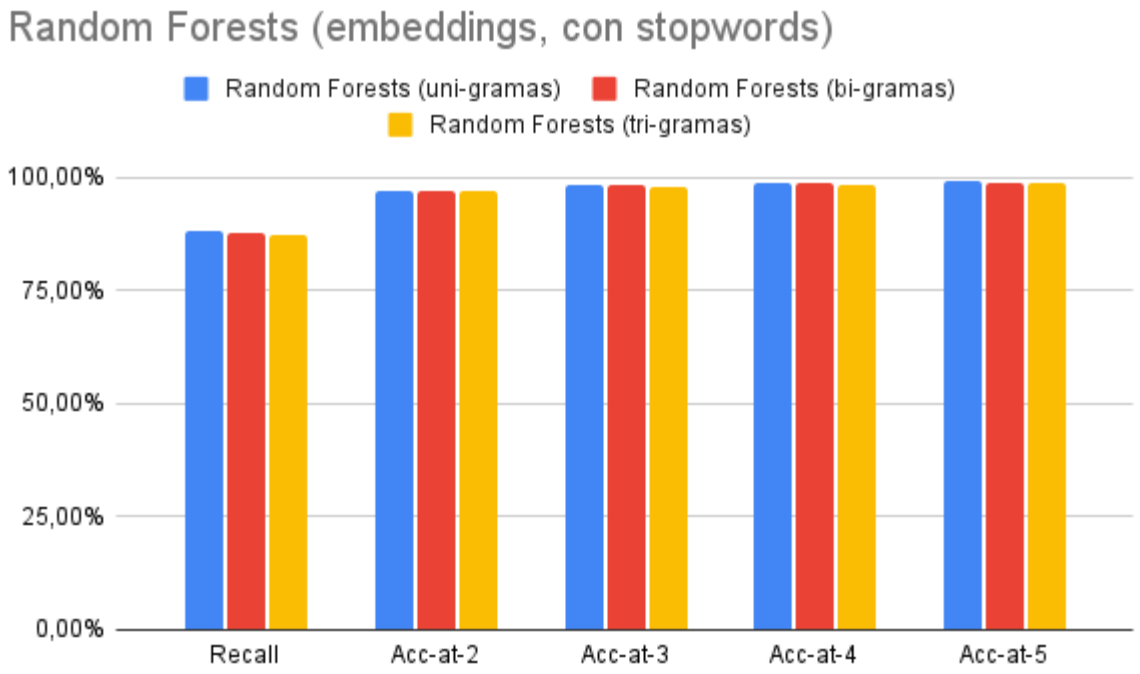


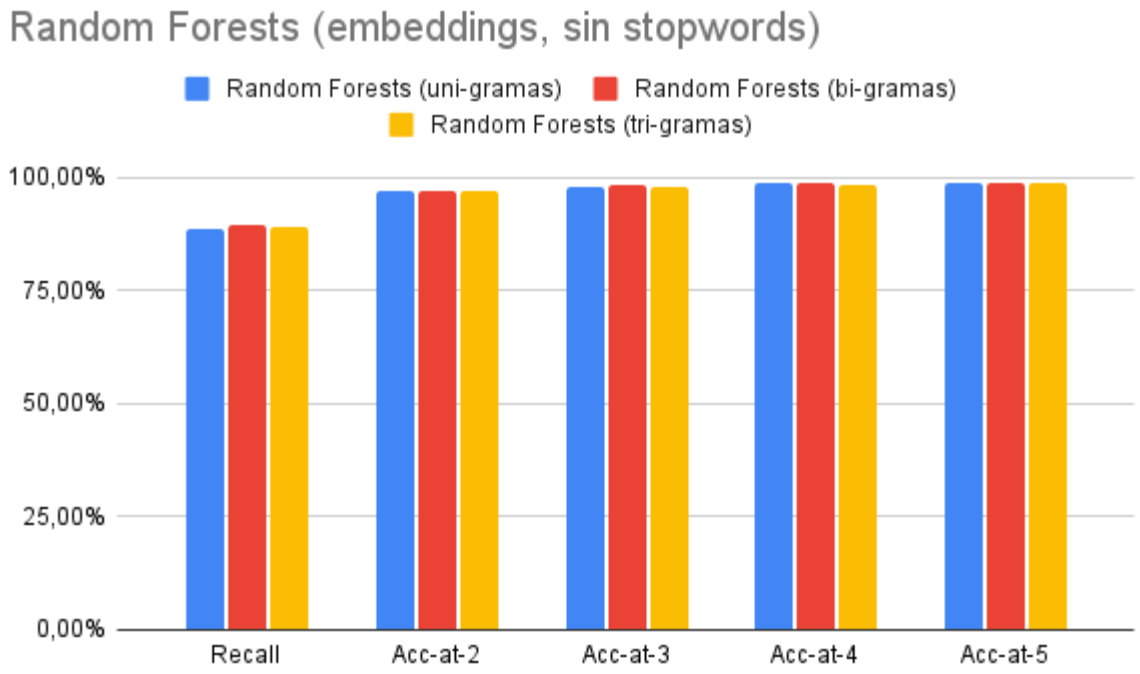
Resultados experimentales





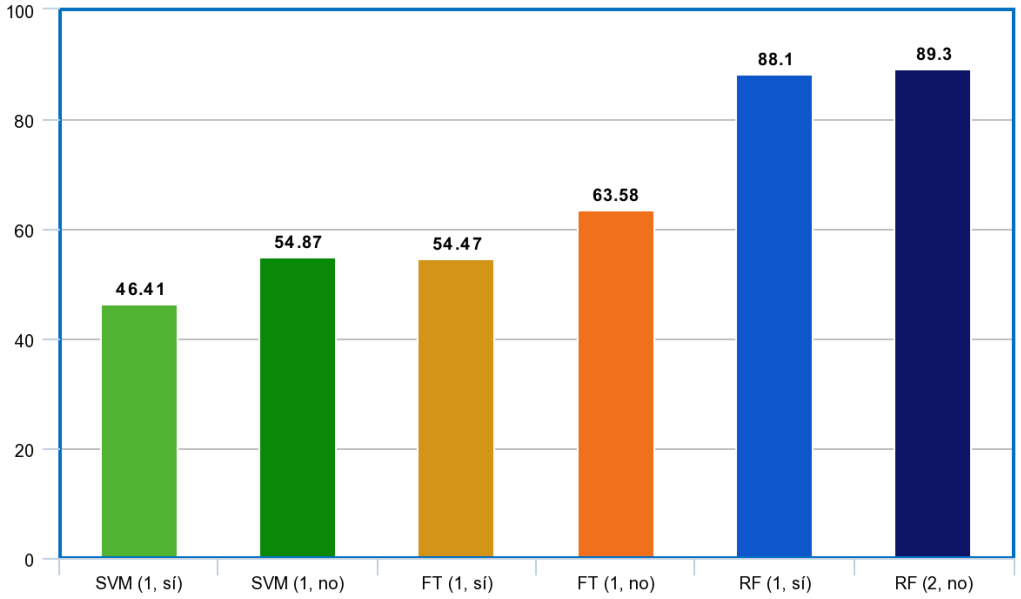
Resultados experimentales



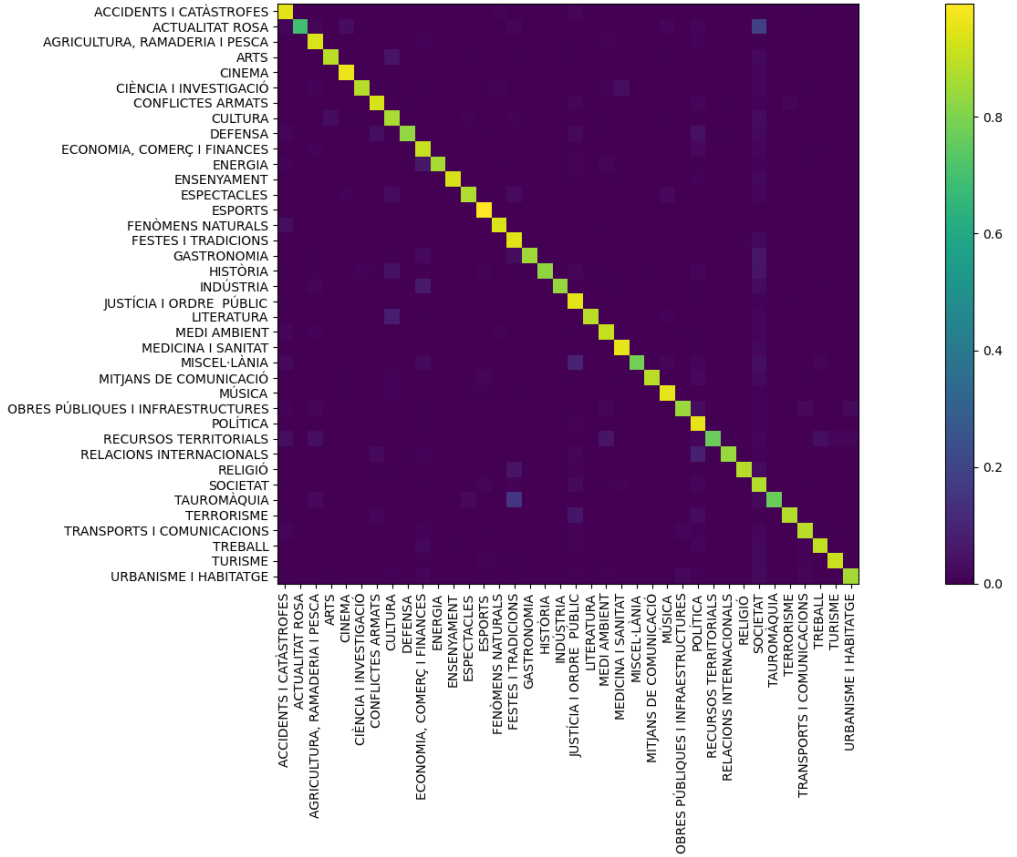


Resultados experimentales

Mejor valor de *recall* macro por modelo



Matriz de confusión para RF(2, no)



Resultados experimentales

Índice

Conclusiones:

- Se ha entrenado con éxito diversos modelos.
- Se ha entrenado con éxito con TF-IDF y *embeddings*.
- Se ha determinado que sí es posible la realización de la herramienta.

Trabajo futuro:

- Realización de *hyperparameter tuning*.
- Generar otros modelos de *embeddings* o uso de *transformers*.
- Implementación de la herramienta.

Clasificación de textos basada en redes neuronales

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Curso 2020-2021

Autor: Mario Campos Mocholí
Tutores: Encarnación Segarra Soriano
Lluís Felip Hurtado Oliver
Emilio Sanchis Arnal