



INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO (ESCOM)



Tecnologías de procesamiento de Lenguaje Natural

Práctica 5. Análisis de
sentimientos, polaridad u opinión.

Integrantes:

Gutierrez Cuello

Christian Alfredo

Boleta: 2022630718

Grupo: 5BV1

Profesor: Ituriel Enrique Flores Estrada

PRÁCTICA 5

RESUMEN

En la práctica se llevo acabo la clasificación de análisis de sentimientos para las opiniones dadas por lo clientes de un restaurante, se usaron varios modelos con el entrenamiento y una codiación especifica de la practica, además de haber hecho el análisis exploratorio de datos, y la limpieza que se debian hacer en los textos obtenidos del dataset, esto con el finde obtener un corpus para poder tanto entrenar los modelos como hacer las predicciones obre los mismos.

En cuanto a la forma en que se hicieron los análisis de sentimientos, se utilizaron 3 tecnicas, por diccionarios, por modelos de machine learning, y por modelos de Deep learning, usando capas de redes neuronales.

Los resultados demuestran que hay variedades según como se lleve acabo el entrenamiento de los modelos, la forma, los hiperparametros y el tipo de técnicas usada.

INTRODUCCIÓN

En esta práctica, nos enfocamos en el análisis de sentimientos de textos, con el objetivo de predecir etiquetas de sentimiento (positivo, neutral y negativo) mediante el uso de diferentes enfoques y técnicas de procesamiento de lenguaje natural. A lo largo del desarrollo, se aplicaron técnicas de normalización y limpieza de texto para mejorar la precisión y el rendimiento de los modelos utilizados.

Inicialmente, se llevó a cabo un Análisis Exploratorio de Datos (EDA) para comprender mejor el conjunto de datos de reseñas de productos alimenticios de Amazon. Este análisis permitió identificar las características y dimensiones clave del dataset, así como determinar qué columnas eran relevantes y cuáles podían ser eliminadas. Posteriormente, se realizó una conversión de las calificaciones en etiquetas de sentimiento, y se verificó el balance de clases para asegurar la equidad en la cantidad de observaciones para cada sentimiento.

El preprocesamiento de los datos incluyó la limpieza de texto mediante la eliminación de comentarios, metadatos, números telefónicos, direcciones, signos de puntuación, términos alfanuméricos y stopwords. Estas técnicas de limpieza fueron justificadas y aplicadas para reducir el ruido en los datos y mejorar la calidad de los análisis posteriores.

En cuanto a la normalización y transformación de los datos, se aplicaron diversas técnicas, tales como la conversión de términos a minúsculas, el stemming y la lematización, la tokenización y la vectorización de reseñas utilizando técnicas como TF-IDF y One Hot-Encoding, dependiendo del enfoque específico.

Se experimentó con tres enfoques distintos para el análisis de sentimientos: diccionarios, modelos de aprendizaje de máquina y redes neuronales. Para el enfoque basado en diccionarios, se utilizaron los lexicones Harvard IV-4 y Opinion Lexicon. En el caso de los modelos de aprendizaje de máquina, se implementaron regresión logística, árboles de decisión y máquinas de soporte vectorial. Finalmente, para el enfoque de redes neuronales, se emplearon capas de embeddings preconstruidas y aprendidas a partir del corpus de documentos.

Cada uno de estos enfoques fue evaluado y comparado utilizando la técnica de validación cruzada (k-fold cross-validation) para asegurar la robustez y fiabilidad de los resultados obtenidos. Los resultados de cada análisis fueron presentados de manera detallada, mostrando la efectividad de cada técnica y modelo en la predicción de etiquetas de sentimiento

DESARROLLO DE LA PRÁCTICA

Primero, hablemos un poco sobre los algoritmos y enfoques usados en el análisis de sentimientos:

El análisis de sentimientos es un campo en crecimiento dentro del procesamiento de lenguaje natural, que permite determinar la polaridad de un texto (positivo, neutral o negativo). En esta práctica, utilizamos tres enfoques distintos: diccionarios, modelos de aprendizaje de máquina y redes neuronales.

Análisis basado en diccionarios:

Los diccionarios de sentimiento, como Harvard IV-4 y Opinion Lexicon, son colecciones de palabras etiquetadas con valores de polaridad. Estos diccionarios asignan un puntaje positivo o negativo a cada palabra, lo que permite calcular un puntaje total para el texto. Por ejemplo, si un texto contiene muchas palabras con puntuaciones negativas, se clasifica como negativo. Esta técnica es simple y rápida, pero su precisión puede verse limitada por la cobertura del diccionario y su capacidad para manejar palabras fuera del vocabulario. (Sawant & Priyanka , s.f.)

Modelos de aprendizaje de máquina:

Para el enfoque de aprendizaje de máquina, utilizamos algoritmos como regresión logística (LR), árboles de decisión (DT) y máquinas de soporte vectorial (SVM). Estos modelos requieren un preprocesamiento extenso del texto, incluyendo tokenización, lematización y vectorización utilizando técnicas como TF-IDF (Term Frequency-Inverse Document Frequency). El peso TF-IDF es una medida estadística que evalúa la importancia de una palabra en un documento dentro de un corpus. Aumenta proporcionalmente al número de veces que una palabra aparece en el documento y es atenuado por la frecuencia de la palabra en el corpus, proporcionando así un balance entre términos relevantes y comunes. (Dr.K.Sindhura)

- **Regresión Logística (LR):** Este modelo es un clasificador lineal que estima la probabilidad de que un ejemplo pertenezca a una clase determinada. Es eficiente y fácil de interpretar.
- **Árboles de Decisión (DT):** Este algoritmo crea un modelo de decisiones basado en características del texto, permitiendo la clasificación a través de un conjunto de reglas.
- **Máquinas de Soporte Vectorial (SVM):** Este método es eficaz en espacios de alta dimensión y utiliza funciones de núcleo para manejar datos no lineales, separando las clases mediante un hiperplano óptimo.

Redes neuronales:

En el enfoque basado en redes neuronales, se utilizaron embeddings de palabras y capas de redes neuronales profundas. Los embeddings de palabras, como Word2Vec y GloVe, convierten las palabras en vectores de alta dimensionalidad que capturan su significado contextual. Existen dos tipos principales de embeddings:

- **Embeddings preentrenados:** Estos son vectores preconstruidos a partir de grandes corpus de texto, que se utilizan directamente para representar palabras en nuestro modelo.
- **Embeddings aprendidos:** Estos vectores se entrenan conjuntamente con el modelo en el corpus específico, ajustándose mejor a las particularidades del conjunto de datos utilizado. (Awargal, s.f.)

Para la clasificación de sentimientos, se implementaron capas de redes neuronales utilizando frameworks como Keras y TensorFlow.

ADQUISICIÓN DE DATOS

Como se había solicitado en la practica se obtuvieron los datos de:

<https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews>, ahí pudimos encontrar un dataset con la siguiente estructura:

Ilustración 1 Dataset de Reviews

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Nota: Fuente Elaboración propia en python

ANALISIS EXPLORATORIO DE DATOS

Como se menciona anteriormente, se usaron tareas de exploración que nos brindaran información sobre el contenido del dataset en genral, para posteriormente trabajar sobre las opiniones que se encuentran en el dataset:

Identificamos la cantidad de filas y columnas:

Numero de filas: 568454

Numero de columnas: 10

1. Id
2. ProductId
3. UserId
4. ProfileName
5. HelpfulnessNumerator
6. HelpfulnessDenominator
7. Score
8. Time
9. Summary
10. Text

Identificamos los tipos de datos de cada columan:

La columna 'Id' es numérica.

