Máxima Entropía en "Rotten Tomatoes Reviews"

Naoto Aguilar Alberto Rodríguez

Universidad de Santiago de Chile. https://www.usach.cl/

Abstract. Esta experiencia corresponde a la clasificación de texto a través de la máxima entropía. El dataset utilizado es perteneciente a la página "Rotten Tomatoes", la cual se encarga de entregar reviews de cine y televisión, a través de comentarios que definen si una película es buena con la clasificación de un tomate "Fresh" o si es mala con la clasificación de un tomate "Rotten". Se plantea utilizar la técnica de clasificación de máxima entropía en los comentarios perteneciente a las reviews para posteriormente entrenar un modelo y que este nos permita estimar si es que un comentario pertenece a una review "Fresh" o "Rotten". Se finaliza este informe con las conclusiones pertinentes, en base al modelo generado con su precisión, recall (Sensibilidad) y F1 respectivo.

Keywords: Máxima Entropía · R · Procesamiento de lenguaje natural

1 Introducción

Rotten Tomatoes es un sitio web estadounidense de reviews para cine y televisión. donde su objetivo es obtener y promediar las reseñas de distintos críticos connotados, para clasificar el valor positivo o negativo de una película o serie. Esto lo hace a través de dos tipos de categorías. El tomate rojo "Fresh" que significa el valor de un tomate fresco, es decir, al valor positivo de una review, al contrario, si el tomate es verde "Rotten" significa el valor de un tomate podrido, que corresponde a una valoración negativa de una review, además cada review contiene un comentario asociado. En esta experiencia se busca procesar el lenguaje natural de dichos comentarios para poder estimar y clasificar la valoración de dicha review, es decir, que en base al comentario de una review estimar si se define una película o serie como "Fresh" o "Rotten", esto con ayuda del método de **Máxima Entropía** utilizando una serie de paquetes entregados por R, que permiten el correcto procesamiento de los datos.

Esta experiencia se divide en los siguientes puntos: los datos utilizados, el método, resultados de la investigación, discusión sobre los resultados y conclusiones obtenida.

Los objetivos de la experiencia son:

- Comprender y presentar el problema de clasificación de texto (categorización)
- Realizar pre-procesamiento de texto para problemas de clasificación.
- Selección de muestra para entrenamiento y test

- Realizar proceso de calibración de parámetros del algoritmo de máxima entropía.
- Comprender de forma práctica el funcionamiento del método de máxima entropía mediante la configuración de sus parámetros.
- Evaluar el rendimiento del algoritmo mediante sus índices "precisión", "recall" y "F1", definiendo un subconjunto de categorías "relevantes" para dicho propósito.

2 Métodos y Datos de la investigación

2.1 Método

Se utiliza el Método de la Máxima Entropía para clasificar, el cual se entiende como un tipo de inferencia estadística encargada de modelas datos conocidos y no suponer datos desconocidos. De acuerdo al físico Ed. T. Jaynes, la máxima entropía se define como la estimación menos sesgada posible sobre la información proporcionada, la cual se construye a partir de la teoría de la información con el objetivo de establecer distribuciones de probabilidad sobre la base de conocimiento parcial [1]. Dicha distribución se selecciona a partir de las condiciones de satisfacción respecto a las restricciones impuesta por el modelo.

El principio de Máxima Entropía es útil solo cuando se aplica información comprobable y tiene diferentes aplicaciones. Para la actual investigación, el método se utiliza con el objetivo de clasificar el lenguaje natural a partir de la extracción de características de ciertas observaciones, las cuales se combinan linealmente, como se indica en la siguiente ecuación:

$$P(c|x) = \frac{1}{Z} exp(\sum w_i f_i) \tag{1}$$

Donde w_i corresponde al peso de la característica e indica la relevancia para clasificar. f_i y Z se definen como el factor de normalización para que las probabilidades sumen 1, siendo f_i la encargada de asociar una característica de un documento i con una clase, mientras Z se encarga de normalizar los valores entre 0 y 1.

