

Análisis de Sentimientos Aplicando Aprendizaje Profundo

Ma. Luisa Argáez Salcido

Maestría en Ciencia de Datos

Facultad de Físico-Matemáticas, UANL

4to. Tetramestre — Procesamiento y clasificación de datos

Resumen

Debido a la gran cantidad de reseñas de productos generadas no es posible analizarlas y tomar decisiones acertadas rápidamente. El análisis de sentimientos es una técnica que ayuda a conocer si determinado texto presenta una emoción positiva o negativa. Por tanto, se contruyó una Red Neuronal Recuerrente de Memoria de Largo y Corto Plazo que clasifico las reseñas respecto al tipo de emoción que presentaban obteniendo una exactitud del 74,3% con el conjunto de prueba. Esto permite ayudar a empresas que cuentan con reseñas a tomar mejores decisiones.

Introducción

Actualmente la generación de datos ha sido impulsada por el acelerado desarrollo del internet de las cosas, lo cual ha provocado que se genere gran cantidad de información y es fundamental recopilarla, limpiarla y procesarla para extraer aquellos indicadores que faciliten y mejoren la toma decisiones haciéndola más rápida y certera (1).

Uno de los medios en donde se genera gran cantidad de información es a través del texto, el cual es posible encontrarlo en diversos formatos como reseñas. En estas es se expresa una emoción u opinión respecto a alguna persona, película o producto y es importante conocer si la emoción que predomina es positiva o negativa respecto al objeto para tomar mejores decisiones. Sin embargo, en ocasiones no es posible analizar todas las reseñas de una forma eficiente por el gran tamaño de la información dificultando tomar decisiones correctas para las empresas (5).

A pesar de que trabajar con texto puede representar un reto, existen múltiples beneficios que las empresas pueden obtener a través del estudio del texto como la aplicación del análisis de sentimientos. Entre algunos ejemplos de la aplicación del análisis de sentimientos se encuentran el conocer si el lanzamiento de un producto esta cumpliendo con las expectativas de los clientes, o bien si un producto existente en el mercado sigue teniendo las mismas opiniones de cuando se lanzó.

La información utilizada para el presente análisis proviene de una empresa popular llamada "Amazon", la cual actualmente es una empresa de comercio electrónico de origen estadounidense que brinda una gran cantidad y variedad de productos. Dichos productos son evaluados por sus compradores por medio de reseñas y calificaciones.

En este poster, propone una metodología para identificar y clasificar las reseñas por medio de la emoción que presentan, la cual puede ser positiva o negativa, comparando diferentes tipos de preprocesamiento de los datos como lo son "Lematización"(Lematization) y de "Análisis de datos raíz"(Stemming"). La Lematización propone convertir las palabras a su forma base; mientras que el análisis de datos raíz reduce la palabra a su raíz eliminando sufijos o prefijos.

El modelo que se utilizó fue una Red Neuronal Recurrente ("Recurrent Neural NetworkRNN) llamada Memoria de Largo y Corto Plazo ("Long Short Term Memory"LSTM). Este tipo de RNN tiene como objetivo resolver el problema del desvanecimiento del gradiente (2).

1. Objetivo

1.1. Objetivo principal

Analizar, procesar y clasificar las reseñas conforme al tipo de emoción que expresa utilizando una RNN de tipo LSTM, la cual presenta un rendimiento en la exactitud de al menos del 70 % en el conjunto de prueba. Este objetivo se llevará a cabo bajo dos perspectivas de preprocesamiento de los datos, utilizando las técnicas de preprocesamiento de "Lematización" y de "Análisis de datos raíz".

1.2. Objetivos secundarios

El primer objetivo secundario es limpiar la base de datos de tal manera que se remuevan e intercambiar aquellos caracteres que no son deseables en el texto como signos de puntuación, caracteres que no son alfanuméricos, entre otros. Además de convertir el texto en minúsculas y eliminar aquellas palabras que no agregan significado al texto. El segundo objetivo es analizar el rendimiento del algoritmo de aprendizaje profundo mediante "Stemming" y lematización del texto. Finalmente, es realizar una arquitectura de una RNN de tipo LSTM, tal que logre una buena separación entre ambos grupos de emociones representadas en el texto.

2. Metodología

2.1. Análisis Exploratorio de Datos

Se estudió un conjunto de datos de un millón de muestras, el cual consiste en reseñas etiquetadas como positiva o negativa. Se analizaron varias métricas por reseña como la longitud de caracteres implicados en la texto, cuantas palabras contenia la reseña y la longitud promedio de las palabras. Esto se contrastó después de ser procesado al omitir las palabras que no agregan valor a la oración, también conocidas en inglés como "Stopwords" para ambos tipos de emoción. A continuación se observa el antes y el después utilizando gráficos de violín.