La columna 'ProductId' es categórica.

La columna 'UserId' es categórica.

La columna 'ProfileName' es categórica.

La columna 'HelpfulnessNumerator' es numérica.

La columna 'HelpfulnessDenominator' es numérica.

La columna 'Score' es numérica.

La columna 'Time' es numérica.

La columna 'Summary' es categórica.
La columna 'Text' es categórica.

Tipos de dato de cada columna:

Id	int64
ProductId	object
UserId	object
ProfileName	object
HelpfulnessNumerator	int64
HelpfulnessDenominator	int64
Score	int64
Time	int64
Summary	object
Text	object

dtype: object

Tambien se identificaron la cantidad de instancias unicas de cada columna:

Columna: Id, Numero de instancias unicas: 568454
Columna: ProductId, Numero de instancias unicas: 74258
Columna: UserId, Numero de instancias unicas: 256059
Columna: ProfileName, Numero de instancias unicas: 218415
Columna: HelpfulnessNumerator, Numero de instancias unicas: 231
Columna: HelpfulnessDenominator, Numero de instancias unicas: 234
Columna: Score, Numero de instancias unicas: 5
Columna: Time, Numero de instancias unicas: 3168
Columna: Summary, Numero de instancias unicas: 295742
Columna: Text, Numero de instancias unicas: 393579

En resumen tenemos:

DATSET de 10 columnas

ID: Números enteros que identifican de manera única cada entrada en el conjunto de datos.

ProductId: Identificadores categóricos de productos, almacenados como cadenas de caracteres.

UserId: Identificadores categóricos de usuarios, almacenados como cadenas de caracteres.

ProfileName: Nombres de perfiles de usuarios, almacenados como cadenas de caracteres.

HelpfulnessNumerator: Números enteros que representan la cantidad de usuarios que encontraron útil la opinión.

HelpfulnessDenominator: Números enteros que indican la cantidad total de usuarios que dieron su opinión sobre si la revisión fue útil o no.

Score: Valoraciones del producto en una escala del 1 al 5, almacenadas como números enteros.

Time: Fecha y hora de la opinión, almacenadas como números enteros.

Summary: Resúmenes de las opiniones, almacenados como cadenas de caracteres.

Text: Opiniones completas, almacenadas como cadenas de caracteres.

PREPROCESAMIENTO

A continuación como se nos fue indicado nos deshicimos de las columnas que no aportaban información para nuestro análisis:

Dimensiones necesarias: 4

Dimensiones que se puede prescindir: 6

Dado que solo estamos realizando un análisis de sentimientos nos es conveniente usar únicamente las columnas:

"Score" dado que es el puntaje que da la opinion al producto en cuestion.

"Text" dado que contienen las opiniones que deseamos analizar.

Aunque son prescindibles para el análisis de sentimientos, podriamos mantener las siguientes columnas para poder interpretar los datos posteriormente si se llegase a necesitar:

"Product Id" dado que es el producto del que se esta dando la opinion.

"User Id" dado que es el usuario que da dicha opinion.

Obteniendo la estructura siguiente de nuestro dataset:

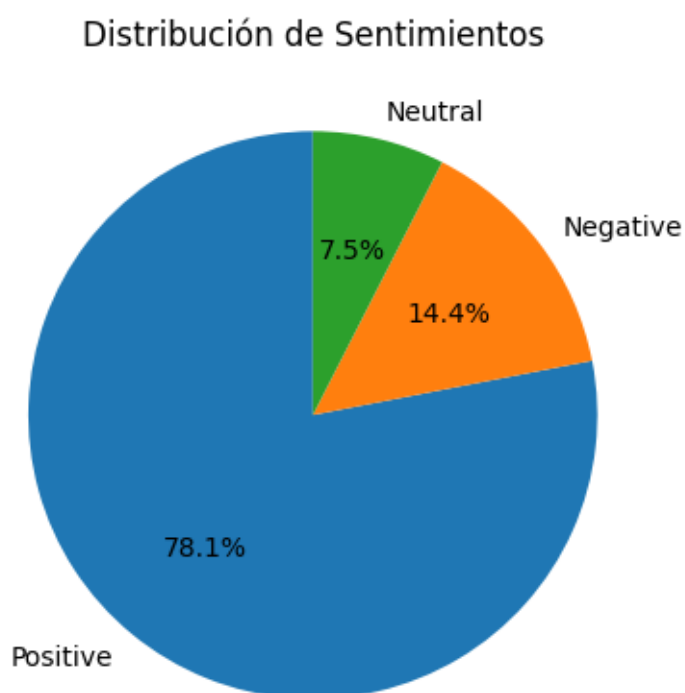
Ilustración 3 Dataset sin columnas innecesarias

	Summary	Text	Sentiment
0	Good Quality Dog Food	I have bought several of the Vitality canned d...	Positive
1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	Negative
2	"Delight" says it all	This is a confection that has been around a fe...	Positive
3	Cough Medicine	If you are looking for the secret ingredient i...	Negative
4	Great taffy	Great taffy at a great price. There was a wid...	Positive

Nota: Fuente Elaboración propia en python

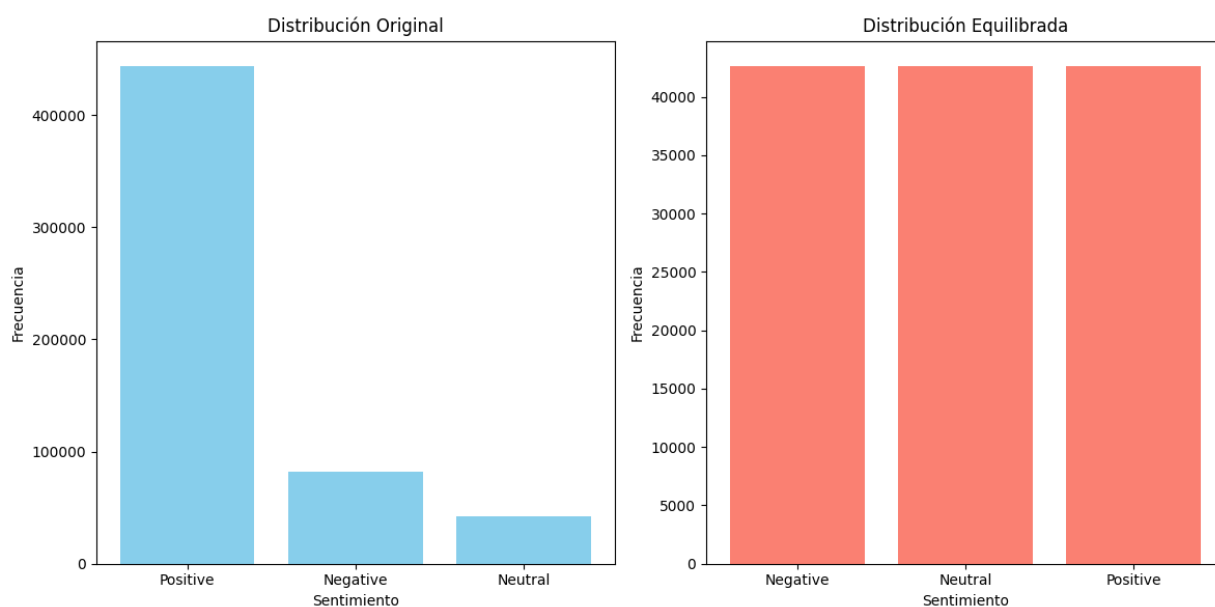
De igual forma se cuantifico la cantidad de sentimientos que habia en el dataset con ayuda de la columna sentiment:

