Universidad Politécnica
de Madrid

Escuela Técnica Superior de

Ingenieros Informáticos

Máster en Ingeniería Informática

Sistemas Inteligentes

Procesamiento Natural del Lenguaje

Análisis de Sentimiento de las Reseñas de Amazon en Productos Electrónicos

Autor: Omar Mokrani Gallego

Madrid, Junio de 2021

Contenido

Introducción

Este trabajo consiste en la realización de un análisis de sentimientos de las opiniones que los clientes publican en distintos productos de Amazon.

Se ha utilizado la base de datos de productos de Datafiniti para la extracción de las valoraciones en formato *csv*. Datafiniti nos proporciona acceso instantáneo a una recopilación de datos de miles de páginas web para crear bases de datos estandarizadas con el principal objetivo de aportar, a la comunidad de desarrolladores, información valiosa de los consumidores que ayudarán a crear modelos de aprendizaje automático en un ámbito mayoritariamente comercial. Este *csv* cuenta con los siguientes datos: información descriptiva del producto, valoración, comentarios del cliente y fecha de publicación entre otros.

La finalidad de este trabajo consiste en interpretar estas opiniones, mediante un análisis de sentimientos desarrollado en R, para descubrir cuales son las palabras que se asocian a un sentimiento positivo del consumidor y cuales a uno negativo, lo que nos permitirá establecer un modelo que clasifique los productos, no solo en función de la puntuación de la valoración, sino que tenga en alta estima los comentarios escritos por los clientes.

Información del Conjunto de Datos

Los datos utilizados para este trabajo se pueden consultar en el archivo "AmazonReviewsDS.csv". Este archivo se compone de una lista de más de 28.000 reseñas en inglés, que los clientes han publicado para productos de Amazon como pueden ser *Kindle*, *Fire TV Stick* y muchos más. La fecha de estas reseñas se encuentra entre febrero y abril de 2019.

Con el ánimo de preparar los datos con una estructura correcta que nos permita realizar el análisis de sentimientos, se ha extraído del conjunto la información más relevante, es decir, se han extraído los datos asociados a los campos id, name, reviews.rating y reviews, y posteriormente se ha creado un dataset con cada una de las palabras que se han encontrado en los comentarios de cada producto para trabajar con más facilidad y analizar cada una de las palabras encontradas en el comentario.

