Análisis de Sentimientos Aplicando Aprendizaje Profundo

Ma. Luisa Argáez Salcido

Maestría en Ciencia de Datos

Facultad de Físico-Matemáticas, UANL

Tetramestre 4 — Procesamiento y clasificación de datos

Resumen

Se escribe hasta el final

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 en 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 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. Otra aplicación es en las inversiones, en donde la minería de opiniones brinda información valiosa de cómo las personas en general y expertos de la bolsa perciben los activos y las situaciones globales; un ejemplo es como la decisión de algún presidente puede trascender en algún conflicto afectando las inversiones de dicho país.

La información utilizada para el análisis proviende de empresa de comercio bastante popular llamada .Amazon". Actualmente, Amazon es una empresa de comercio electrónico de origen estadounidense que brinda una gran cantidad y variedad de productos. A través de los años, ha crecido exponencialmente y actualmente también ofrece servicios basados en la web, realiza dispositivos electrónicos y ofrece servicios de auto-publicación. 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.

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 que tenga un rendimiento en la exactitud de al menos del $70\,\%$ en el conjunto de prueba.

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 finalmente que todos los vectores que representan al texto, sean de la misma longitud utilizando *padding*.

Finalmente, es realizar una arquitectura de la red recurrente de memoria de largo y corto plazo tal que logre una buena separación entre 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 realizaron 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 como *stopwords* para ambos tipos de emoción. A continuación se observa el antes y el después utilizando gráficos de violín.

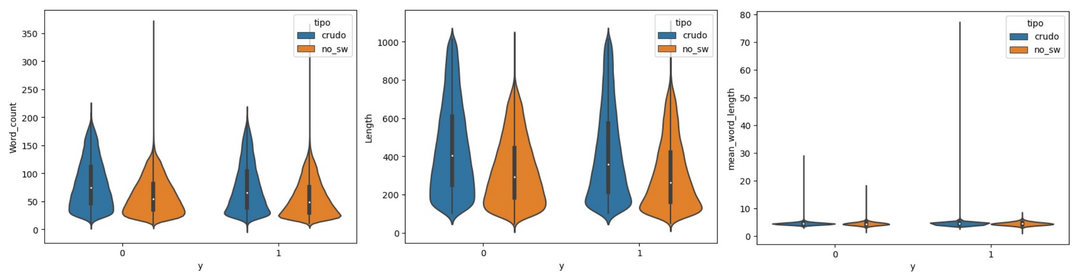

Figura 1: Gráficos de violín para comparación de datos crudos y después de omitir *stopwords* por clase

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 luegar 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.