Ilustración 4 Distribución de sentimientos



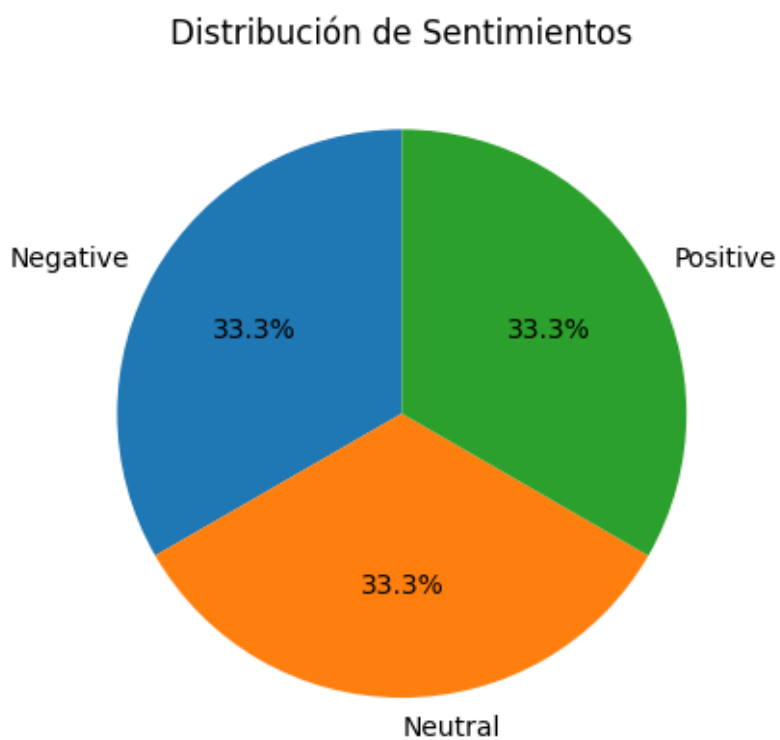
Nota: Fuente Elaboración propia en python

Ilustración 2 Antes y después del balanceo



Nota: Fuente Elaboración propia en Python

Ilustración 5 Antes y después del balanceo

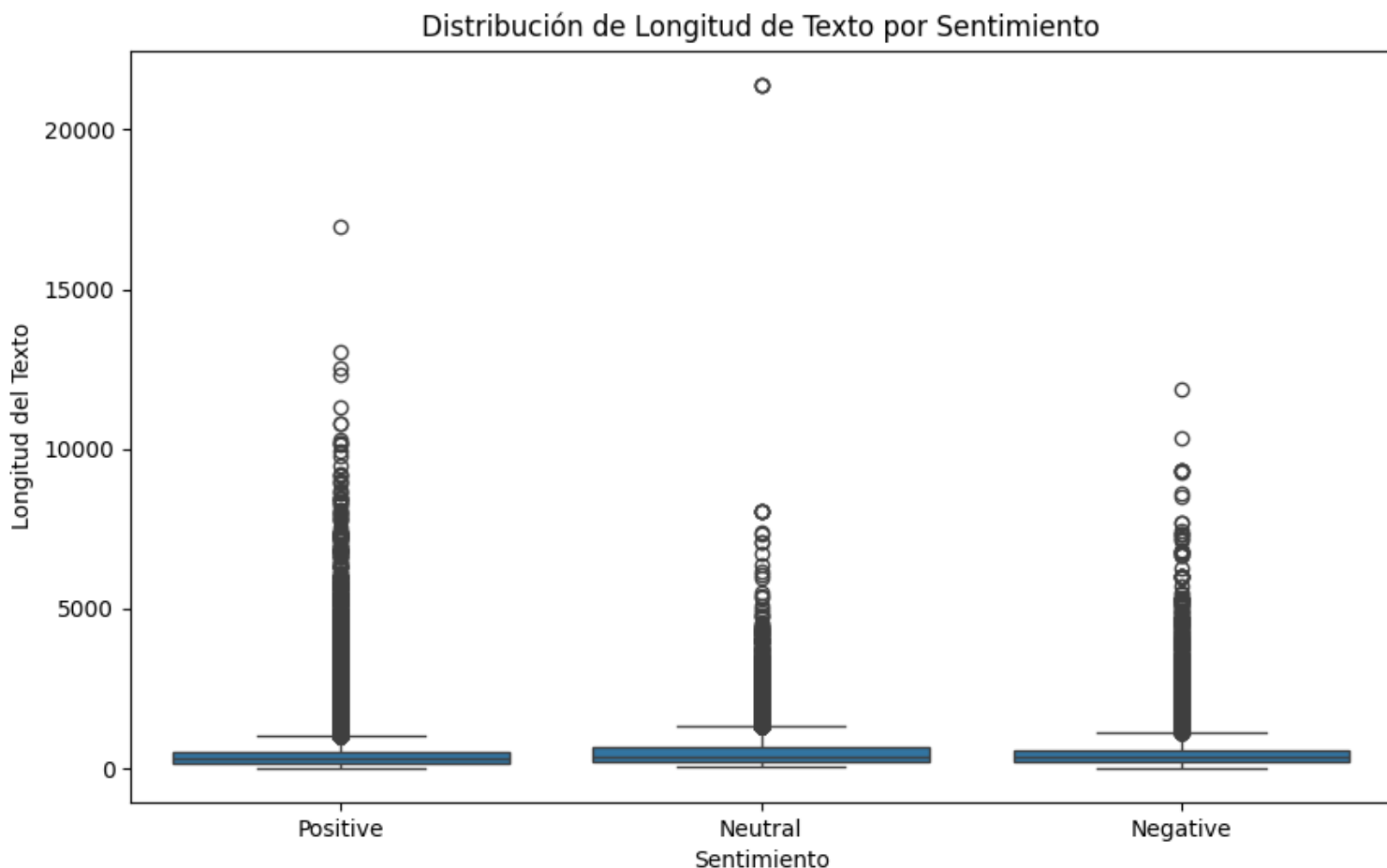


Nota: Fuente Elaboración propia en python

Tambien buscamos las palabras que conformaban cada sentimiento y su frecuencia, para darnos una idea de la distribución de palabras para nuestra aplicación:



Tambien observamos la longitud de texto por sentimiento, esto con la itención de observar si habia una diferencia grande la cual podria afectar nuestros entrenamiento:



Nota: Fuente Elaboración propia en python

Y para tener el dato, la cantidad de cada clase:

```
Sentiment
Positive      443777
Negative      82037
Neutral       42640
Name: count, dtype: int64
```

Hicimos el balanceo en cuanto a cantidad de clases para aligerar la dispersión del dataset, de forma general quedando:

Ilustración 3 Dataset con el balanceo de clases aplicado

	ProductId	UserId	Text	Sentiment	Text_Length
0	B001GB7IJI	A2AVRC44POYX0A	i love swedish fish but was horrified when i r...	Negative	312
1	B000G7YO2M	ALHLA8UE7JFQT	i ordered this for my mom - and my mom raved a...	Positive	216
2	B003VXL0V6	A26RIJE57A1XZE	bought this on the recommendation of my son. ...	Neutral	201
3	B004Y4RC4U	A3C1H6WWL9Q9KQ	this stuff is terrible...not much more to say....	Negative	167
4	B000EGZ97Y	A2LM2FT5JBZX8S	this is great rice, tender when cooked and sti...	Neutral	210

