CLASIFICACIÓN DE TEXTO CON MÚLTIPLES ETIQUETAS

PRESENTADO POR

ANGIE JOYA - 2322609

SHEILA VALENCIA - 2243011

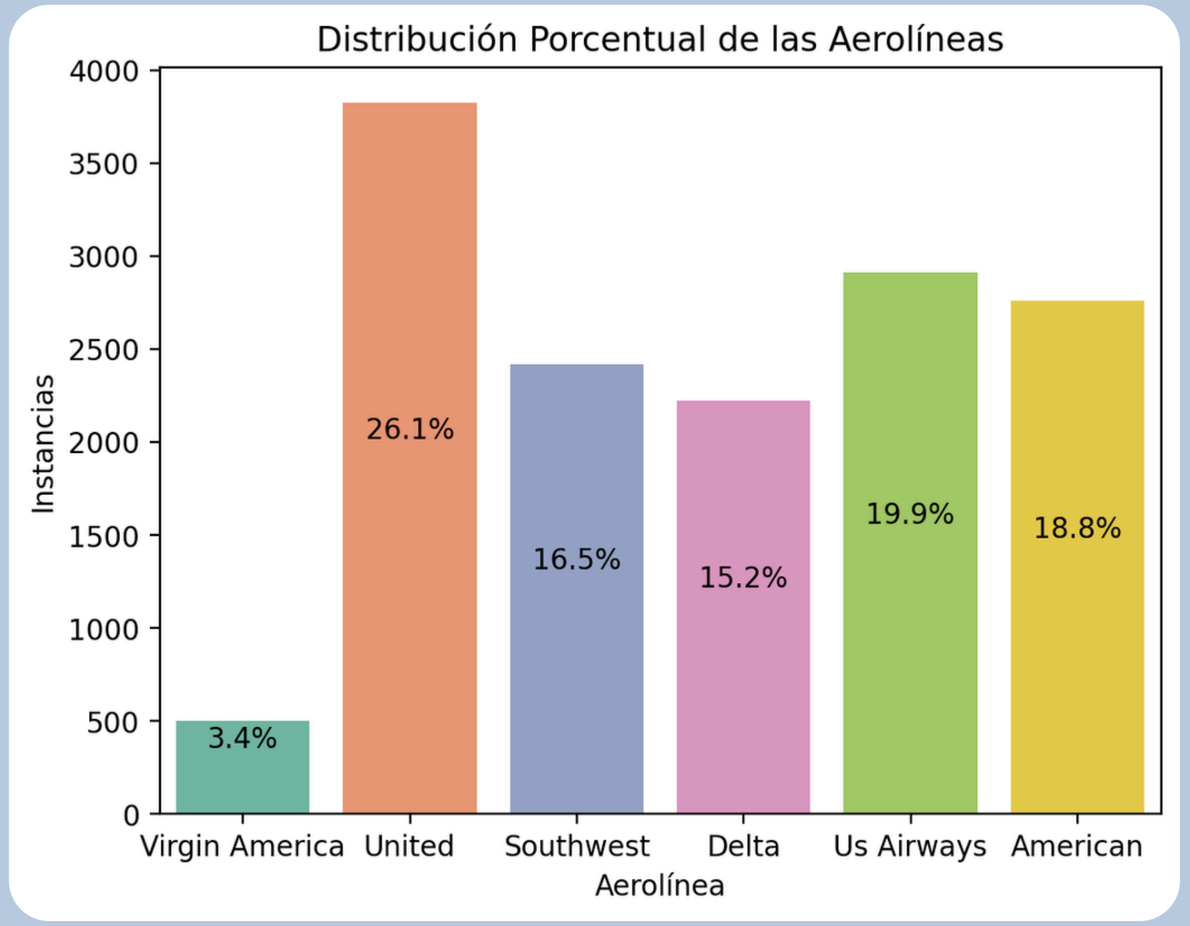
DATASET

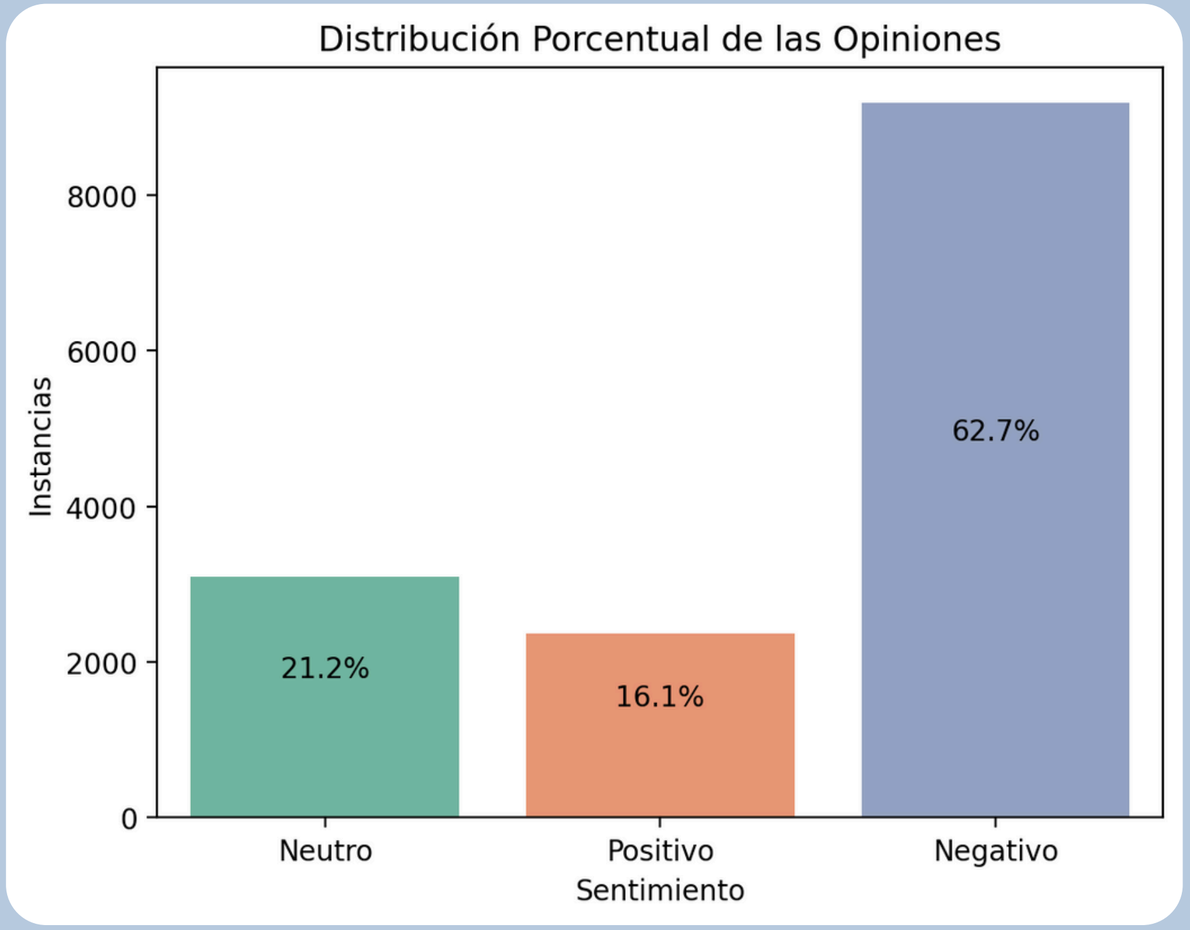
TWITTER US AIRLINE SENTIMENT

DESCRIPCIÓN	Consta de 14.640 publicaciones de Twitter realizadas en el año 2015, las cuales están relacionadas con la opinión sobre seis de las principales aerolíneas de los Estados Unidos. Estas opiniones se clasifican como positivas, neutrales o negativas.
DISTRIBUCIÓN	9.178 - Negativas 3.099 - Neutras 2.363- Positivas

