Conclusions
Neural Network
TV Modeling
CV Modeling
BoW
EDA
About

PROCESAMIENTO DEL LENGUAJE NATURAL

OBJETIVO:

Este proyecto tiene como objetivo realizar **Sentiment Analysis** sobre un conjunto de datos provistos por los usuarios de la plataforma Yelp, usando varios algoritmos de Machine Learning.

DATOS DE ORIGEN

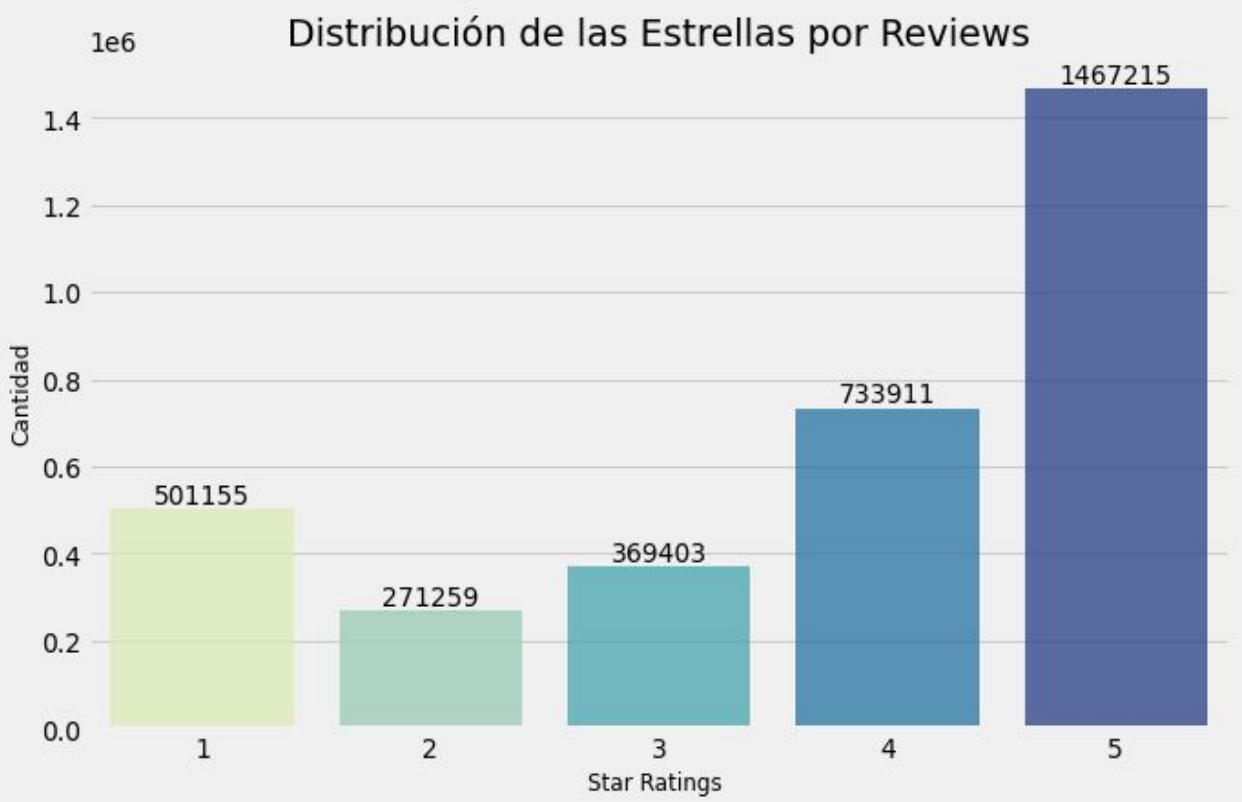
Para clasificar los comentarios en
POSITIVOS ó NEGATIVOS,
se usarán los siguientes atributos:

stars	cantidad de estrellas otorgadas por el usuario en referencia a la review
text	revisión realizada por el usuario sobre un determinado negocio
cool	cantidad de votos por haber sido una review “genial”
funny	cantidad de votos por haber sido una review “divertida”
useful	cantidad de votos por haber sido una review “útil”

CANTIDAD DE INFORMACIÓN PROCESADA:

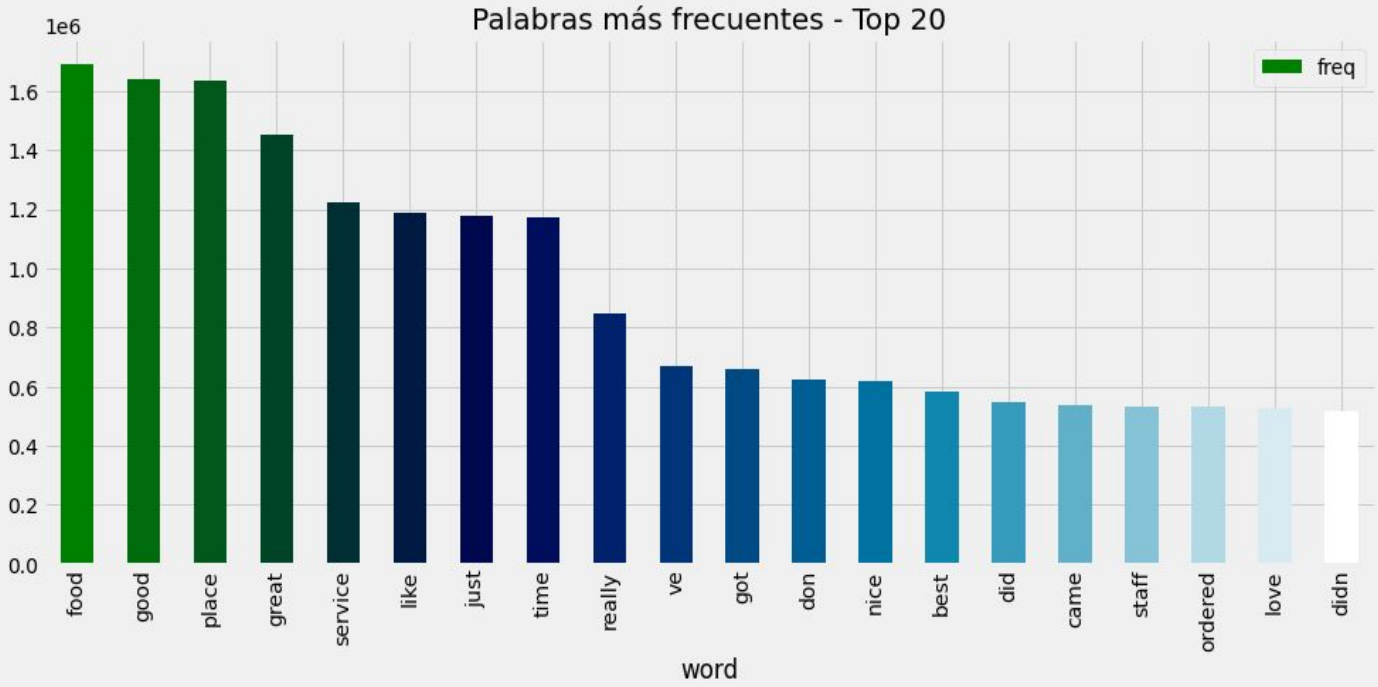
- ✔ ATRIBUTOS: 5 COLUMNAS
- ✔ REGISTROS: 30.000k FILAS

DISTRIBUCIÓN DE LA VARIABLE STARS



Por cada revisión un usuario dio una puntuación de 1 a 5 estrellas. Para pronosticar si una revisión es "positiva" o "negativa", tomaremos la variable de texto como predictor y la variable de estrellas como objetivo (target).

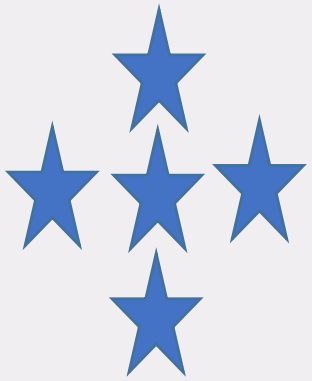
LAS 20 PALABRAS MÁS IMPORTANTES



Podemos encontrar que no importa la estrella que obtenga un lugar, las palabras más frecuentes son comida, servicio y algunas otras palabras (bueno, estupendo, etc.) que se utilizan para describir la calidad de los mismos.