Nota: Fuente Elaboración propia en python

LIMPIEZA DE DATOS

Con ayuda de herramientas como Regex eliminamos palabras o datos que pudieran llegar a proporcionar ruido para el paso de modelado y entrenamiento, primero viendolo de forma manual para darnos una idea de que cambios deberian hacerse:

```
# Se ajusta el maximo de caracteres que se pueden mostrar para ver las opiniones completas
pd.set_option('display.max_colwidth', 500)

# Imprime las primeras 20 filas del DataFrame con el texto truncado
print(dfBalanced['Text'].head(50))

# Se revierte el maximo de caracteres a mostrar
pd.reset_option('display.max_colwidth')
```

```
0
i love swedish fish but was horrified when i received a bag of deceased fish! the bag was torn open in the box. some were
hard as rocks. others had melted into a bloody red mess. what are you made of that you can house both extremes in one ba
g? i'll never eat one again. i've switched to blue gummy sharks.
1
i ordered this for my mom - and my mom raved about it. because of the way it is packaged, she has even been able to share
some biscotti with her friends! this biscotti is addictive - bet you can't just eat 'one'...
2
bought this on the recommendation of my son. not a big fan of flavored coffee, but a cup a day of this lightly flavored co
conut coffee puts me in the islands. how can you go wrong with wolfgang puck?
3
this stuff is terrible...not much more to say....i always use the butterball kits for all my bigger meats but this one is j
ust bad..i guess i'll stick to the cajun kit
4
this is great rice, tender when cooked and sticky. but the price is way too expensive. i bought a bigger bag for the nearly
same price at the local market. look around your town for a far more reasonable price.
5
breakdown:<br />5/5 - flavor<br />5/5 - unsweetened<br />2/5 - dryness (very dry, even after baked into macaroons)<br
/>0/5 - texture (the worst, while the flakes are the right size they are not the texture of coconut, and very difficult to
eat)<br /><br />while the coconut taste is there, i mean wonderful real coconut not just coconut flavor, this product leave
```

Nota: Fuente Elaboración propia en python

- Obteniendo las siguientes conclusiones:
 - <\br> se presenta en varios comentarios y no aporta información: se eliminan
- Emoticonos, depende de la tecnica de analisis de snetimiento a usar, por ejemplo para los diccionarios que ya estan predefinidos, si no tienen o reconocen algun emoticono podria bajar la eficiencia, se porbara con ambos
- Contracciones del idioma ingles, se ampliaran
- cadenas de html, links y correos serán eliminados, se eliminaran al no agregar info
- simbolos y numeros, se eliminaran al no agregar info

Aplicamos normalización como stopwords, conversión a minúsculas, lematización y stemming

TRANSFORMACIÓN DE DATOS

Se aplicaron las codifiación necesarias y solicitadas, que se aplicaran en etapas posteriores.

MODELADO Y RESULTADOS

A conitnuación se presentán cada uno de los modelos entrenados con nuestro dataset y corpus procesado, junto con los resultados obtenido a partir de cada predicción resultantes sobre los bloques de testeo

Diccionarios

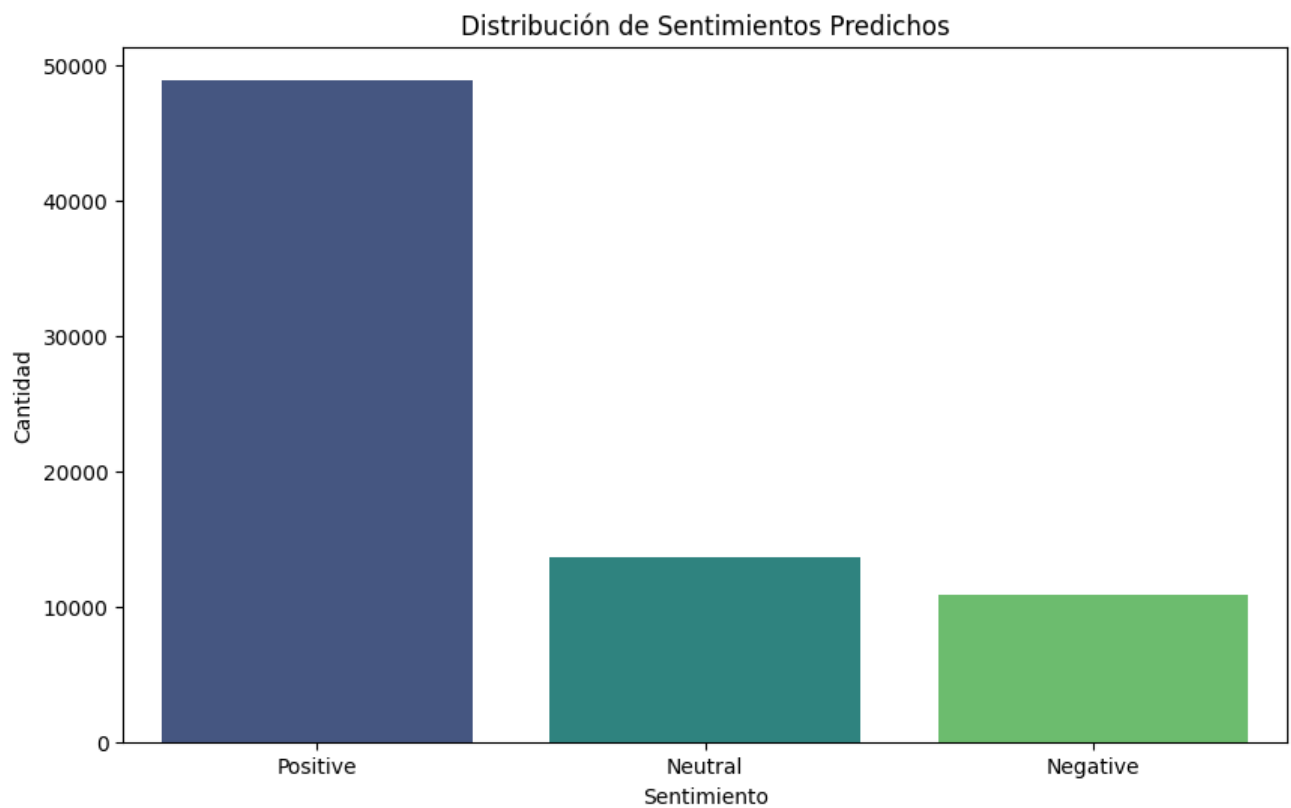
HARVARD

Ilustración 5 DATASET con la column de predicciones hechas por HARVARD

	Summary	Text	Sentiment	Text_Length	Text_lower	Text_stemmed	Text_lemmatized	Text_tokenized	SentimentH
0	Cheaper buying direct..	Buy it from guru directly much cheaper and al...	Negative	134.0	buy it from guru directly much cheaper and al...	buy it from guru directli much cheaper and als...	buy it from guru directly much cheaper and als...	['buy', 'it', 'from', 'guru', 'directly', 'muc...	Neutral
1	Pure?	Maybe theres no sugar but that does that make ...	Negative	559.0	maybe theres no sugar but that does that make ...	mayb there no sugar but that doe that make it ...	maybe there no sugar but that doe that make it...	['maybe', 'theres', 'no', 'sugar', 'but', 'tha...	Positive
2	little for the prie	I purchased this for my wife for Valentines Da...	Neutral	160.0	i purchased this for my wife for valentines da...	i purchas thi for my wife for valentin day my ...	i purchased this for my wife for valentine day...	['i', 'purchased', 'this', 'for', 'my', 'wife'...	Neutral
3	Good heat	Initially there was an element of the flavor t...	Neutral	361.0	initially there was an element of the flavor t...	initi there wa an element of the flavor that s...	initially there wa an element of the flavor th...	['initially', 'there', 'was', 'an', 'element', '...	Negative
4	Not the best pop chip flavor	While the chip is still light and crispy the f...	Neutral	202.0	while the chip is still light and crispy the f...	while the chip is still light and crisperi the f...	while the chip is still light and crispy the f...	['while', 'the', 'chip', 'is', 'still', 'light...	Positive