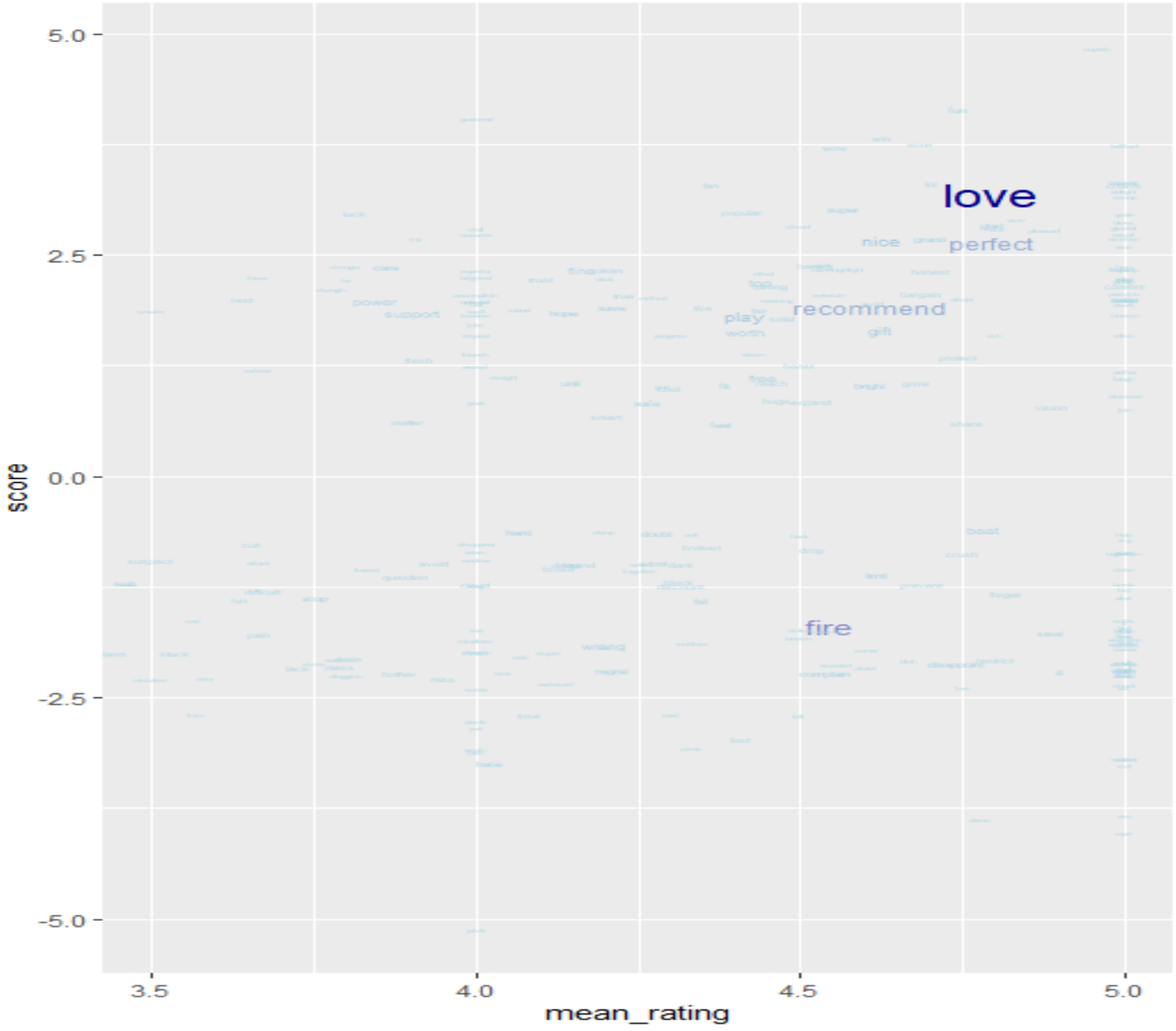
A modo introductorio, se presenta la siguiente nube de frecuencias con las palabras que más se repiten en el conjunto de datos:

Tratamiento del Conjunto de Datos y Desarrollo del Análisis Sentimental

Para el desarrollo de este trabajo se ha utilizado la lista Afinn de palabras en inglés, en el que se clasifican las palabras según el sentimiento en el intervalo [-5,5], siendo -5 el valor más negativo y 5 el más positivo. Una vez declarada esta tabla en el proyecto R, se unificarán la lista creada en el apartado anterior con las palabras encontradas en los comentarios de los productos con la lista Afinn. De esta manera podremos clasificar las palabras de nuestro conjunto de datos en función de la valoración estipulada por la lista Afinn.

Tras desarrollar esta funcionalidad, se pueden obtener distintos gráficos que nos servirán para hacernos una idea del sentimiento que produce en un usuario al leer las distintas palabras encontradas en las reviews.

Palabras más comunes en función de su puntuación



En este gráfico se visualizan con una letra más grande y de color más oscuro las palabras más repetidas y con letra más pequeña y de color mas claro las que con menos frecuencia aparecen.

Estas palabras se sitúan en el gráfico según su puntuación en el intervalo [-5,5] del modelo Afinn (eje y) y de la media de valoración de los productos en cuyos comentarios aparece la palabra a analizar (eje y). Dado a que la mayoría de los productos del conjunto de datos cuentan con una valoración de entre un 3.5 y un 5, con la intención de que el gráfico se visualizara más fácilmente, hemos limitado el eje x, descartando aquellas palabras con valoraciones inferiores a 3.5 en las reseñas de Amazon.