LAS 100 PALABRAS MÁS USADAS PARA REVIEWS CON 3 STARS ó MENOS

LAS 100 PALABRAS MÁS USADAS PARA REVIEWS CON 4 y 5 STARS



CLASIFICADORES MÁS FRECUENTES PARA LAS REVIEWS

Stars 1-2

HORRIBLE

Decepción
Conflicto
Renegar
Lento

Stars 3

NADA

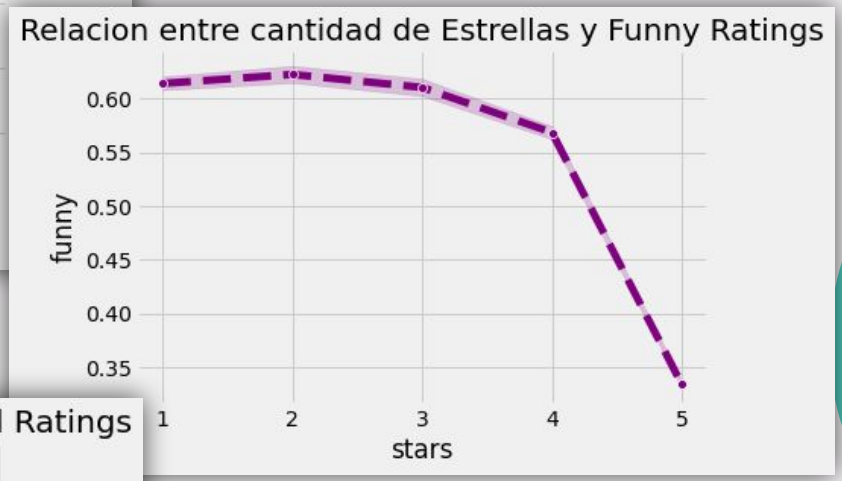
Infame
Problema
Única opción

Stars 4-5

GOOD

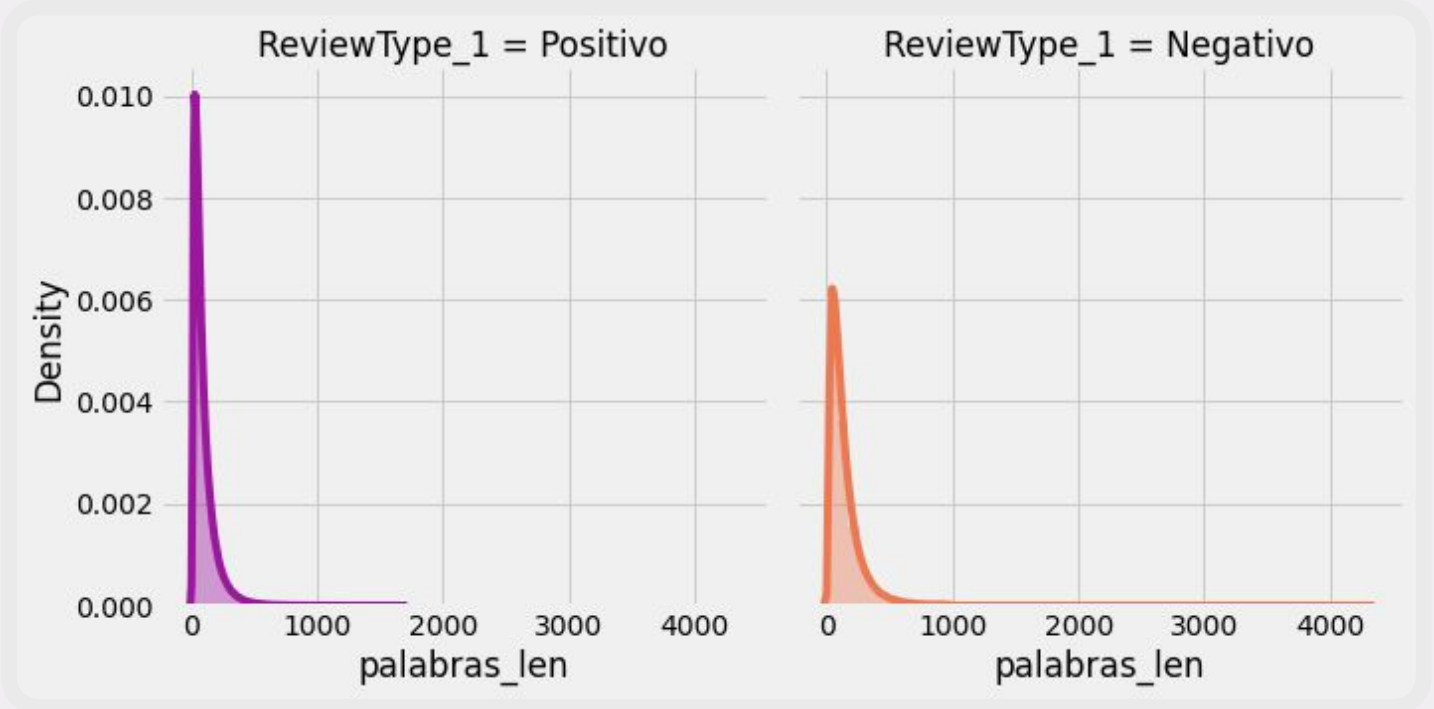
Estupendo
Agradable
Realmente
Siempre

RELACIÓN ENTRE LA CANTIDAD DE STARS Y TIPOS DE VOTACIÓN



- ✔ Las reseñas que tienen una calificación baja (1-2 estrellas) se consideran más "útiles".
- ✔ Las calificaciones con 3 y 4 estrellas han sido votado como 'genial'
- ✔ Las calificaciones más bajas parecen haber sido votadas como "divertidas" en comparación con las reseñas con una calificación de estrellas más alta.

RELACIÓN ENTRE LA CANTIDAD DE STARS Y DURACIÓN DE LA REVIEW



Las personas que tienden a calificar un lugar con 3 estrellas ó menos, tienen en promedio **140** palabras en sus reseñas, mientras que las personas que evalúan un lugar con un comentario de 4-5 estrellas, tienen un promedio de **98** palabras en sus reseñas.

1

ELIMINACIÓN REVIEWS NEUTRAS

Para trabajar con un análisis más polarizado, se eliminan los comentarios con 3 estrellas

2

CLASIFICACIÓN BINARIA

Se toma una muestra de datos, donde se asigna 1 a las reviews consideradas positivas y 0 a las consideradas negativas

3

BALANCEO DE LOS DATOS

Se balancean las clases, para que el desequilibrio entre estrellas, no influya en los resultados del modelo

Eliminación de Ruido

- ✔ Poner texto en minúsculas
- ✔ Tokenizar
- ✔ Quitar números
- ✔ Quitar signos de puntuación
- ✔ Quitar token vacíos
- ✔ Quitar tokens con una letra

Proceso Reviews en otro idioma

- ✔ Se identifican los comentarios en otro idioma
- ✔ Se eliminan los comentarios a los cuales no se les pudo identificar el idioma
- ✔ Se traducen al inglés los comentarios para los cuales se encontró idioma

