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 texto generada en forma de reseñas de productos no es posible analizarla y tomar decisiones acertadas rápidamente. El análisis de sentimientos es una técnica de la rama del procesamiento del lenguaje natural que ayuda a deterinar si texto presenta una emoción positiva o negativa. Por tanto, se ha construido una Red Neuronal Recuerrente llamada de Memoria de Largo y Corto plazo, la cual obtiene una exactitud con el conjunto de prueba del $75\,\%$. Lo cual permite ayudar a empresas que cuenta con reseñas a analizarlas eficientemente para 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, almacenarla, limpiarla y procesarla para extraer aquellos indicadores y elementos que faciliten y mejoren la toma decisiones haciéndola más rápida y certera. En particular, 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, opiniones, noticias, entre otros. Hablando especificamente de las reseñas, es posible decir que en ellas 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 leer, procesar y analizar todas las reseñas de una forma rápida y certera por el gran tamaño de la información, lo cual dificulta tomar decisiones correctas para las empresas.

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 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. Es importante mencionar, que la página web de Amazon publica millones de productos, los cuales son evaluados por sus compradores por medio de reseñas y calificaciones.

En este poster, lo que se intenta resolver con la metodología establecida es identificar y clasificar las reseñas por medio de la emoción que presentan, la cual puede ser positiva o negativa y 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 Neuronal Recurrente (*“Recurrent Neural Network-NN*) 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. En este caso, aprende de las sequencias presentes en las reseñas como texto para lograr captar relaciones en las reseñas que presentan emociones positivas y negativas.

1. Objetivo

1.1. Objetivo principal

Analizar, procesar y clasificar las reseñas conforme al tipo de emoción que presenta utilizando un tipo de red neuronal recurrente llamada memoria de largo y corto plazo, la cual presente 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 “Lematizaciónz 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 o intercambiar aquellos caracteres que no son deseables en el texto como signos de puntuación, enlaces, caracteres que no son alfanuméricos, números, 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 *steamming* y lematización del texto. Además de tokenizarlo y que todos los vectores que representan al texto, sean de la misma longitud utilizando *padding*.

Finalmente, es realizar una arquitectura de una Red Recurrente de Memoria de Largo y Corto Plazo, también conocidas en inglés como *Long Short Term Memory, LSTM*), tal que logre una buena separación entre ambos grupos de emociones representadas en el texto.

2. Metodología

Se estudió un conjunto de datos el cual consiste en una columna que contiene las reseñas de productos seguido de otra columna que contiene la etiqueta marcando ”1”si representa una emoción positiva o un ”0”si indica una emoción negativa. El tamaño de la muestra es de 4 millones de registros.

2.1. Análisis Exploratorio de Datos

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 contrasto 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.

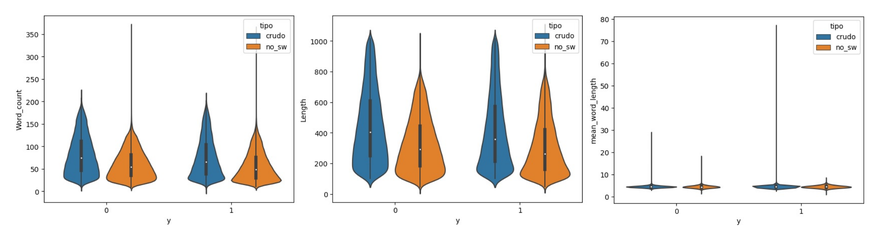

Figura 1: Gráficos de violín para comparación de datos crudos y después de omitir *stopwords* por clase

2.2. Limpieza de datos

Los datos crudos fueron procesados con el fin de resumir y concentrar solo la información que es relevante para la reseña. Para lo cual se siguieron los siguientes pasos: en primer lugar fue necesario convertir todo el texto a minusculas, eliminación de cualquier caracter que no sea un dígito, letra o espacio en blanco. Posteriormente, se suprimieron las palabras que no son relevantes en el texto y que no suman información a la reseña, este tipo de palabras son conocidas como *stopwords*.

2.3. Procesamiento de los datos

Una vez limpio el texto, se procedió a realizar dos preprocesamientos de datos: *steamming* y lematización. El steamming consiste en eliminar los afijos 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. 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 construyo utilizando la libreria de “Keras”. Se apilaron las siguientes capas. En primer lugar, se agrega una capa de incrustación (“Embedding”), posteriormente se apilaron dos capas “LSTM”de 50 y 10 unidades cada una. Después se agrego una capa de abandono (“Dropout”) con una tasa del $50\,\%$. Luego se agregó una capa de 8 neuronas y finalmente, una capa de salida sigmoidal con una neurona. El modelo se compilo utilizando el optimizador “Adam”, la función de pérdida “Entropía cruzada binaria”(“binary crossentropy”), y se evaluó el rendimiento del modelo con la métrica de exactitud. El modelo se entrenó durante 20 épocas con datos de entrenamiento y se monitoreo con la métrica de pérdida y exactitud con un conjunto de validación. Se utilizo una técnica que consis en que si la pérdida en el conjunto de validación no mejora durante 3 épocas consecutivas, la tasa de aprendizaje se reduce a la mitad. Esto ayuda a mejorar el entrenamiento del modelo, especialmente si se encuentra en una región de convergencia difícil en el espacio de optimización. Esta metodología se realizo dos veces, una donde los datos eran preprocesados utilizando “Stemmingz otra donde los datos eran preprocesados 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.

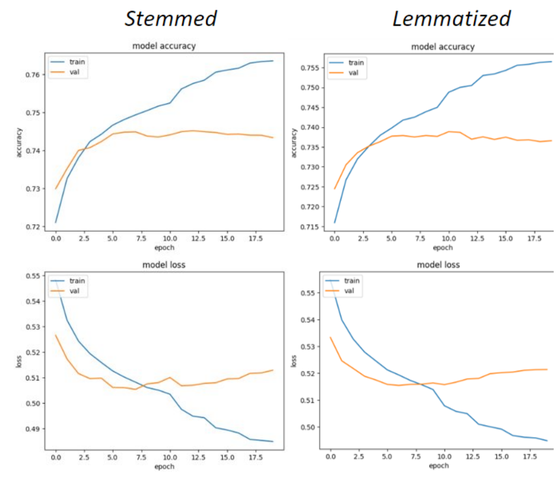

Figura 2: Comparación del modelo con el preprocesamiento de *Stemmed* y *Lemmatization*

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 “Stemmed”. Además que con la metodología de “Lematización”también se logró cumplir con este objetivo principal. En este caso se observo menos sobreentrenamiento bajo el preprocesamiento de análisis de datos raíz. 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 exahustiva, 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 epócas.