Fuente: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

DISTRIBUCIÓN DE LOS DATOS





EJEMPLOS


Jack
@blackjackpro911
@VirginAmerica amazing to me that we can't get any cold air from the vents. #VX358 #noair #worstflightever #roasted
5:05 AM · Feb 24, 2015
5 Retweets 2 Quote Tweets 17 Likes


Ellen
@urno12
@united thanx so much. You followed through and emailed me a $1000 ticket voucher. #unitedairlines they do care
11:28 PM · Feb 23, 2015
1 Retweet 8 Likes


Krystal~
@kristagermanis
@SouthwestAir What is the best credit card to use/ open to get miles with y'all?
6:46 AM · Feb 22, 2015
5 Retweets 1 Quote Tweet 12 Likes

PREPROCESAMIENTO

OPINIONES

Transformar las opiniones en vectores numéricos (one hot encoding)

negativo = 0 = [1,0,0]
neutral = 1 = [0,1,0]
positivo = 2 = [0,0,1]

TWEETS

Eliminar palabras vacías (a, the...)

Eliminar tags (@), links y # de hashtags

No eliminar negaciones (no, not)

Convertir a minúsculas

Convertir verbos conjugados a su raíz

Tokenizar (vector de palabras)

REPRESENTACIÓN NUMÉRICA

TF - IDF

Da un valor numérico proporcional al número de veces que una palabra aparece en los tweets

PARÁMETROS

Procesar tokens en vez de String

Elimina las palabras que ocurran menos de 5 veces

Vocabulario máximo de 2.000 palabras

@VirginAmerica Your **chat support** is **not working** on your **site** http://t.co/vhp2GtDWPk

['**chat**' , '**support**' , '**not**' , '**work**' , '**site**']

[0.0 , 0.0 , ... , **0.608** , 0.0 , ... , **0.189** , 0.0 , ... , **0.456** , 0.0 , ... , **0.514** , 0.0 , ... , **0.347** , 0.0 , ...]

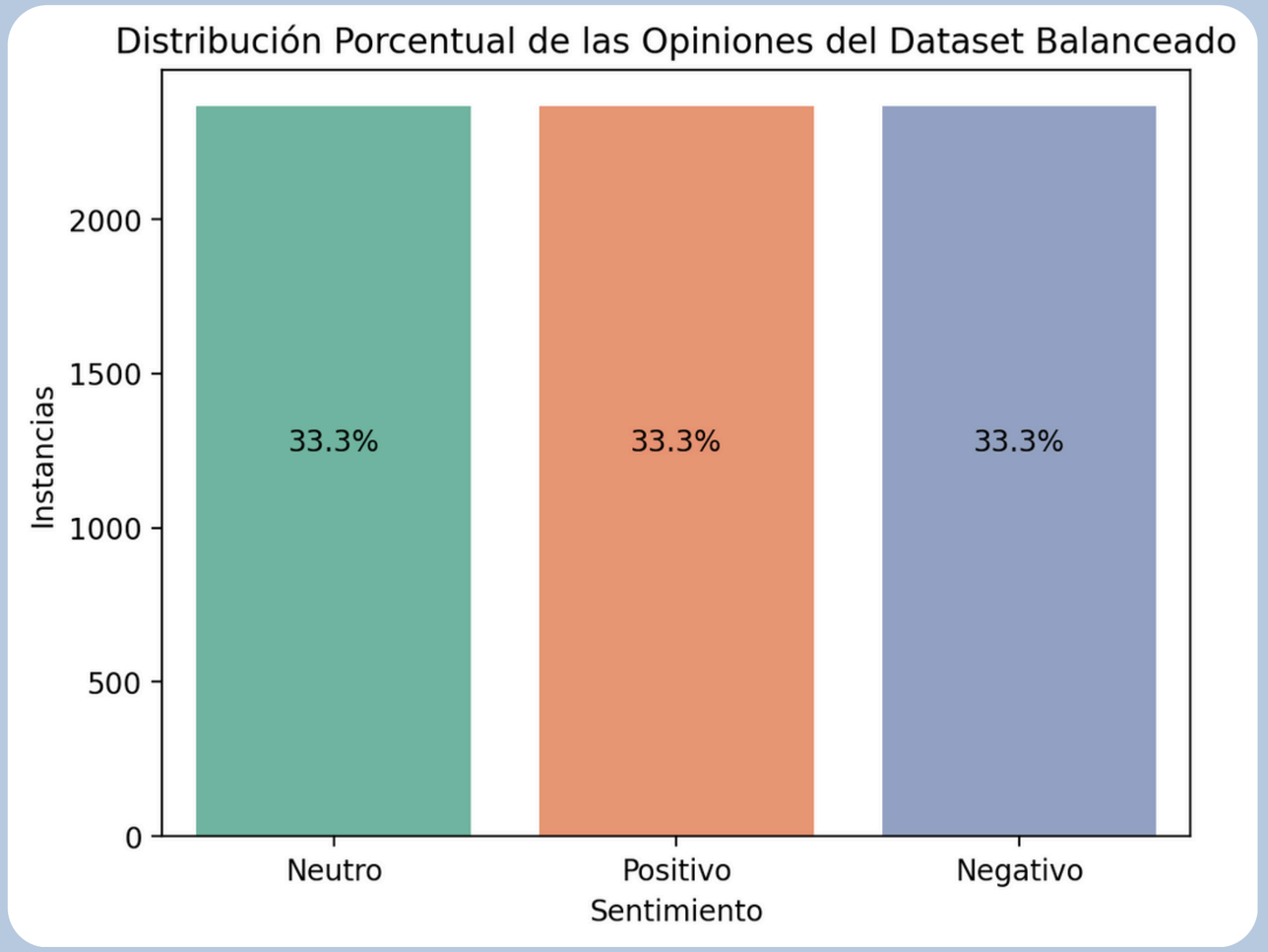
MUESTREO

SUBMUESTREO CLUSTER CENTROIDS

Crea un subconjunto de las clases mayoritarias, donde cada muestra es reemplazada por el centroide de un clúster obtenido mediante clustering. (generalmente k-means)

Todas las clases tienen la misma cantidad (2.363)