Separar Set Train & Test

- ✔ Se toman una muestra para entrenar los modelos y otra para prueba

IMPLEMENTACIÓN COUNT VECTORIZER

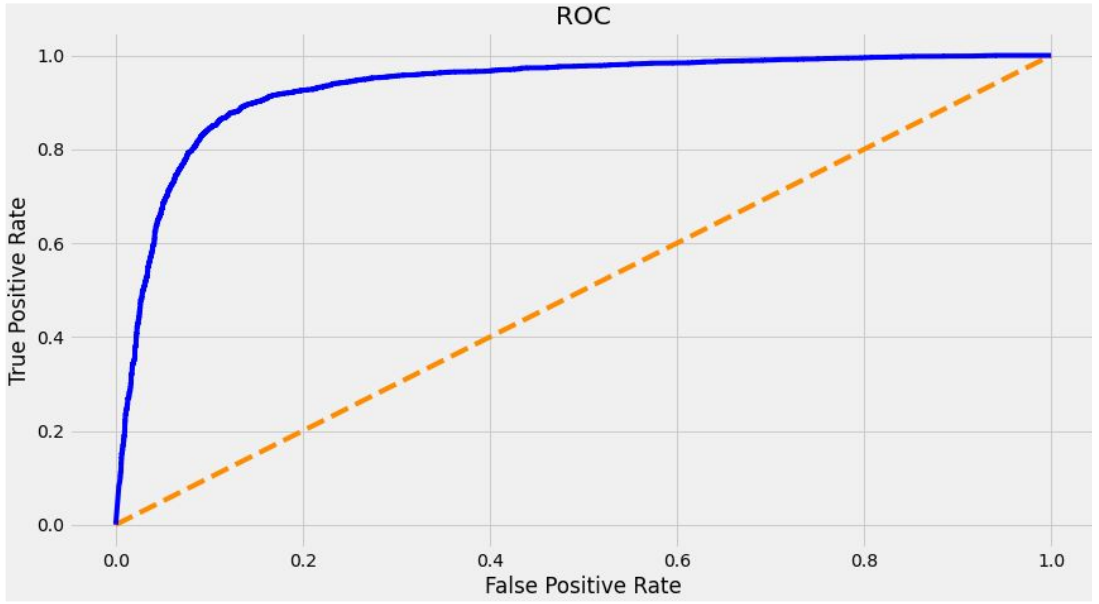
Naive Bayes

KNN (12 vecinos)

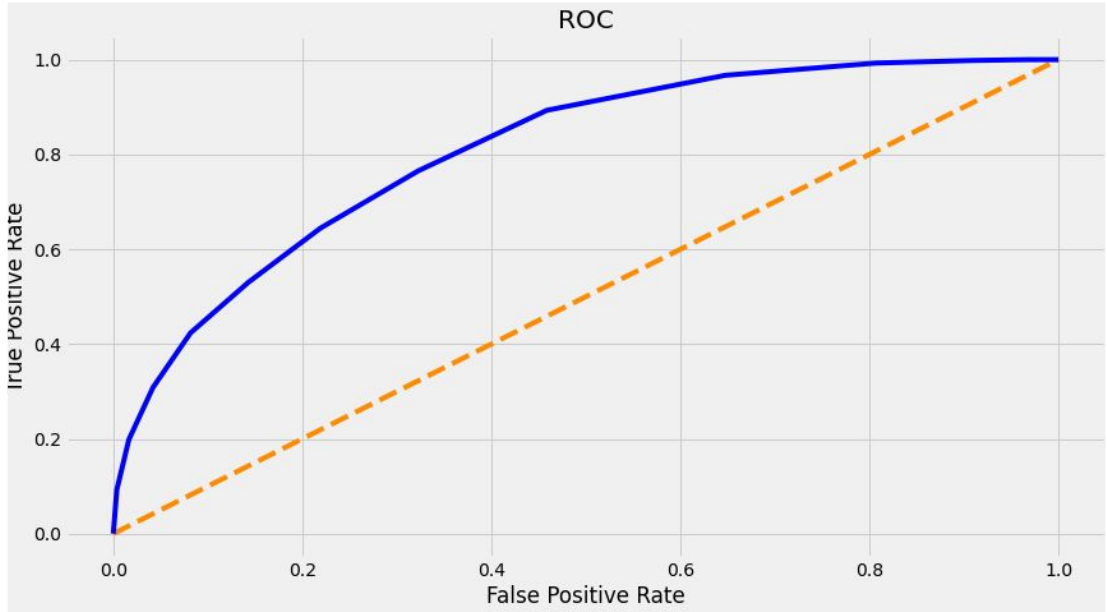
METRICS	**NAIVE BAYES**	KNN
✔ Accuracy	**0.88**	0.71
✔ Precisión	**0.86**	0.75
✔ Recall	**0.90**	0.64
✔ F1 score	**0.88**	0.69

Naive Bayes

CURVA ROC



KNN (12 vecinos)