2.2 Datos de la investigación

Este conjunto de datos es obtenido mediante la página oficial de Rotten Tomatoes y representan una visión crítica de las películas y series de televisión realizadas. El objetivo del conjunto de datos es clasificar aquellas películas que tengan una mejor calificación por parte de los críticos. Esta calificación es evaluada en según dos categorías presentes en la base de datos: Fresh y Rotten. En particular, todas las críticas realizadas estan en inglés y el conjunto de datos consta solo con 2 variables. Los atributos del dataset son:

- **Freshness:** Tipo de crítica realizada. Para una película o serie considerada buena por los críticos se dice que está Fresca (Fresh), mientras que una película o serie considerara mala está podrida (Rotten).

– **Review:** Texto de la crítica realizada a la película o serie, donde se dan las apreciaciones sobre cada obra.

De las 480.000 observaciones realizadas, hay un 50% reviews consideradas como Fresh y un 50% como Rotten (ver Figura 1). En cuanto a la ausencia de datos, no existe datos faltantes en ninguno de los 2 atributos.

Se realiza un pre-procesamiento a los datos, en el atributo de **Review**, ya que el otro atributo son las clases, estas son necesarias ya que permiten mejorar el modelo antes de entrenarlo y evaluarlo. A continuación se muestran todos los filtros realizados (se utilizó la función *Corpus* de R, que nos permite explorar el texto y filas en particular):

– **Caracteres:** Eliminación de comillas, puntos, apóstrofes, etc.
– **Palabras:** Eliminación de artículos y preposiciones.
– **Sufijos:** Eliminación de sufijos, terminaciones de palabras para dejar la raíz.
– **Minúsculas y números:** Se eliminan números y se deja todo en minúsculas.

3 Resultados de la investigación

Dado que el procesado para encontrar los parámetros de ajuste para el modelo de máxima entropía, requiere un costo bastante alto de tiempo, se aplica validación cruzada con el fin de considerar un volumen menor de datos, que para efectos del taller se realizaron pruebas entre un 10% a un 20% de la totalidad de los datos.
En primer lugar, para una configuración del $l1 - regularizer = 0.0$ y $l2 - regularizer = 1.0$ con un 80% de los datos para entrenar, se obtienen los siguientes valores de eficiencia en la predicción con los datos de prueba al 20% del total de la muestra. Para los siguientes casos se intercalan las proporciones de entrenamiento y pruebas (20% y 80% respectivamente) y se varía el tamaño del conjunto de datos, respetando las mismas configuraciones de umbrales.
Los umbrales presentados en las siguientes tablas, hacen referencia al valor de calibración para considerar aquellos documentos que son relevantes para el cálculos de la eficiencia, por tanto a menor valor, tendremos una mayor flexibilidad para considerar si un dato es considerado correcto en la predicción.

Table 1. Tabla de eficiencia para TR-data 80% y Test-data 20%. Dataset 60.000.

Umbral	Precisión	Recall	F1
0,51	0,503	0,986	0,666
0,53	0,504	0,957	0,660
0,55	0,505	0,930	0,655
0,60	0,508	0,860	0,638
0,65	0,507	0,778	0,614
0,70	0,507	0,699	0,588
0,75	0,509	0,621	0,560
0,80	0,512	0,534	0,523

Table 2. Tabla de eficiencia para TR-data 20% y Test-data 80%. Dataset 60.000

Umbral	Precisión	Recall	F1
0,51	0,512	0,983	0,673
0,53	0,512	0,949	0,665
0,55	0,514	0,916	0,658
0,60	0,516	0,834	0,638
0,65	0,518	0,750	0,613
0,70	0,520	0,663	0,583
0,75	0,523	0,574	0,547
0,80	0,525	0,479	0,501

Table 3. Tabla de eficiencia para TR-data 80% y Test-data 20%. Dataset 100.000.