Tras analizar este gráfico, nos damos cuenta de que existe una gran cantidad de palabras que tienen una valoración muy negativa en comparación de la puntuación que han recibido en Amazon. Es el caso del término “fire” que dispone de una puntuación en la lista Afinn negativa, frente a la valoración media que han recibido los comentarios en los que esta incluida esta palabra, siendo de un valor aproximado de 4,5/5.

Palabras positivas más comunes

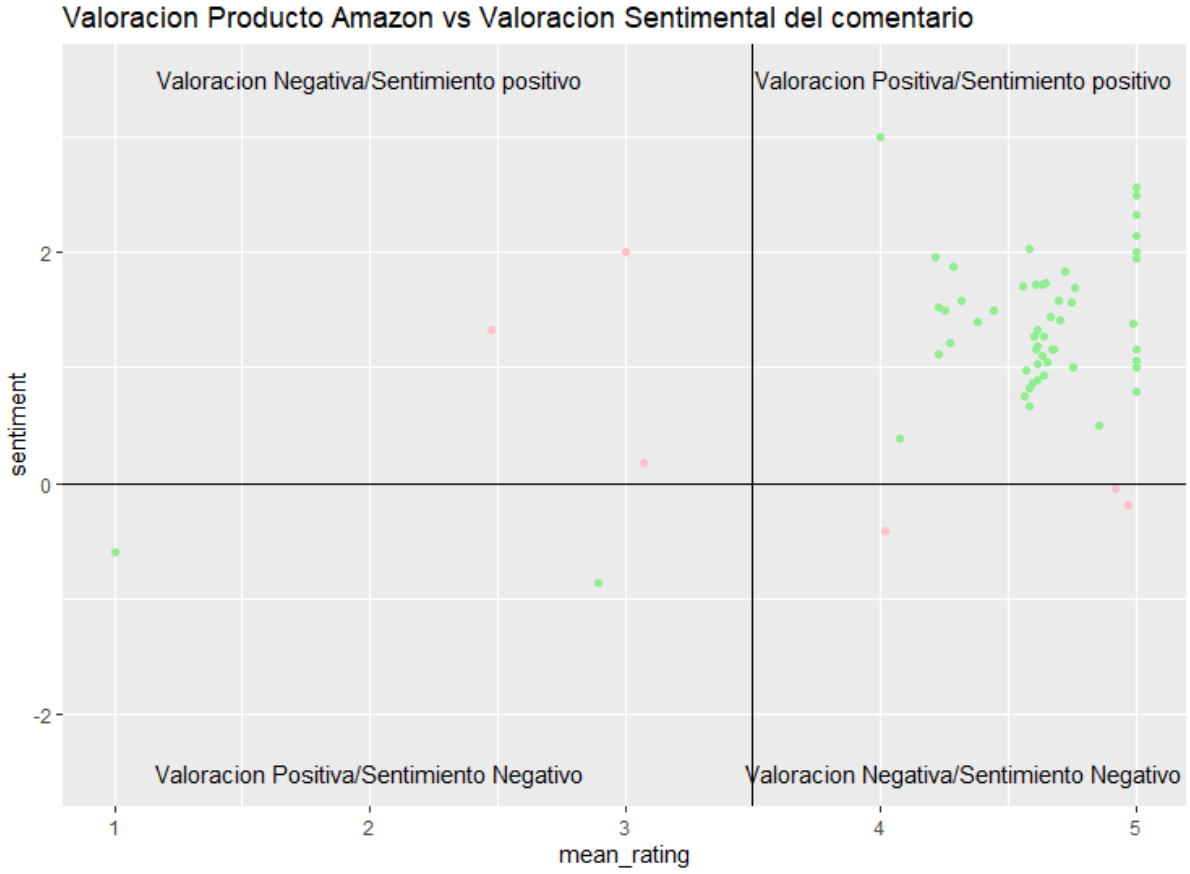
En el gráfico se aprecia la frecuencia de las palabras utilizadas en los comentarios de los productos con un carácter positivo. A mayor tamaño de fuente, mayor es la frecuencia y viceversa. La condición que se ha tenido en cuenta para la identificación de palabras buenas o de carácter positivo ha sido la de seleccionar la palabra cuando la valoración de la palabra se sitúe por encima de la valoración media de los productos en los que se encuentran.