Total de 7.089



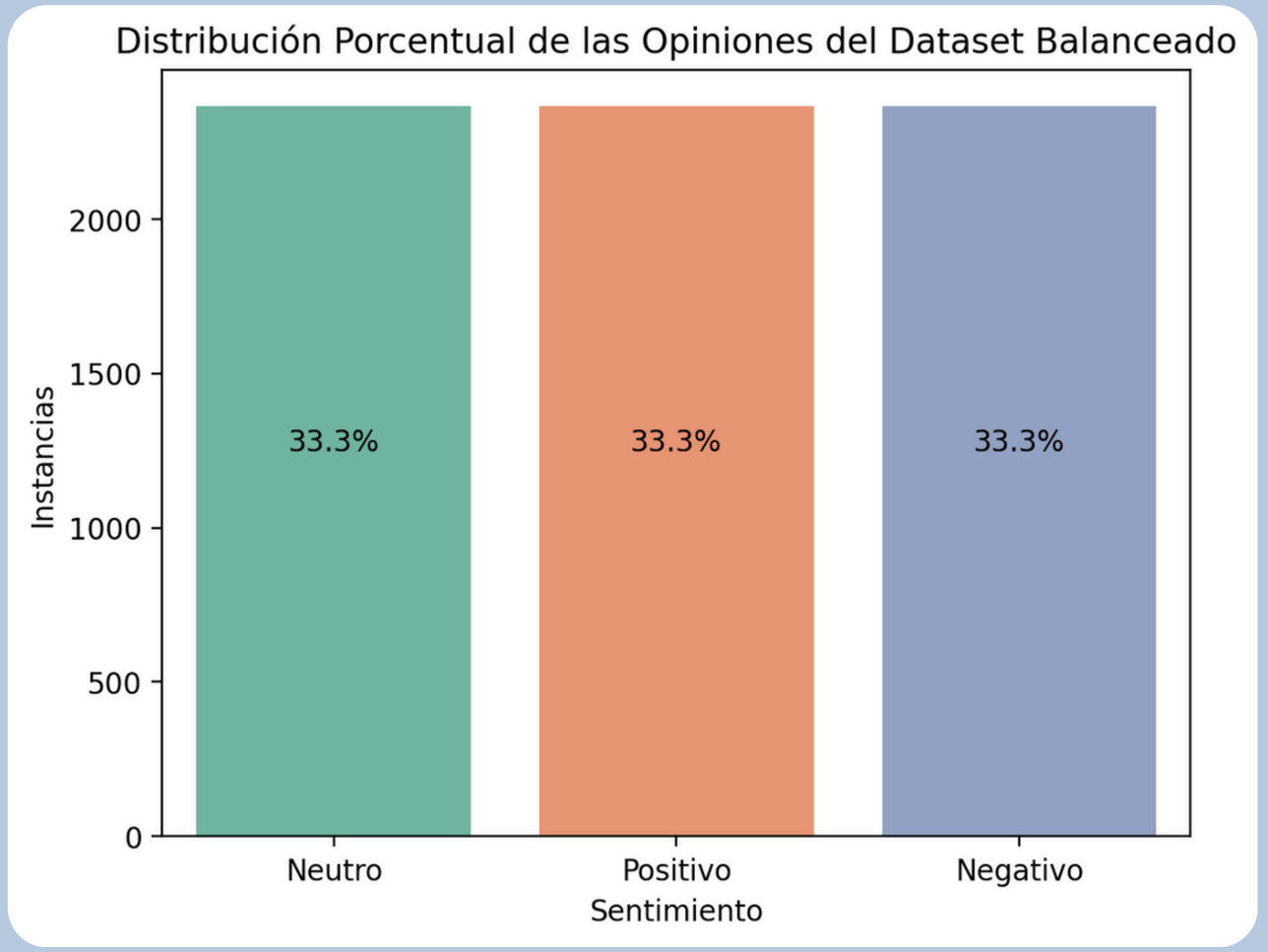
MUESTREO

SUBMUESTREO NEARMISS

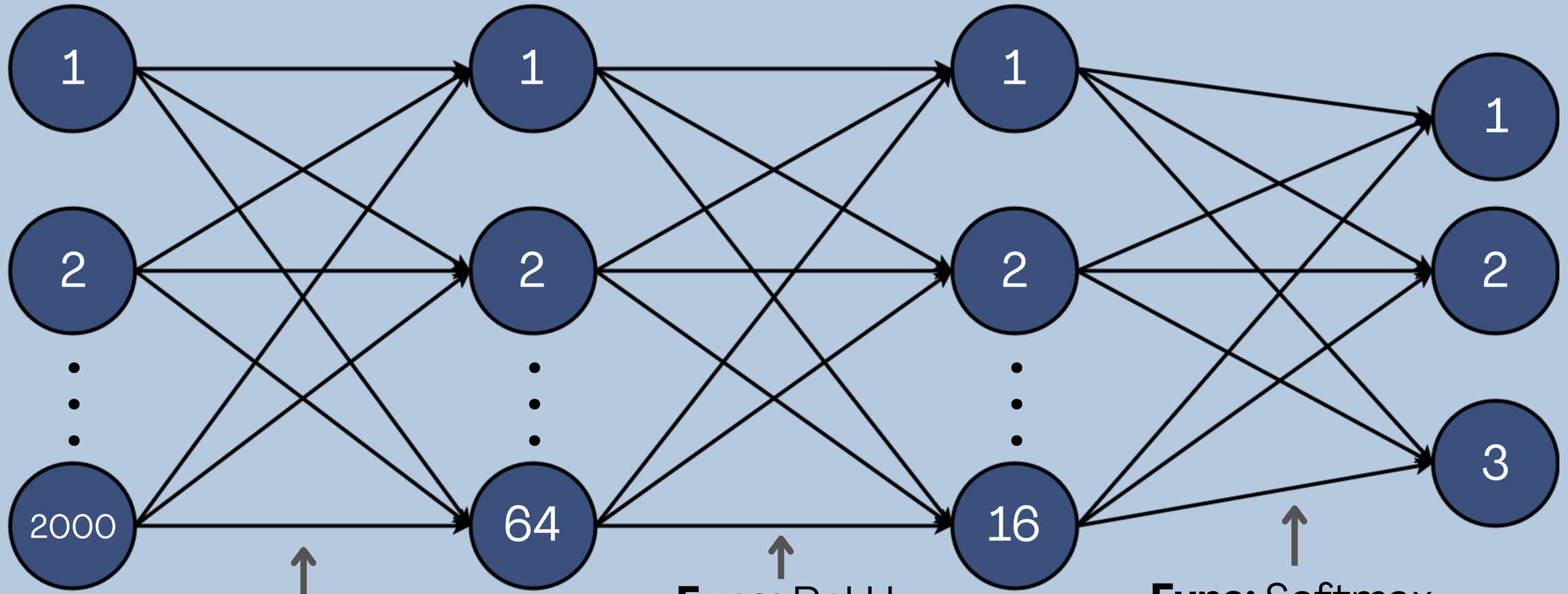
Selecciona ejemplos de la clase mayoritaria que estén cerca de los ejemplos de la clase minoritaria

Todas las clases tienen la misma cantidad (2.363)