Umbral	Precisión	Recall	F1
0,51	0,506	0,986	0,669
0,53	0,506	0,956	0,661
0,55	0,505	0,927	0,654
0,60	0,506	0,852	0,635
0,65	0,507	0,779	0,614
0,70	0,509	0,704	0,590
0,75	0,507	0,618	0,557
0,80	0,506	0,530	0,518

Table 4. Tabla de eficiencia para TR-data 20% y Test-data 80%. Dataset 100.000.

Umbral	Precisión	Recall	F1
0,51	0,508	0,984	0,670
0,53	0,507	0,953	0,662
0,55	0,508	0,923	0,656
0,60	0,509	0,844	0,635
0,65	0,510	0,765	0,612
0,70	0,511	0,682	0,584
0,75	0,514	0,597	0,552
0,80	0,516	0,505	0,510

Table 5. Tabla de eficiencia para TR-data 80% y Test-data 20%. Dataset 200.000.

Umbral	Precisión	Recall	F1
0,51	0,505	0,985	0,668
0,53	0,505	0,956	0,661
0,55	0,505	0,928	0,654
0,60	0,505	0,856	0,636
0,65	0,506	0,785	0,616
0,70	0,508	0,711	0,593
0,75	0,510	0,633	0,565
0,80	0,512	0,547	0,529

Table 6. Tabla de eficiencia para TR-data 20% y Test-data 80%. Dataset 200.000.

Umbral	Precisión	Recall	F1
0,51	0,509	0,984	0,671
0,53	0,510	0,955	0,665
0,55	0,510	0,926	0,658
0,60	0,510	0,851	0,639
0,65	0,512	0,776	0,618
0,70	0,515	0,697	0,592
0,75	0,515	0,614	0,561
0,80	0,517	0,525	0,521

Table 7. Tabla de eficiencia para TR-data 80% y Test-data 20%. Dataset 8.000.

Umbral	Precisión	Recall	F1
0,51	0,502	0,977	0,531
0,53	0,503	0,967	0,508
0,55	0,514	0,923	0,477
0,60	0,501	0,885	0,420
0,65	0,519	0,843	0,336
0,70	0,507	0,678	0,209
0,75	0,502	0,540	0,167
0,80	0,509	0,430	0,095

4 Discusión

De acuerdo a los rendimientos presentados en el apartado de resultados en las tablas 1, 2, 3, 4, 5 y 6, se logra evidenciar que al variar los umbrales en cuanto a la selección de documentos relevantes, tanto el *recall* como el $F1$, se incrementan conforme se aumenta el umbral, esto es algo lógico, dado que a mayor valor se restringe la cantidad de casos que se consideran como efectivamente válidos en la predicción del documento como una observación rotten o fresh.
Para el caso de la precisión, se logra notar que en general se mantiene entorno al $0,5$, dado que el conjunto de datos está exactamente balanceada y por tanto se refleja en el valor de este indicador.
Cabe señalar que al variar el tamaño del conjunto de entrenamiento, como el de pruebas y así como también las proporciones entre ellas, no presenta diferencias significativas entre cada una de las configuraciones. Al considerar el umbral $0,51$ correspondiente al valor más flexible dentro de las pruebas presentadas, vemos que el mejor rendimiento se obtiene para el conjunto de la tabla 2, obteniéndose un $F1$ de $0,673$, mientras que el peor valor se obtiene para el conjunto presentado en la tabla 1 con un $F1$ de $0,666$, con una diferencia del $\sim 3\%$.
Luego si se compara de acuerdo al tamaño del conjunto de datos, y para las mismas proporciones para los datos entrenamiento y prueba, notamos que la diferencia es de $\sim 1,5\%$ y $\sim 0,6\%$ entre las tablas 1-3 y 2-4 respectivamente en cuanto al $F1$. Por lo tanto, no se obtiene cambios significativos al incrementar el conjunto de datos ni al cambiar las proporciones del conjunto de entrenamiento y pruebas.
Al observar los resultados obtenidos para un valor mayor de datos, presentadas en las tablas 5 y 6, se logra observar que también se obtienen eficiencias similares a las tablas anteriormente analizadas. Por tanto, dado que el costo computacional de analizar mayor cantidad de datos, es que se detiene las pruebas respecto al incremento de datos dentro del modelo de máxima entropía, al no presentar mejoras significativas en los indicadores.
Finalmente, se destaca que para valores de conjunto de datos menores a diez mil, los valores del *recall* y $F1$, comienzan a decaer de forma importante, como se puede observar en la tabla 7, donde para un umbral de $0,51$, se obtienen diferencias entorno al $\sim 20\%$ del $F1$ respecto a las configuraciones con mayor volumen de datos.
Esto sugiere que el modelo no necesariamente podría mejorar con una mayor cantidad de datos o bien con diferentes proporciones de entrenamiento y pruebas, en este caso, se evidencia que es fuertemente importante el pre-procesamiento de los textos y en particular sobre las revisiones de los usuarios de acuerdo a una película.