IMPLEMENTACIÓN TFIDF VECTORIZER

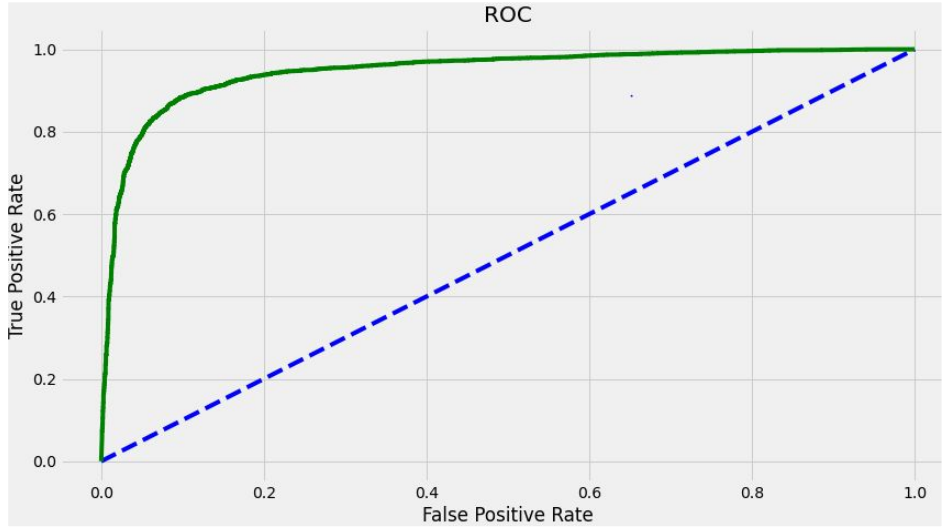
Naive Bayes

KNN (12 vecinos)

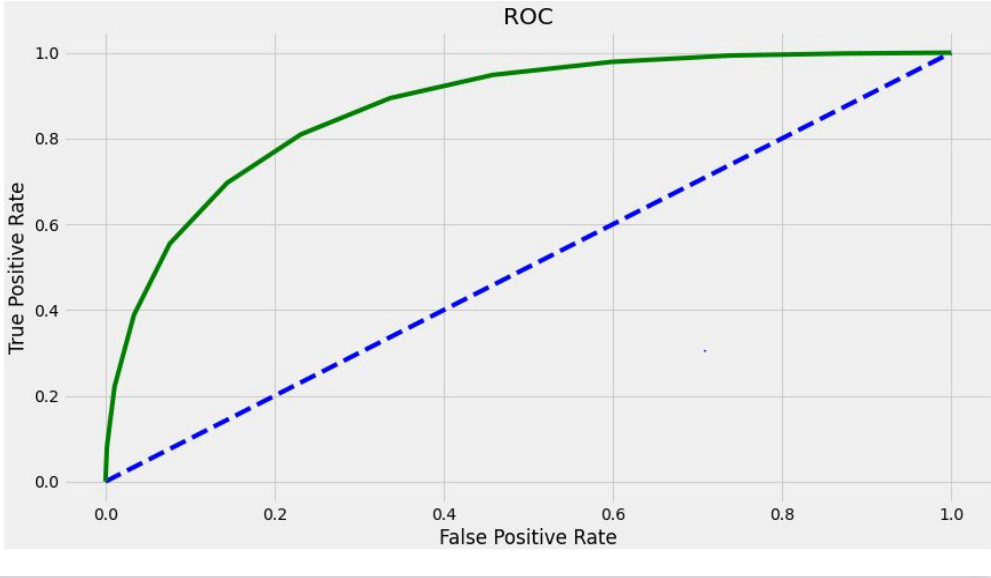
METRICS	**NAIVE BAYES**	KNN
✔ Accuracy	**0.89**	0.78
✔ Precisión	**0.90**	0.83
✔ Recall	**0.87**	0.70
✔ F1 score	**089**	0.76

Naive Bayes

CURVA ROC



KNN (12 vecinos)