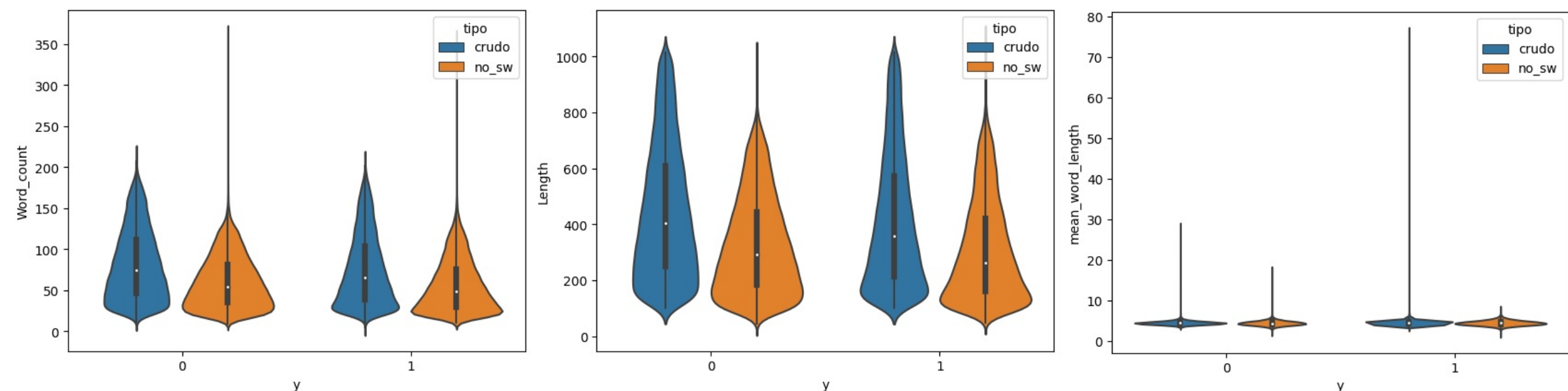


Figura 1: Gráficos de violín para comparación de datos crudos y después de omitir "Stopwords" por clase

2.2. Limpieza de datos

Los datos crudos fueron procesados con el fin de concentrar solo la información que es relevante en la reseña. Para lo cual se siguieron los siguientes pasos: en primer lugar se convirtió todo el texto a

minúsculas, se eliminaron caracteres que no fueran dígito, letra o espacio en blanco. Posteriormente, se suprimieron las palabras conocidas como stopwords.

2.3. Procesamiento de los datos

Una vez limpio el texto, se procedió a realizar dos preprocesamientos de datos: "Stemming" y lematización. El "Stemming" consiste en eliminar los prefijos y sufijos de las palabras con el objetivo de representarla como su raíz, sin embargo esto puede generar palabras no válidas. Por otro lado, la lematización se basa en reducir las variantes de las palabras a su raíz común o lexema. Además, se separo el conjunto de datos en entrenamiento y prueba en un porcentaje de 80 % y 20 % respectivamente.

2.4. Modelo "Long Short Term Memory

El modelo se construyó utilizando la librería de "Keras" en Python. El modelo tuvo la siguiente arquitectura: primero, se agregó una capa de incrustación ("Embedding"), posteriormente se apilaron dos capas "LSTM" de 50 y 10 unidades cada una. Después se agregó una capa de abandono ("Dropout") con una tasa del 50 %. Luego se agregó una capa de 8 neuronas y finalmente, una capa de salida sigmoideal con una neurona. El modelo se compilo utilizando el optimizador "Adam", la función de pérdida "Entropía cruzada binaria"("Binary Crossentropy"), se evaluó el rendimiento del modelo con la métrica de exactitud. El modelo se entrenó durante 20 épocas con un lote de 32 muestras con datos de entrenamiento y se monitoreo con la métrica de pérdida y exactitud con un conjunto de validación. Se utilizó una técnica que consiste en si la pérdida en el conjunto de validación no mejora durante 3 épocas consecutivas, la tasa de aprendizaje se reduce a la mitad (3) (4). Esta metodología se realizo dos veces, una donde los datos eran preprocesados utilizando "Stemming" y con lematización.

3. Resultados

En primer lugar, se observa que realizar una limpieza de datos es un paso fundamental para el rendimiento del modelo ya que logra eliminar aquellas palabras que no suman valor además de reducir la cantidad de palabras a procesar. En segundo lugar, se observa que los datos bajo la metodología de "Análisis de datos raíz" logró una exactitud del 74,34 % con el conjunto de prueba y un 76 % con el conjunto de entrenamiento. Por otro lado, es posible ver que los datos bajo la metodología de "Lematización" logró una exactitud del 73,66 % con el conjunto de prueba y un 75,65 % con el conjunto de entrenamiento.