5 Conclusiones

Realizando esta experiencia se ha comprendido el problema de clasificación de lenguaje natural, además se demostró el uso del método de la máxima entropía, la cual permitió crear un modelo de clasificación de texto, entrenándolo y evaluando con muestras entre 10% y 20% del dataset original. El modelo de clasificación se basó en los comentarios pertenecientes a cada review del dataset de Rotten Tomatoes, logrando estimar de manera correcta la clasificación positiva de "Fresh" y negativa "Rotten" para los datos de evaluación. Cabe destacar que el pre-procesamiento pertinente de los datos, logró considerablemente mejorar la evaluación del modelo.
En general la eficiencia del modelo no mejora significativamente a medida que se utiliza un mayor volumen de datos, tanto para el entrenamiento como para las pruebas, así como también en sus proporciones, obtiendose valores de $F1$ entorno a $\sim 0,67$ en todas las configuraciones de volumen de datos mayores a 60.000, sin embargo para conjuntos mas reducidos, se comienza a observar que la eficiencia del modelo decae de forma importante, entorno al 20% respecto al $F1$ para un umbral de $0,51$ y 8.000 datos con una proporción de 80% y 20% para el conjunto de entrenamiento y pruebas respectivamente.

References

1. Sotolongo-Costa, O., González, A., Brouers, F. (2009). Distribuciones estadísticas "generalizadas" a partir del principio de máxima entropía, 26(2B), 262–266. http://www.revistacubanadefisica.org/RCFextradata/OldFiles/2009/vol.26-No.2B/RCF-26-2B-2009-262.pdf

6 Anexos

A A Análisis Estadístico

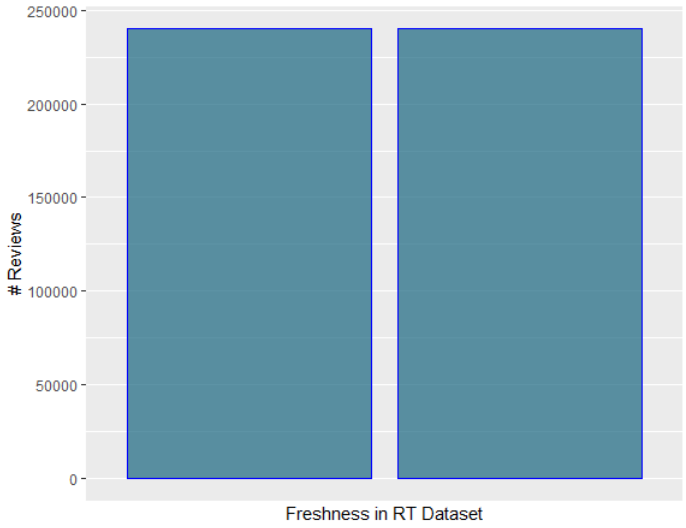

Fig. 1. Cantidades de cada clase en el dataset