Total de 7.089



ARQUITECTURA DE LA RED NEURONAL



ARQUITECTURA DE LA RED NEURONAL

FUNCIONES DE ACTIVACIÓN

Capas Ocultas: ReLU
Capa de Salida: Softmax

FUNCIÓN DE PÉRDIDA

Categorical Crossentropy

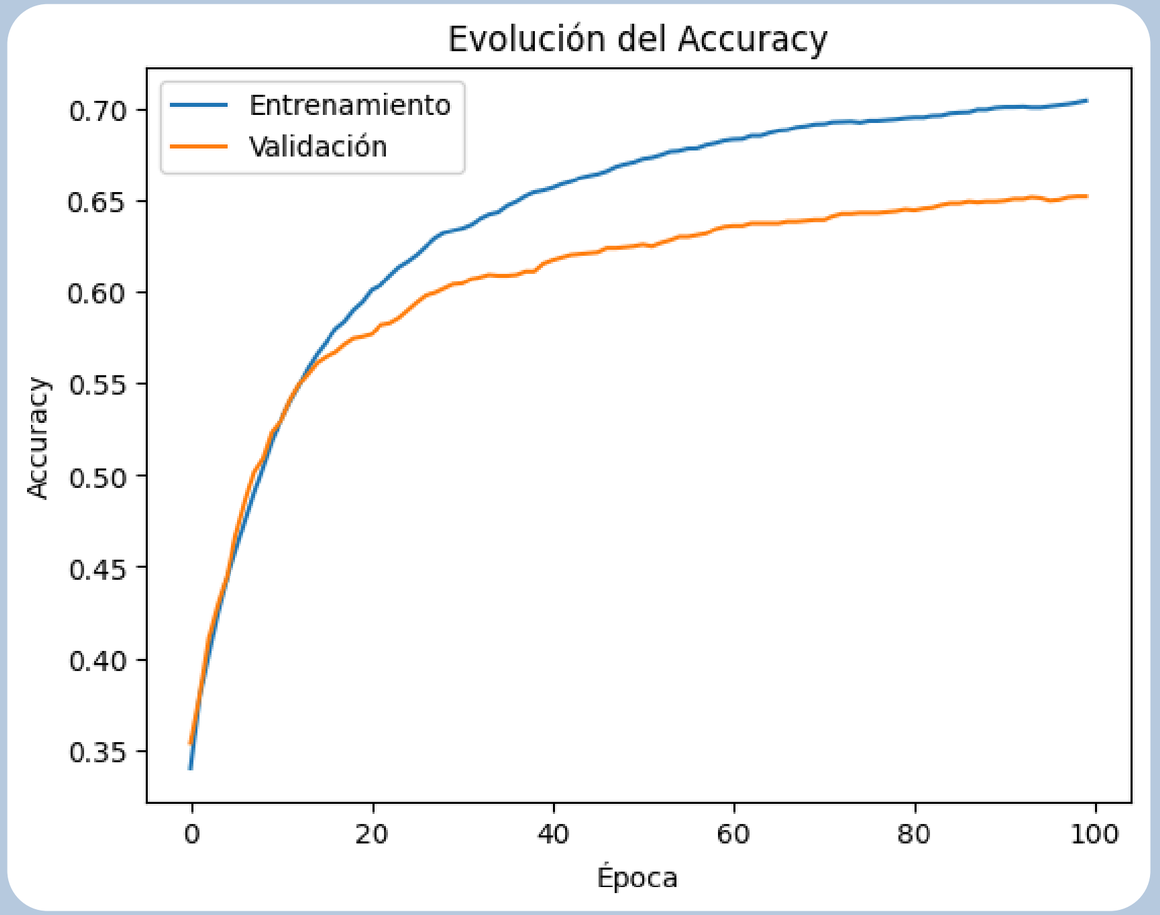
REGULADORES

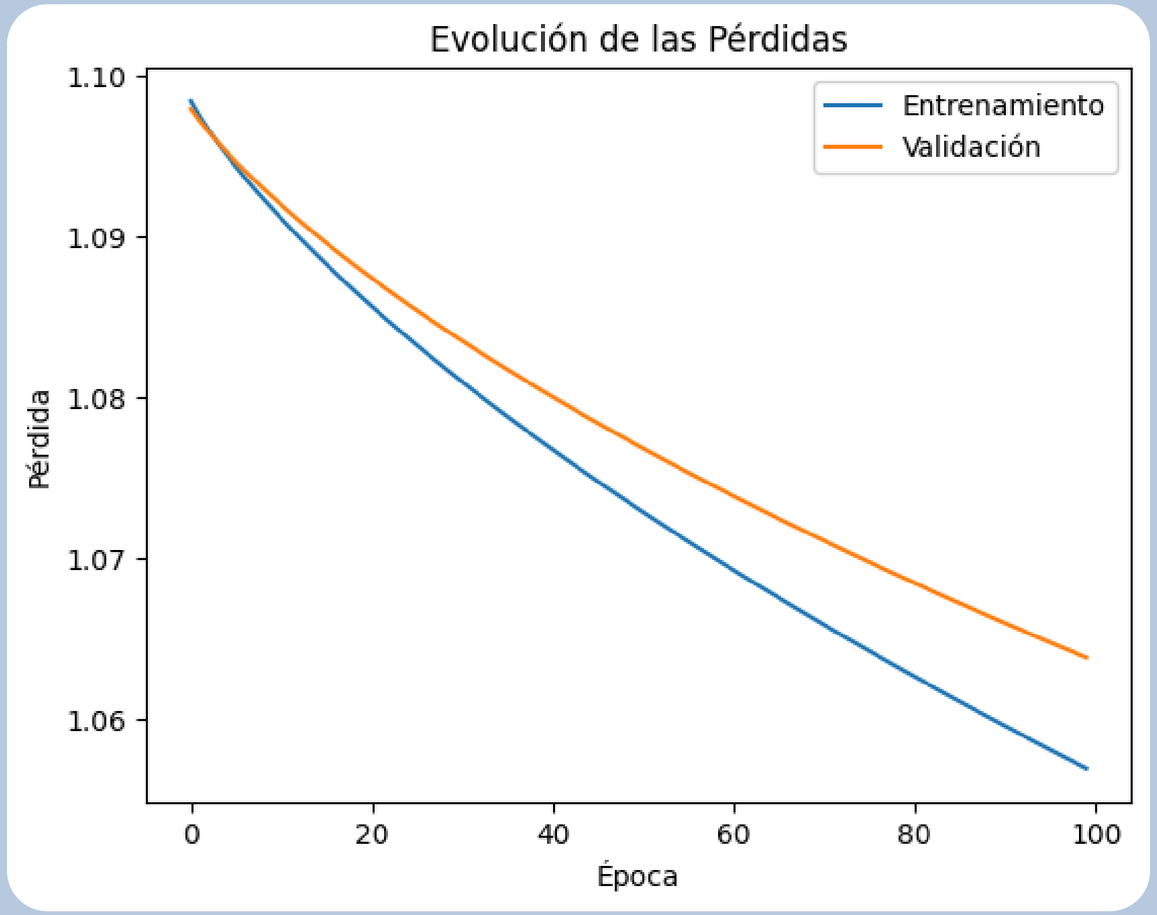
L2: 0.001 (En las capas ocultas)
Dropout: 0.3 (En capa oculta y de salida)
Early Stopping: 4