Nota: Fuente Elaboración propia en Python

Ilustración 9 RESULTADOS DE HARVARD



Nota: Fuente Elaboración propia en python

LEXICON

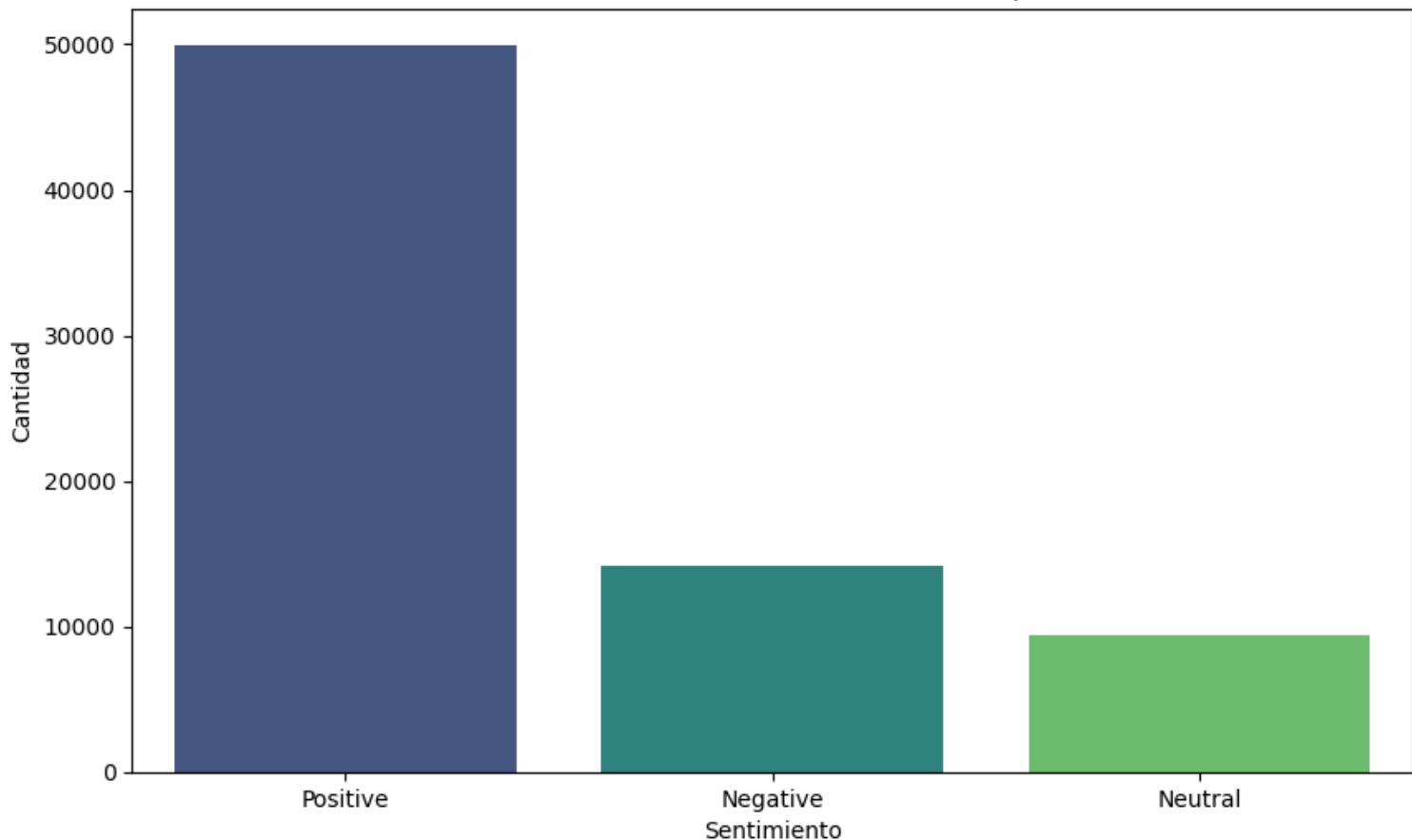
Ilustración 6 DATASET con la column de predicciones hechas por LEXICON

	Summary	Text	Sentiment	Text_Length	Text_lower	Text_stemmed	Text_lemmatized	Text_tokenized	SentimentOL
0	Cheaper buying direct..	Buy it from guru directly much cheaper and al...	Negative	134.0	buy it from guru directly much cheaper and al...	buy it from guru directli much cheaper and als...	buy it from guru directly much cheaper and als...	['buy', 'it', 'from', 'guru', 'directly', 'muc...	Positive
1	Pure?	Maybe theres no sugar but that does that make ...	Negative	559.0	maybe theres no sugar but that does that make ...	mayb there no sugar but that doe that make it ...	maybe there no sugar but that doe that make it...	['maybe', 'theres', 'no', 'sugar', 'but', 'tha...	Negative
2	little for the prie	I purchased this for my wife for Valentines Da...	Neutral	160.0	i purchased this for my wife for valentines da...	i purchas thi for my wife for valentin day my ...	i purchased this for my wife for valentine day...	['i', 'purchased', 'this', 'for', 'my', 'wife'...	Neutral
3	Good heat	Initially there was an element of the flavor t...	Neutral	361.0	initially there was an element of the flavor t...	initi there wa an element of the flavor that s...	initially there wa an element of the flavor th...	['initially', 'there', 'was', 'an', 'element', '...	Positive
4	Not the best pop chip flavor	While the chip is still light and crispy the f...	Neutral	202.0	while the chip is still light and crispy the f...	while the chip is still light and crspi the f...	while the chip is still light and crispy the f...	['while', 'the', 'chip', 'is', 'still', 'light...	Positive