Conclusiones

Tras analizar los resultados obtenidos, se debe investigar como el sentimiento de las palabras individuales afecta a la valoración general del producto. Para ello se ha desarrollado un script que cumple con lo siguiente:

- Agrupamos los datos por cada producto mediante el identificador de la lista Asinn.
- Establecemos la valoración media del producto.
- Establecemos el sentimiento medio de todas las palabras asociadas a las valoraciones de cada producto.

Relación entre la valoración de los productos y la media de los sentimientos



Mediante esta gráfica podemos concluir que los cuadrantes exitosos son aquellos en los que la reseña y el sentimiento coinciden, ya sea de manera positiva o negativa. En este caso existen muchos más puntos exitosos que desfavorables, pero no hay que olvidarse de las imprecisiones existentes en el mismo. Hay varias palabras que, según nuestro conjunto de datos, se interpretan con un sentimiento positivo pero los comentarios de los productos en los que están contenidas tienen un valor negativo y viceversa, existen valoraciones positivas de productos que contienen palabras catalogadas como sentimiento negativo.

A todo esto, podemos concluir en que existe una relación adecuada entre la valoración de los productos y el sentimiento de las palabras que contiene la reseña del producto de Amazon.

Algoritmos de Bayes Naive y Random Forest

En adición al análisis realizado previamente, se han utilizado los algoritmos de *Bayes Naive* y *Random Forest* para crear un modelo de clasificación que nos permita obtener las valoraciones de los comentarios en función del rating que disponga, así como las palabras contenidas en el mismo.

Para el entrenamiento del modelo se ha tomado un 70% de los datos para el entrenamiento y un 30% para el test.

La matriz de confusión obtenida es la siguiente:

	Reference				
Prediction	1	2	3	4	5
1	33	29	89	137	672
2	0	1	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0

1 Matriz de Confusión Bayes Naive

Con un valor de precisión de un 3.54%, hemos decidido descartar este modelo y probar a realizar el entrenamiento de un modelo, con las mismas proporciones para el entrenamiento y test, utilizando el algoritmo de *Random Forest*.

Los resultados en este caso han sido los siguientes:

	Reference				
Prediction	1	2	3	4	5
1	31	0	0	0	0
2	0	30	0	0	0
3	1	0	82	0	0
4	0	0	1	110	0
5	1	0	6	27	672

2Matriz de Confusión Random Forest

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.93939	1.00000	0.92135	0.8029	1.0000
Specificity	1.00000	1.00000	0.99885	0.9988	0.8824

Con una precisión del 96.25%, podemos afirmar que el algoritmo *Random Forest* resulta se el más adecuado a utilizar en este caso.

Gracias a esta prueba podemos concluir que la gran mayoría de las valoraciones obtenidas a partir de las palabras contenidas en los comentarios coinciden en mayor medida en las clases 2

y 5 disponiendo de una sensibilidad completa, mientras que en el caso de la clase 2 solo hemos conseguido calificar correctamente el 80.29%.

Como estimación, útil para comprobar cuanto mejora este modelo una predicción contra las probabilidades observadas, disponemos de un valor Kappa = 91.85%.

A modo indicador de la probabilidad que tiene un dato predicho, pertenezca a una categoría positiva, cabe destacar que se han obtenido valores predictivos positivos muy altos en comparación a los negativos. Se sitúan todos ellos por encima de 0.9518. Adicionalmente mencionamos que el valor obtenido para la precisión balanceada ha sido en todos los casos superior a un 94.12%.

Código

El código en leguaje R junto con los datos empleados para la realización de este trabajo se pueden encontrar en: https://github.com/OmarMokraniGallego/IntelligentSystems