OPTIMIZADOR

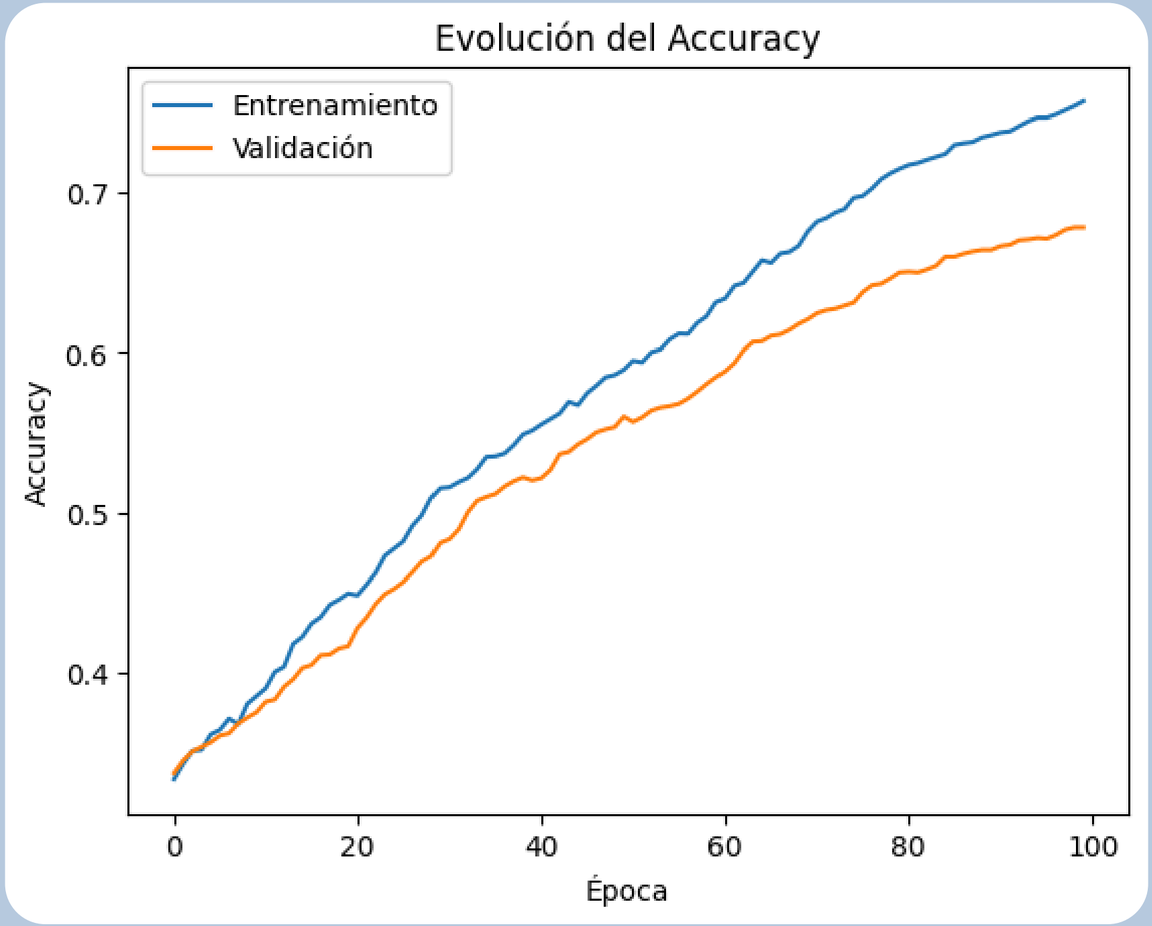
Algoritmo: Adam
Tasa de aprendizaje: 0.00001