Nota: Fuente Elaboración propia en Python

Ilustración 9 RESULTADOS DE HARVARD

Distribución de Sentimientos Predichos usando Opinion Lexicon



Nota: Fuente Elaboración propia en python

Modelos de Machine learning

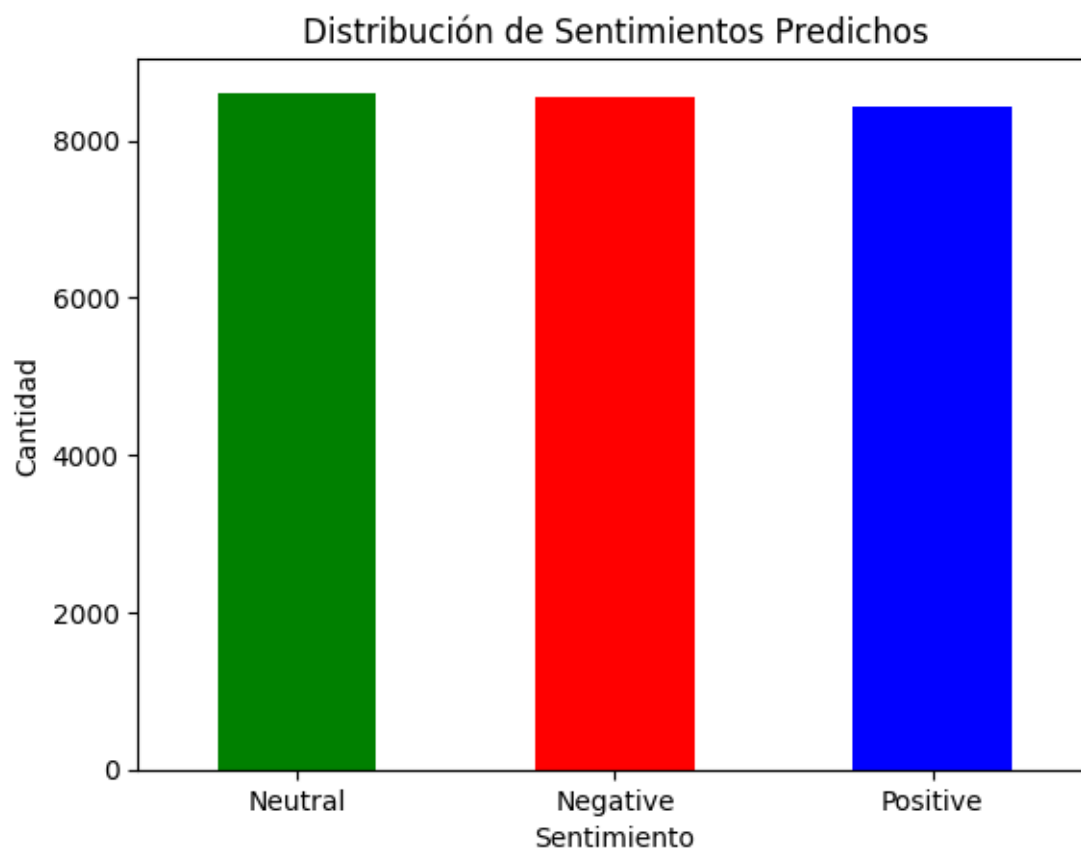
Regresión Logistica

Cross-Validation Scores Logistic Regression with TF-IDF: [0.74864403 0.74947471 0.74825312 0.7473247 0.74409968]

Cantidad de cada sentimiento según las predicciones:

Neutral: 8606
Negative: 8542
Positive: 8434

Ilustración107 Resultado de Regresión Logística



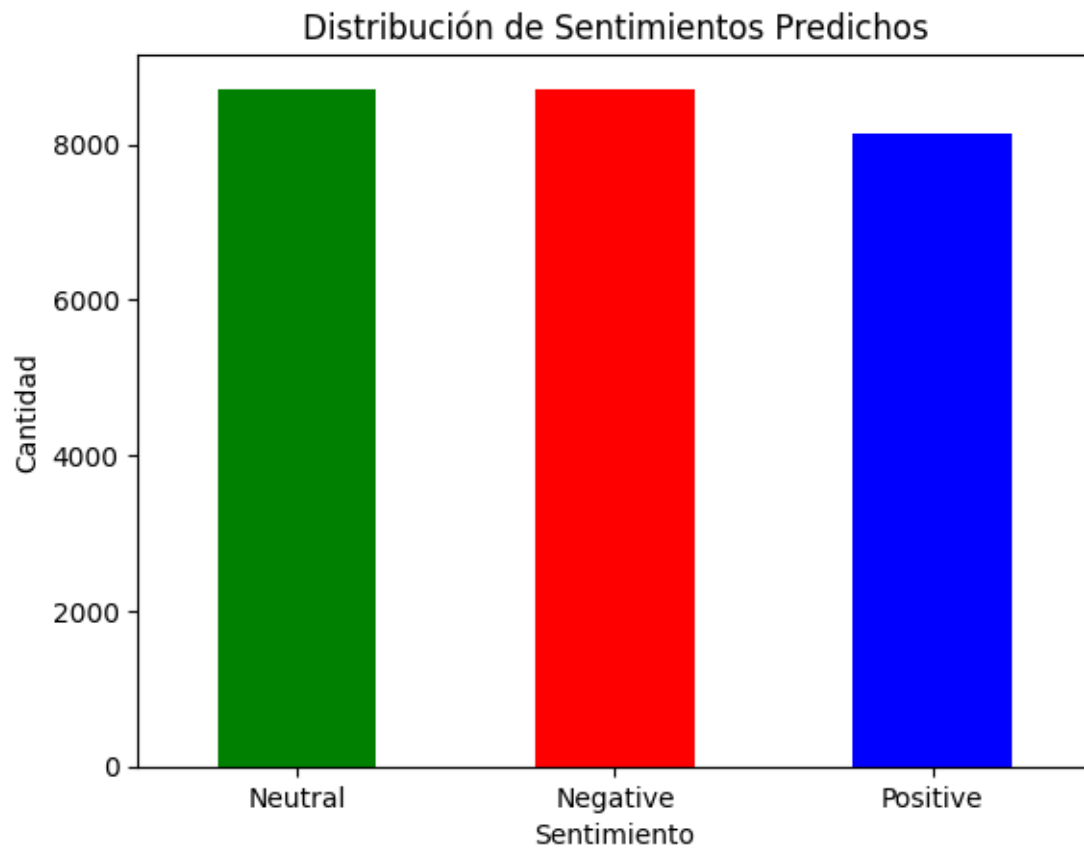
Nota: Fuente Elaboración propia en Python

Arboles de decisión

Cross-Validation Scores Decision Tree with TF-IDF: [0.63879795 0.64500366 0.6354752 0.63469338 0.63361837]

Cantidad de cada sentimiento según las predicciones:

Neutral: 8719
Negative: 8712
Positive: 8151



Nota: Fuente Elaboración propia en Python

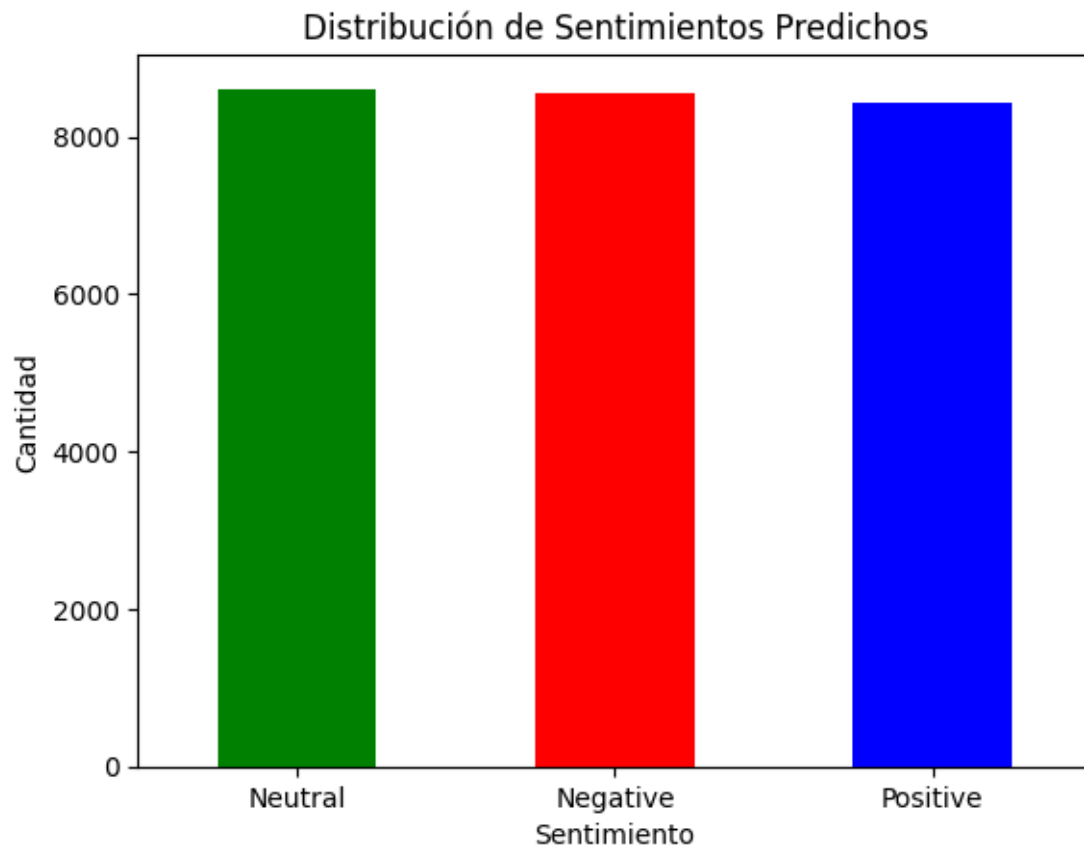
Support Vectorial Machine

Número de filas: 36768

Número de columnas: 8

Cross-Validation Scores SVM with TF-IDF: [0.74808771 0.73380928 0.7336393 0.7417984 0.73818429]