RED NEURONAL

Embedding

Con GloVe de 50D pre entrenado

Conv1D

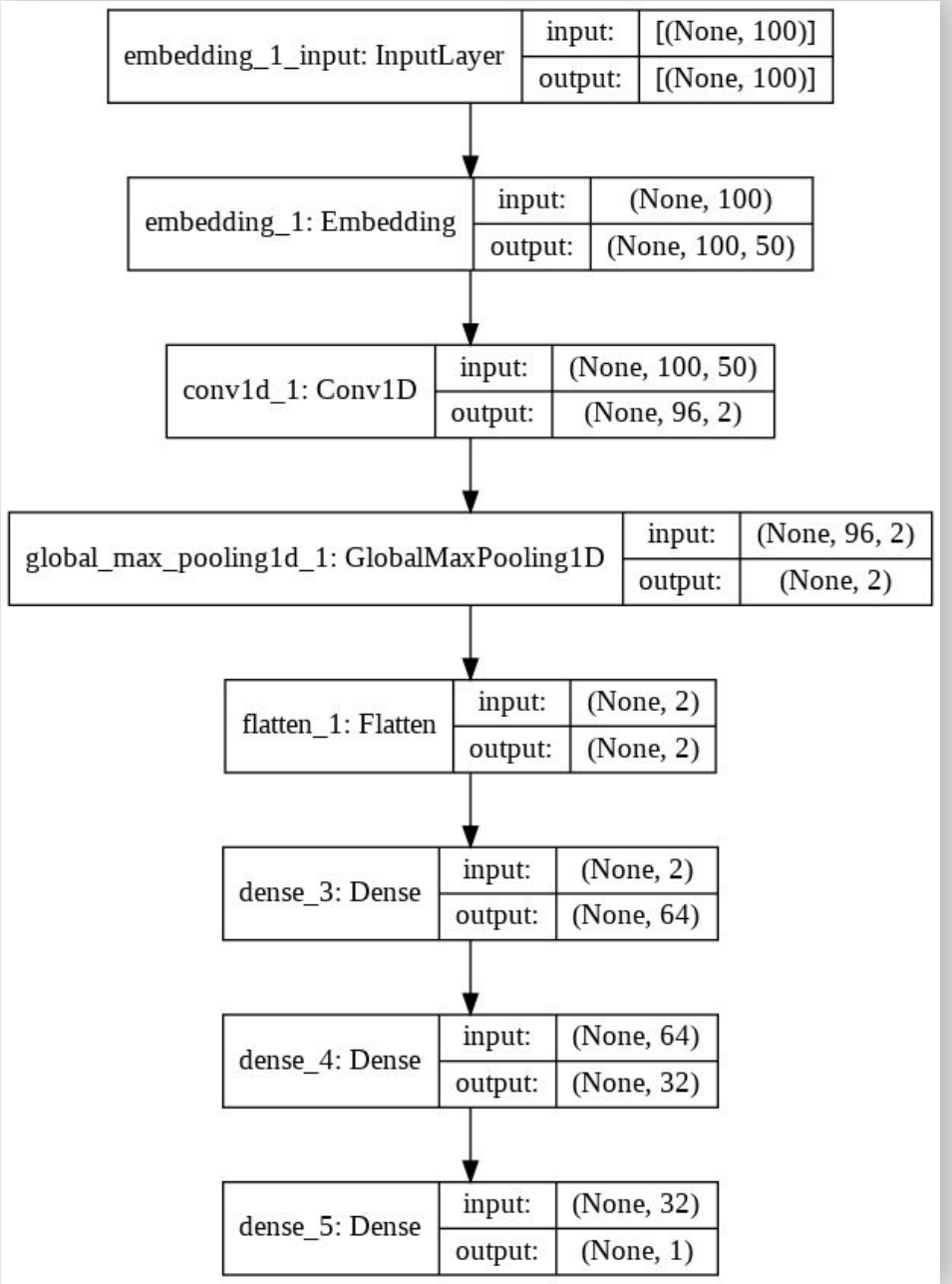
Con 2 filtros y kernel size 5

MaxPooling

Reducción de dimensión

Flatten

Unidimensionalidad para ingreso a capa Densa



RED NEURONAL

Modelo1

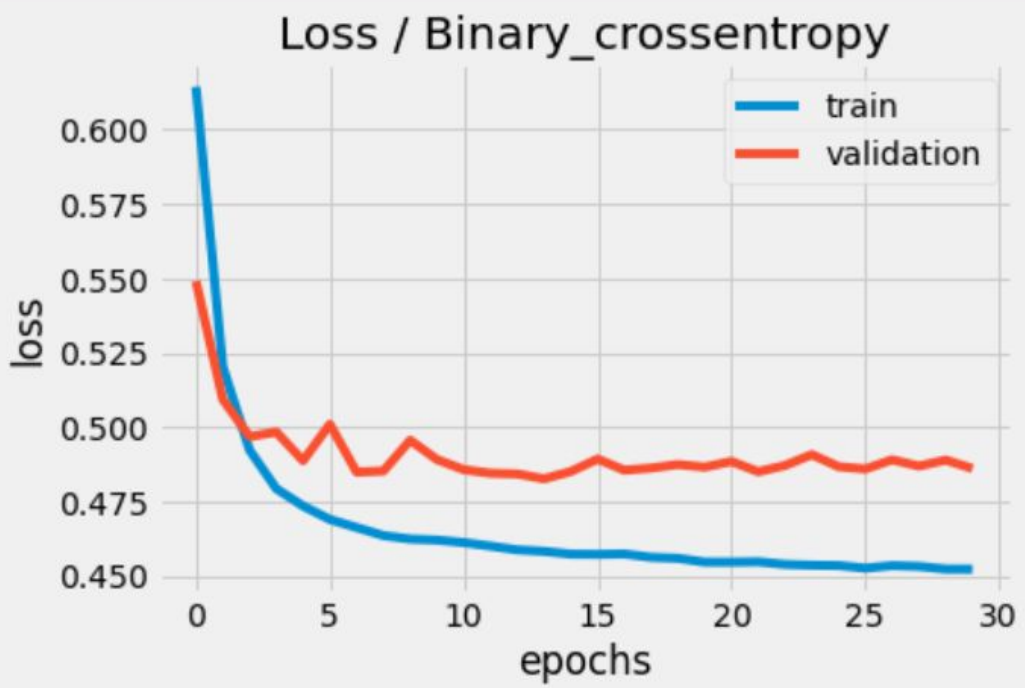
64 - 48 -1 Neuronas
Activación Relu
Conv1D 2 filtros

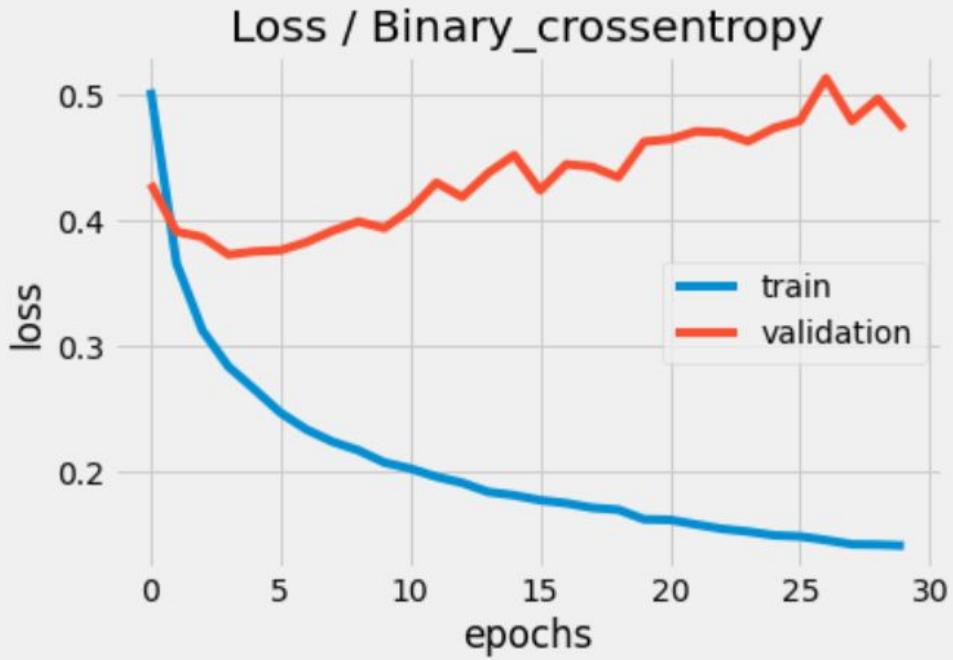
Modelo 2

64 - 48 -1 Neuronas
Activación Relu
Conv1D 20 filtros

METRICS	**MODELO 1**	MODELO 2
✔ Accuracy	**0.76**	0.84
✔ Precisión	**0.76**	0.84
✔ Recall	**0.77**	0.84
✔ F1 score	**0.76**	0.84

Modelo1





Modelo2

Conclusions

Neural Network

TV Modeling

CV Modeling

BoW

EDA

About

Selección del Modelo final

MÉTRICAS	NB (con CV)	NB (con TF)	RED NEURONAL
✔ Accuracy	0.88	0.89	0.76
✔ Precisión	0.86	0.90	0.76
✔ Recall	0.90	0.87	0.77
✔ F1 score	0.88	0.89	0.76

Grupo Nro.4

Integrantes:

Agustina Ghelfi, Cecilia Manoni, Carolina Guzmán, Noelia Ferrero