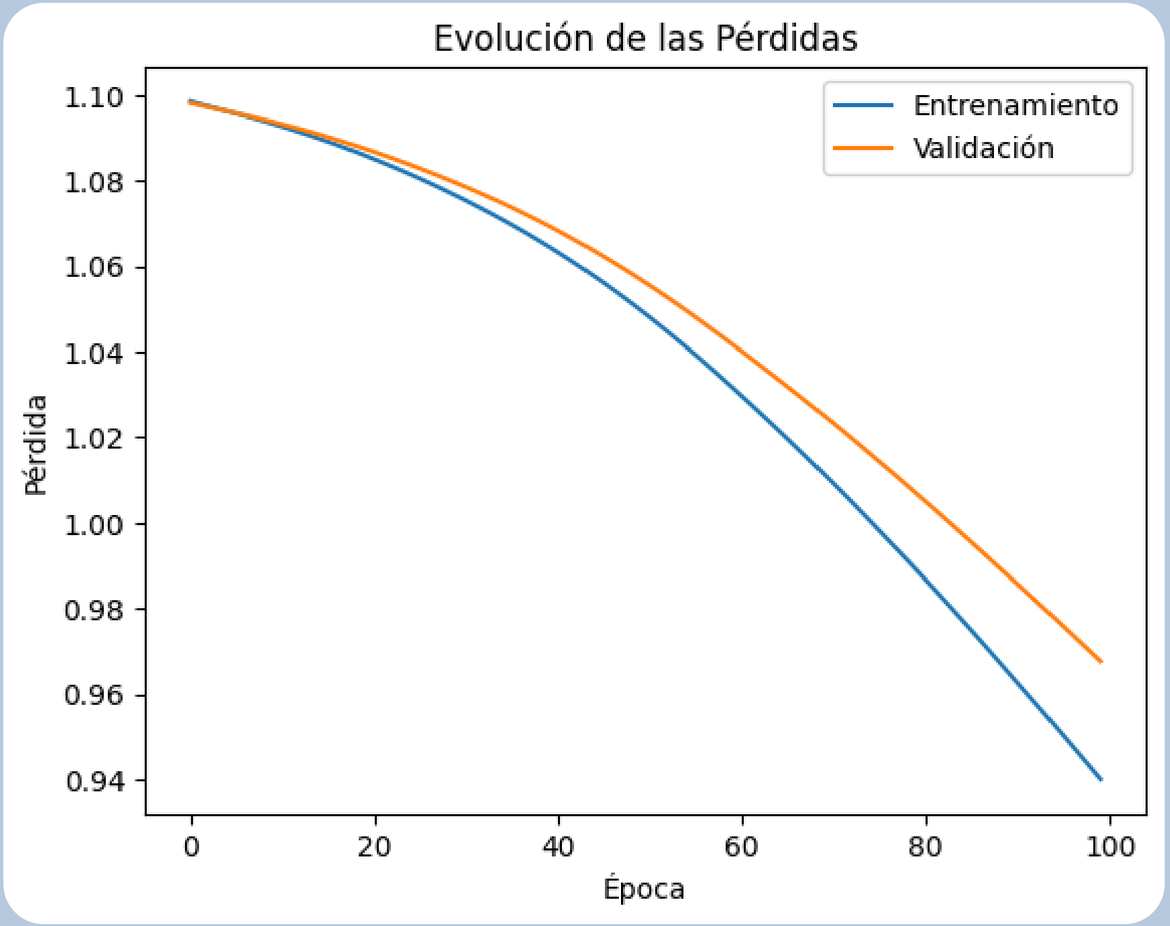
ENTRENAMIENTO - 1 CAPA



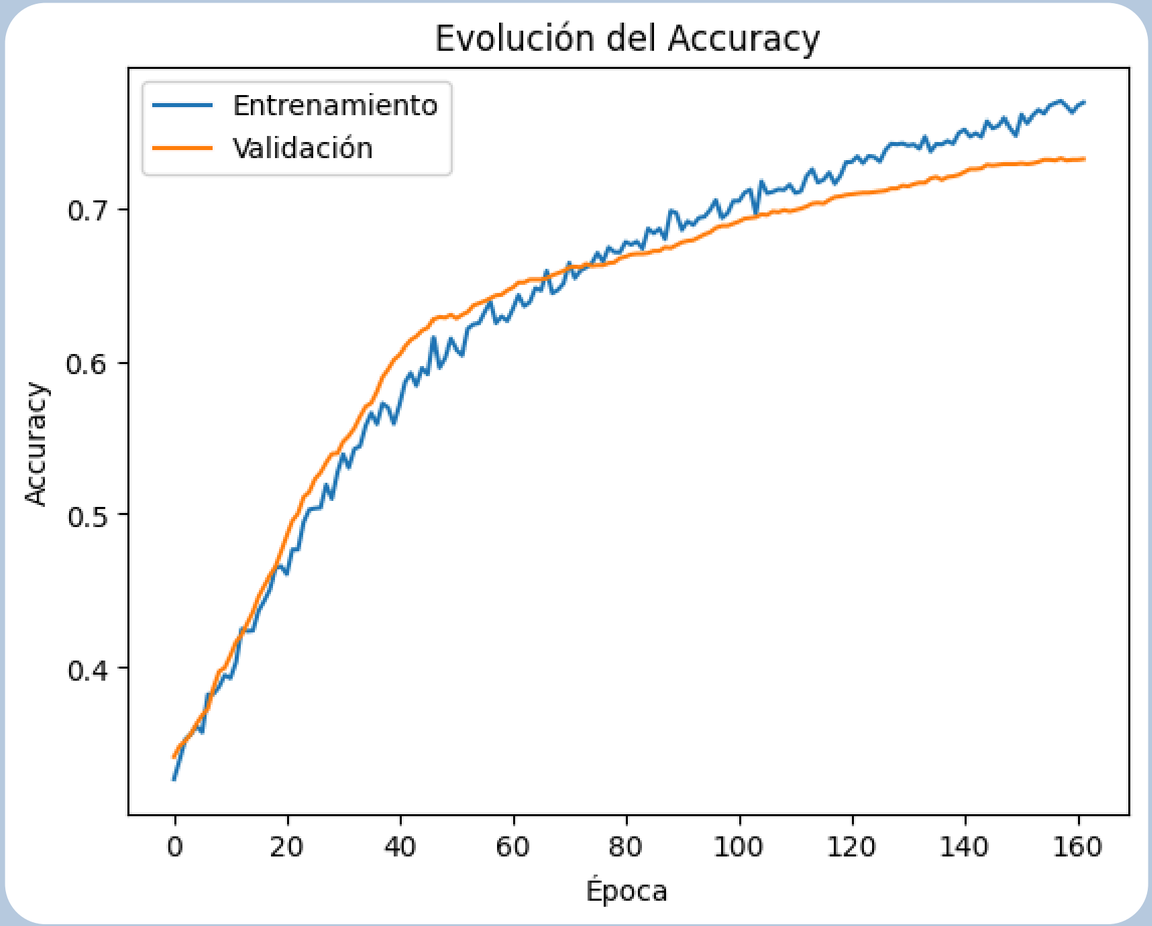


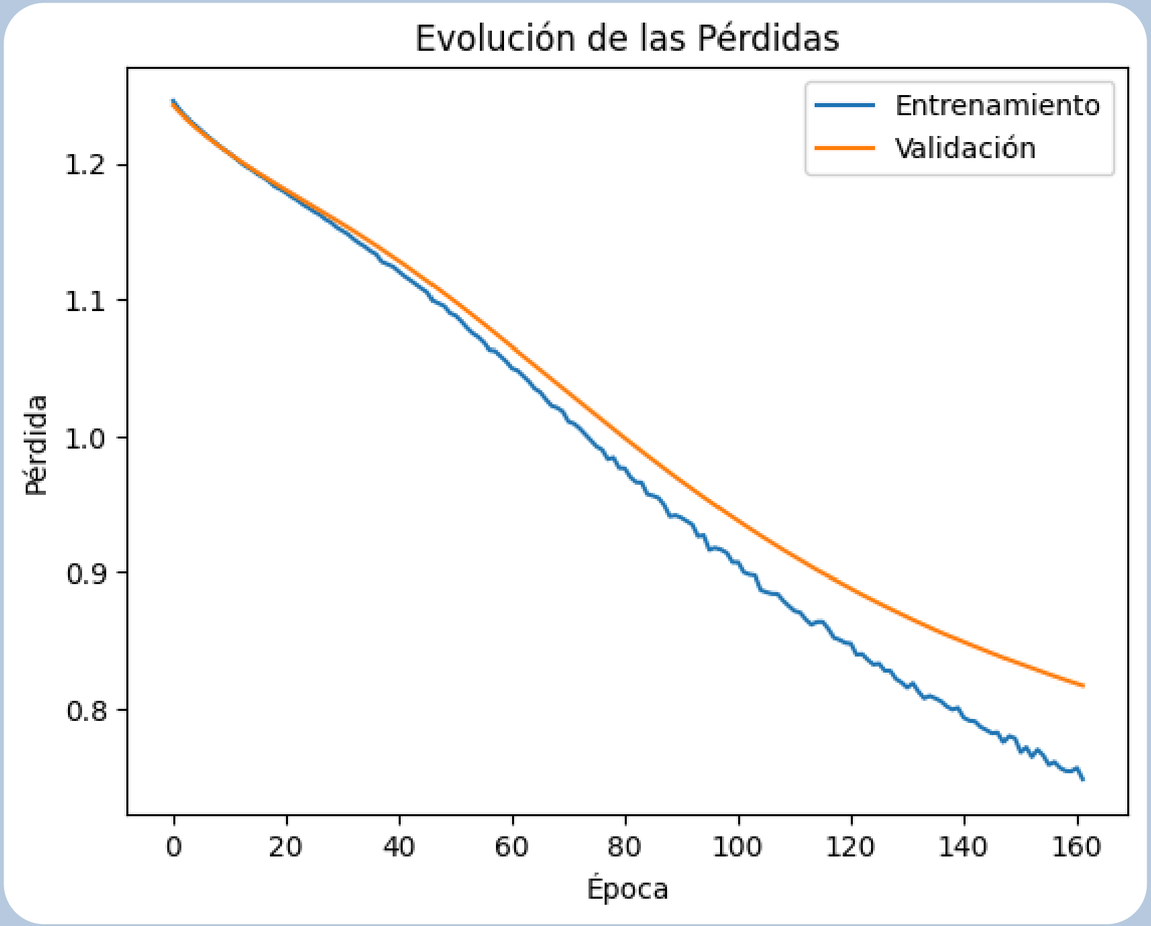
ENTRENAMIENTO - SIN REGULARIZACIÓN



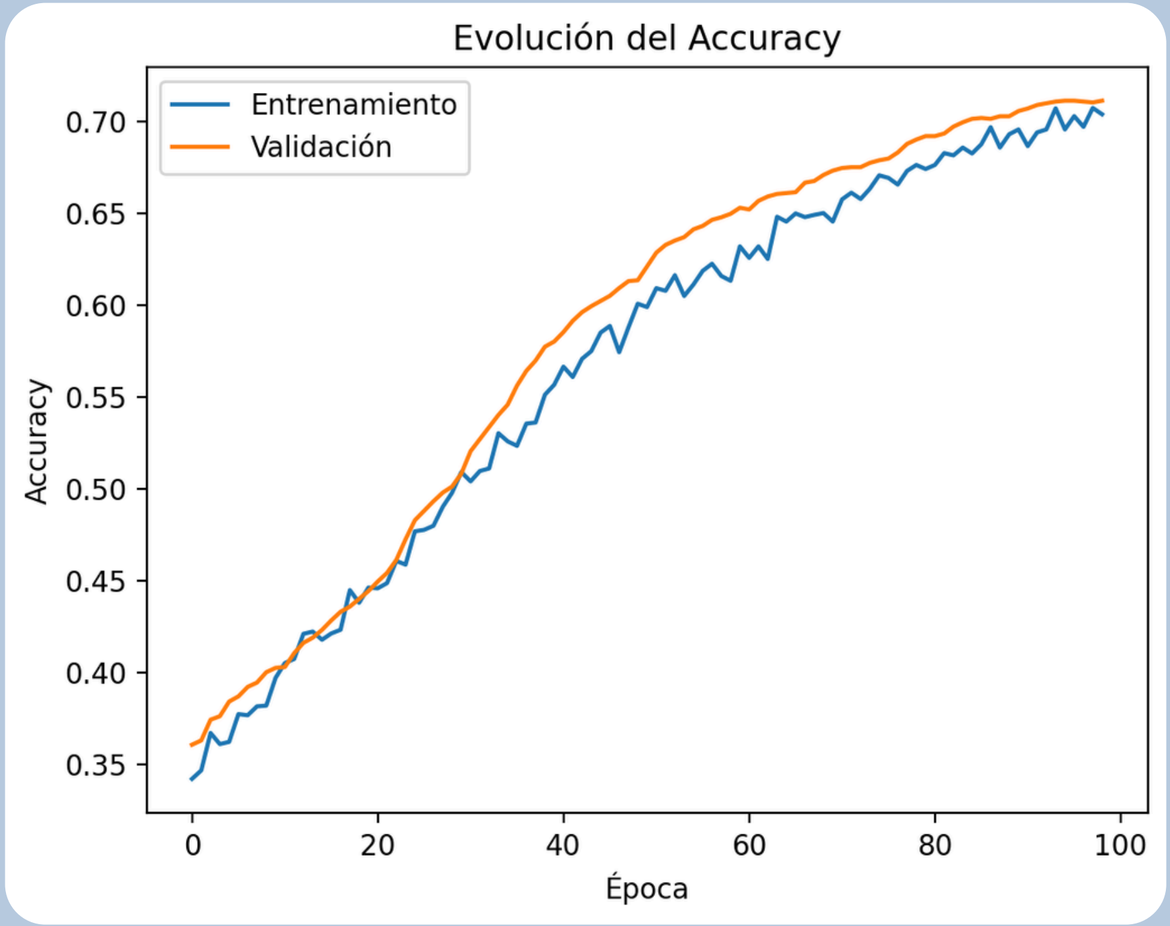


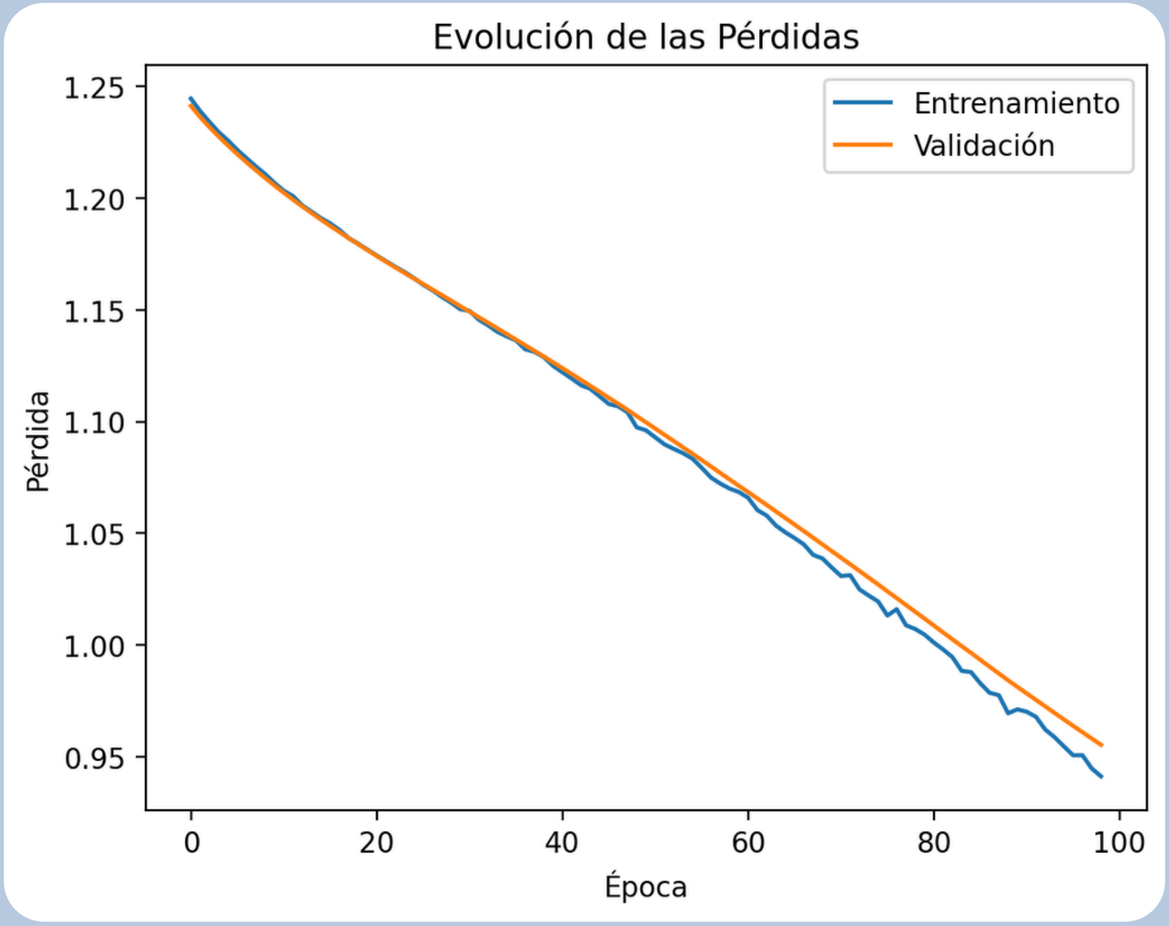
ENTRENAMIENTO - CLUSTER CENTROIDS



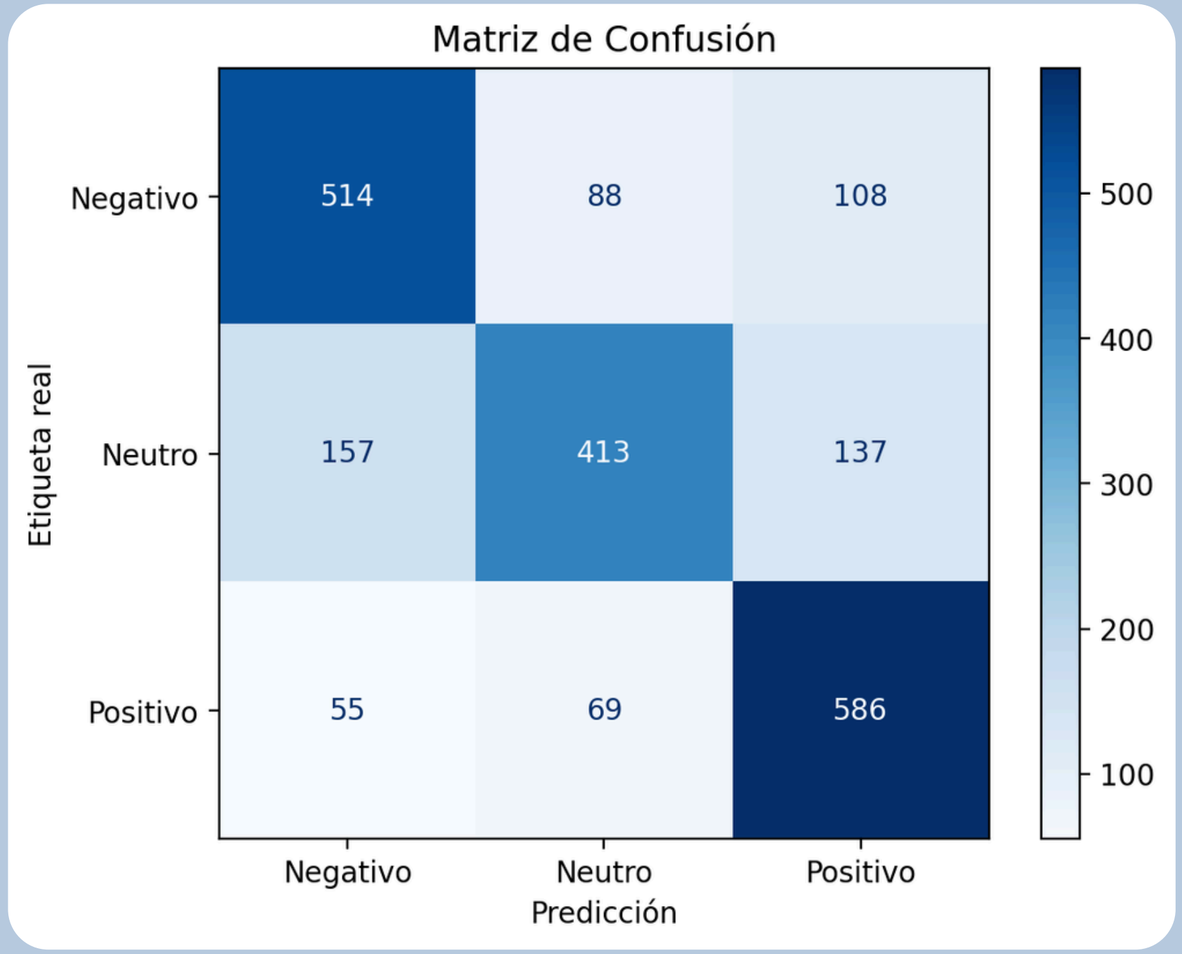