Test Accuracy SVM with TF-IDF: 0.7472124014141963



Nota: Fuente Elaboración propia en Python

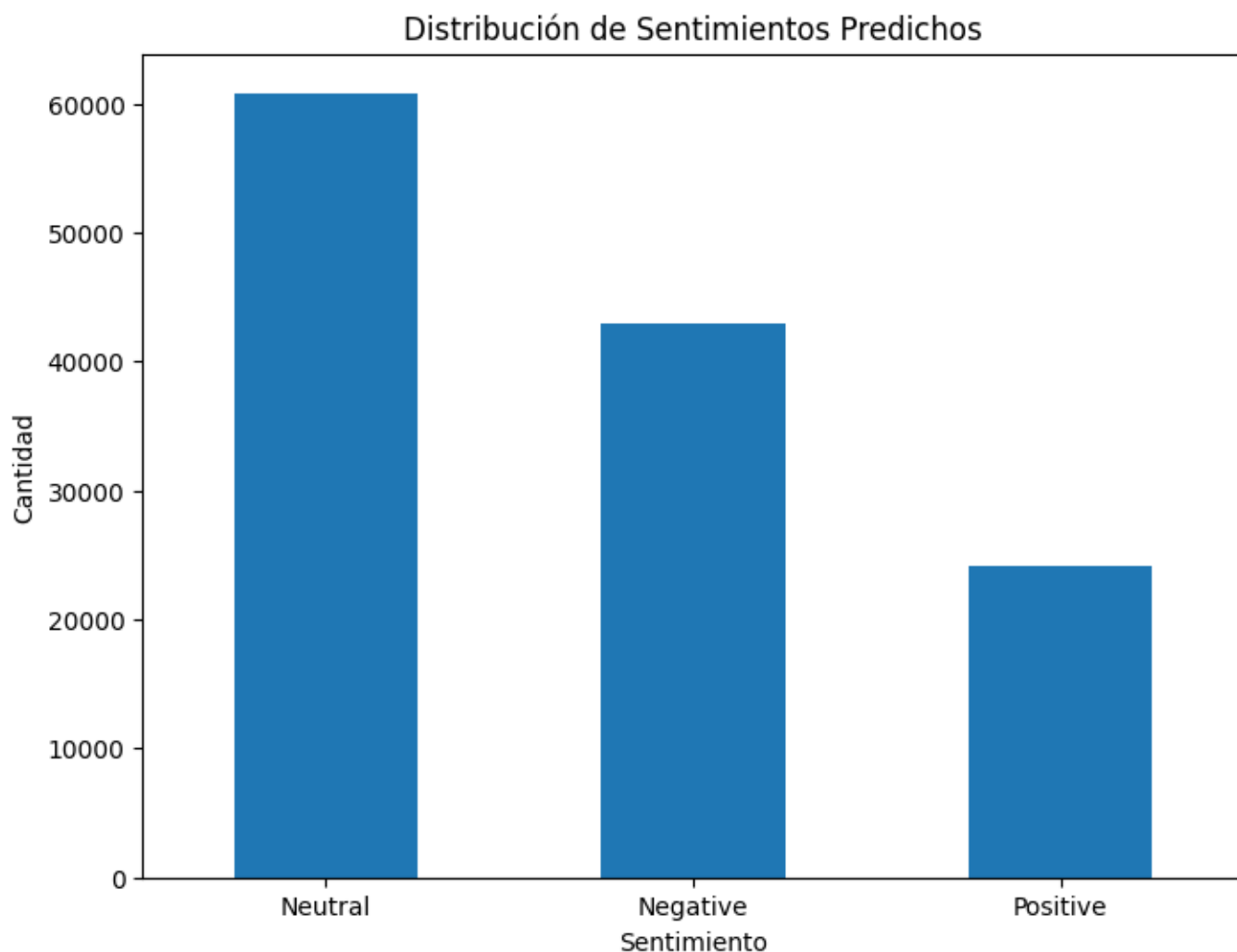
Redes Neuronales

Capa Preentrenada

```
Epoch 1/4
342/342 [=====] - 917s 3s/step - loss: 0.7387 - accuracy: 0.7355
Epoch 2/4
342/342 [=====] - 906s 3s/step - loss: 0.7255 - accuracy: 0.7360
Epoch 3/4
342/342 [=====] - 910s 3s/step - loss: 0.7110 - accuracy: 0.7377
Epoch 4/4
342/342 [=====] - 908s 3s/step - loss: 0.7014 - accuracy: 0.7405
171/171 [=====] - 126s 734ms/step - loss: 0.7015 - accuracy: 0.7418
Fold 3/3:
Tamaño conjunto de entrenamiento: 10928
Tamaño conjunto de validación: 5464
Precisión en el conjunto de validación: 0.7418

129/129 [=====] - 96s 744ms/step - loss: 0.7605 - accuracy: 0.7333
Precisión en el conjunto de prueba: 0.7333
El promedio en kfolds es: 0.7374328772226969

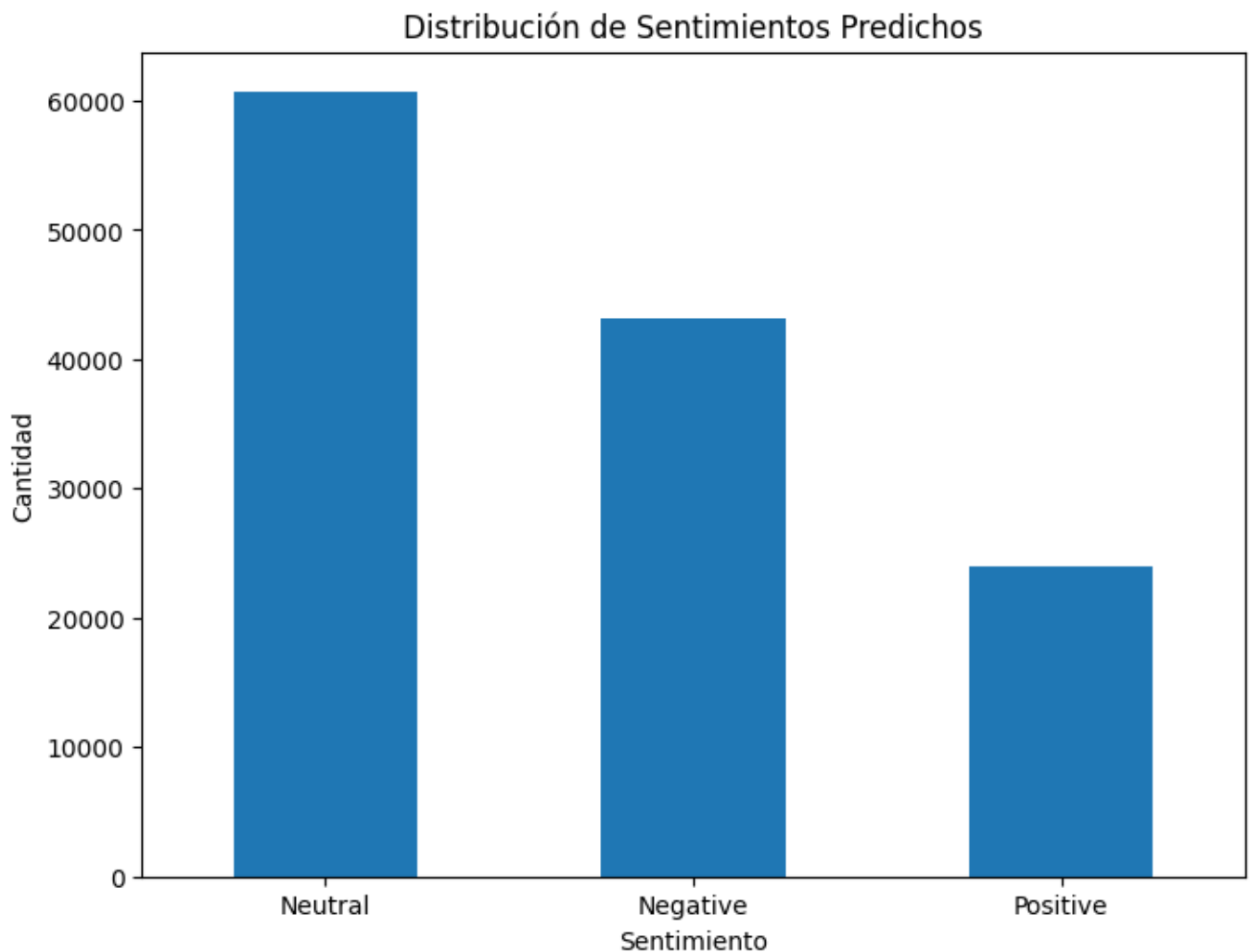
Neutral      60813
Negative     42948
Positive     24159
```