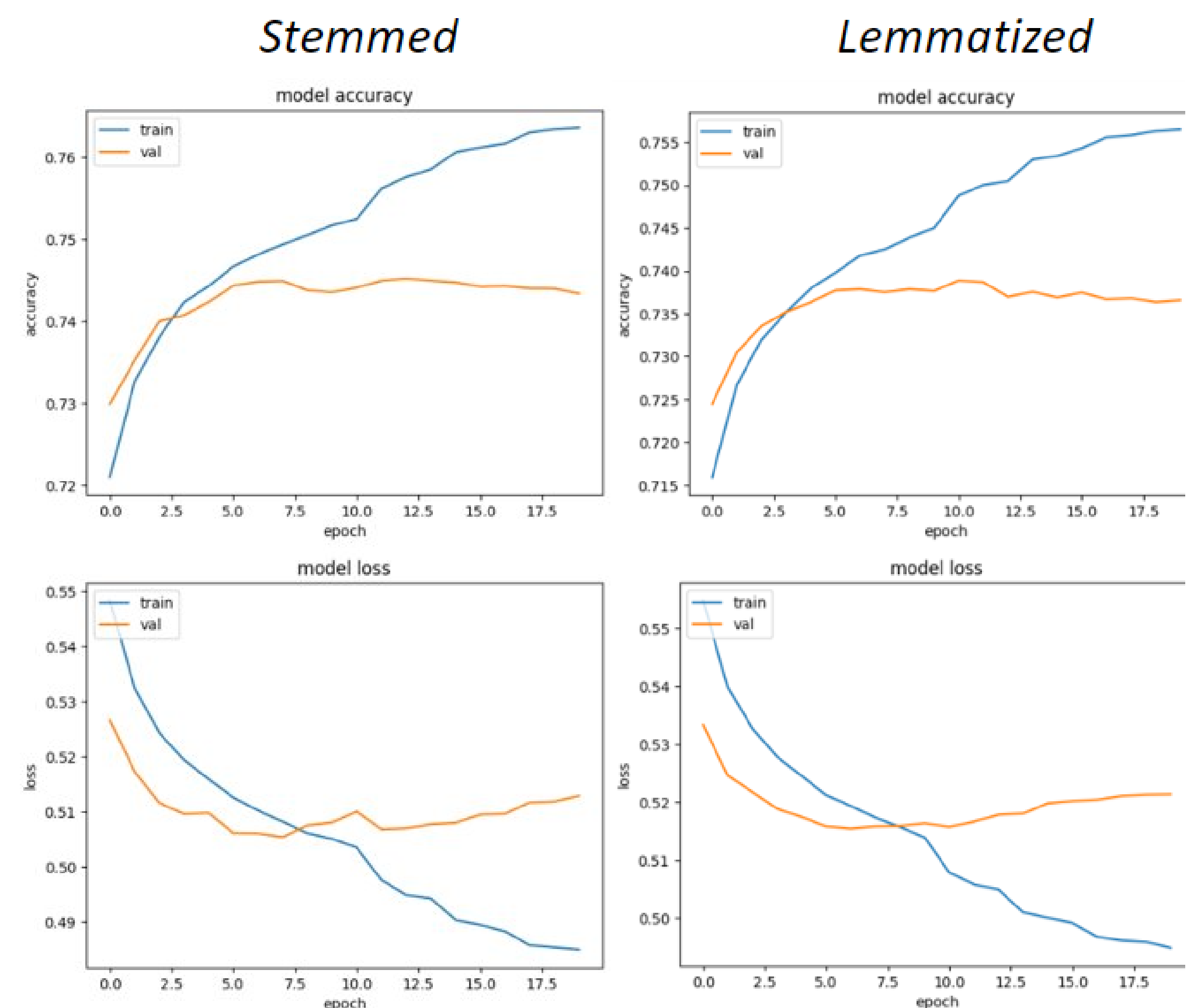


Figura 2: Comparación del modelo con el preprocesamiento de "Stemming" y "Lemmatization"

4. Conclusiones

Se logró el objetivo principal que fue alcanzar una exactitud de más del 70 % con el conjunto de prueba mediante la metodología de "Stemming" y de "Lematización". Sin embargo, la metodología con mejores resultados fue la de "Stemming" con 1,3 % de exactitud mejor. Por otro lado, es importante mencionar que la limpieza, preprocesamiento y la arquitectura del modelo fueron piezas fundamentales para lograr el objetivo principal. En conclusión, según los resultados encontrados es necesario realizar una limpieza exhaustiva, un buen preprocesamiento incluyendo la tokenización y truncamiento de los vectores generados. Finalmente, la arquitectura del modelo LSTM logró buenos resultados utilizando la capa de abandono, sin embargo se logra apreciar un sobreentrenamiento en las últimas épocas.

5. Trabajo a futuro

Se propone explorar otras alternativas con mayor poder de computo para entrenar el modelo con todas las muestras.

6. Bibliografía

Bibliografía - (Click)