ENTRENAMIENTO - NEAR MISS





ENTRENAMIENTO - NEAR MISS



CONCLUSIONES

PARÁMETROS

Los parámetros seleccionados son robustos en modelos de clasificación de texto, ya que ofrecen estabilidad y un rendimiento eficiente.

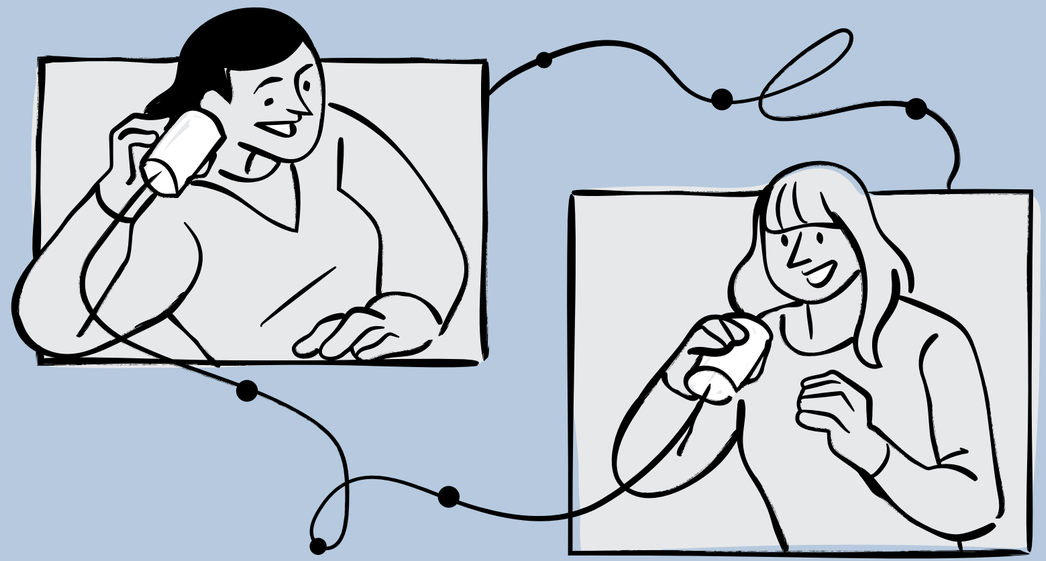
DATASET

Los Tweets no utilizan un lenguaje formal. Contienen contracciones y abreviaturas que complican el preprocesamiento y la clasificación automática.

PRUEBAS

Realizar pruebas con diferentes funciones y algoritmos, tanto de regularización como de optimización, así como variar los parámetros asociados, permite identificar una combinación adecuada

ACCEDER AL NOTEBOOK



¡MUCHAS GRACIAS!