Nota: Fuente Elaboración propia en python

Sobre el mismo corpus

```
Epoch 1/4
569/569 [=====] - 1378s 2s/step - loss: 1.0630 - accuracy: 0.4122
Epoch 2/4
569/569 [=====] - 1387s 2s/step - loss: 1.0519 - accuracy: 0.4219
Epoch 3/4
569/569 [=====] - 1371s 2s/step - loss: 1.0437 - accuracy: 0.4299
Epoch 4/4
569/569 [=====] - 1395s 2s/step - loss: 1.0355 - accuracy: 0.4349
1066/1066 [=====] - 175s 164ms/step - loss: 1.0440 - accuracy: 0.4360
Fold 3/3:
Tamaño conjunto de entrenamiento: 68224
Tamaño conjunto de validación: 34112
Precisión en el conjunto de validación: 0.4360
Neutral      60730
Negative     43190
Positive     24000
```



Nota: Fuente Elaboración propia en python

Conclusión.

Bien, algo bastante curioso es que ninguna distribución se parece entre cada técnica, siendo que parece que la más cercana a la realidad fueron los modelos de machine learning, siendo la más estable; aunque también se debe mencionar que por cuestiones de computo y de tiempo no fue posible darle un mejor entrenamiento a las redes, por lo que podría haber ocasionado que no tuviera mejor rendimiento que las demás.

Los diccionarios parecen tener una preferencia por los sentimientos positivos, lo que podría ser la expresión del poco entendimiento que estas técnicas presentan a cosas como el sarcasmo o situaciones más complejas o abstractas.

Mientras que los modelos de machine learning no presentaron preferencia por algún sentimiento en específico, siendo que lograron ser los más apegados a la realidad, por el balanceo de clases, sin embargo también sería recomendable que para futuras aplicaciones ver como se comporta con otros o nuevos corpus.

Las redes neuronales algo curioso es el hecho de que tienen preferencia por el sentimiento neutral, aunque como se menciona es por la falta de recursos, que no se logro una mayor eficiencia en los modelos.

Como ultima mención queda decir que tanto las tareas de preprocesamiento, transformación, limpieza, modelado presentan tantas variables que deben ser

controladas de forma que nuestro corpus, es decir nuestro datos que al final de cuentas son nuestra representación de la realidad, deben ser armónico con nuestro modelo y la aplicación para poder lograr resultados optimos.

Referencias

- (s.f.).
- AMAZON. (s.f.). *¿Qué es el análisis de opiniones?* Obtenido de AMAZON AWS: <https://aws.amazon.com/es/what-is/sentiment-analysis/>
- Awargal, A. (s.f.). *¿Qué modelos de PLN son más efectivos para el análisis de sentimientos?* Obtenido de LINKEDIN: <https://www.linkedin.com/advice/0/what-nlp-models-most-effective-sentiment-dbtcf>
- Baruni, J. (2020). Keyphrase Extraction from Document Using RAKE and TextRank Algorithms. *Department of Computer Science, Bishop Heber College (Autonomous)*.
- BEMBADE, D. (s.f.). KAGGLE. Obtenido de <https://www.kaggle.com/discussions/general/419664>
- Chandrakant, K. (18 de March de 2024). *Baledung*. Obtenido de <https://www.baeldung.com/cs/automatic-keyword-keyphrase-extraction>
- Cruz, R. D. (17 de 12 de 2023). *MEDIUM*. Obtenido de Sentiment Analysis Using Natural Language Processing (NLP): <https://medium.com/@robdelaacruz/sentiment-analysis-using-natural-language-processing-nlp-3c12b77a73ec>
- Devika, R. (2021). A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. *IEEE Access (Volume: 9)*.
- Dr.K.Sindhura. (s.f.). SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING AND. *JSCT*.
- Fesenko, P. (5 de Octubre de 2023). *Best open-source models for sentiment analysis – Part 1: dictionary models*. Obtenido de MEDIUM: <https://medium.com/@pavlo.fesenko/best-open-source-models-for-sentiment-analysis-part-1-dictionary-models-ece79e617653>
- Hasan, K. S. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. *Human Language Technology Research Institute*.
- Mayo, M. (26 de Octubre de 2022). *Getting Started with Automated Text Summarization*. Obtenido de <https://www.kdnuggets.com/2019/11/getting-started-automated-text-summarization.html>
- Sarkar, D. (2019). *Text Analytics with Python*. Apress.
- Sawant, O., & Priyanka , G. (s.f.). *Linkedin*. Obtenido de ¿Cuáles son las técnicas más efectivas para normalizar datos de texto?: <https://www.linkedin.com/advice/1/what-most-effective-techniques-normalizing-text-s3dkc>
- Shahul. (4 de Septiembre de 2023). *neptune.ai*. Obtenido de <https://neptune.ai/blog/exploratory-data-analysis-natural-language-processing-tools>
- T, P. (26 de Apr de 2024). *analyticsvidhya*. Obtenido de <https://www.analyticsvidhya.com/blog/2022/03/keyword-extraction-methods-from-documents-in-nlp/>
- Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de un red social para python. *Simposio Argentino de Inteligencia Artificial*.
- Wei, D. (s.f.). *Demystifying Neural Networks: Sentiment Analysis*. Obtenido de MEDIUM: <https://medium.com/@weidagang/demystifying-neural-networks-sentiment-analysis-7944e1e97e9d>
- Zhang, L. (s.f.). Deep Learning for Sentiment Analysis: A Survey. *ARXIV*.