Minería de texto aplicado a análisis sentimental para críticas de películas

Pedro Pablo Silva Antilef[1] [0000−1111−2222−3333]

Universidad de Santiago de Chile, Facultad de ingeniería, Departamento de ingeniería informática, Estación Central, Santiago de Chile pedro.silva@usach.cl

Abstract. En el presente trabajo se presenta un análisis sentimental de un *dataset* con críticas de la página dedicada a películas *Rotten Tomatoes*. El *dataset* consta de una columna con un texto que corresponde a la crítica en sí y otra columna con la clasificación dicotómica de este texto, el cual puede ser *rotten* (podrido) o *fresh* (fresco), es decir, un comentario negativo y positivo respectivamente. Para realizar este análisis fue necesario preprocesar la base da datos de críticas, extrayendo sólo la información relevante, con el fin de obtener un mejor resultado de reconocimiento. Finalmente se obtuvo un porcentaje de acierto de 75.14% en la clasificación.

Keywords: Sentiment Analysis · Max entropy · Data Minning.

1 Introducción

La orientación semántica se puede definir como una medida de subjetividad y opinión vertida en un texto. Para esto se busca determinar el factor o polaridad del texto (negativo o positivo) y también la potencia o fuerza (grado de que el texto es positivo o negativo). En general el análisis y automatización de la extracción de la orientación semántica de un palabra, frase, sentencia o documento se denomina análisis sentimental [1].

En [2] se realiza realiza un análisis sentimental de texto obtenidos diariamente desde la red social *Twitter*, los cuales son procesados para ser definidos posteriormente de dos formas diferentes; *OpinionFinder* para definir si un texto es positivo o negativo y otro denominado *Google-Profile of Mood States*, el cual mide el estado de ánimo en 6 categorías: calmado, alerta, seguro, vital, amable y feliz. Con estos datos, los investigadores fueron capaces de predecir el mercado de acciones (Índice bursátil Dow Jones) de Estados Unidos utilizando *Self-Organizing Fuzzy Neural Network*, obteniendo una precisión del 86.7%.

El análisis sentimental ha sido aplicado para obtener información en tiempo real de las apreciaciones de votantes acerca de candidatos a cargos públicos, permitiendo rastrear cómo los votantes se sienten sobre los diferentes hechos y acciones de los candidatos [3].

En el presente informe se preprocesará una base de datos de críticas de películas obtenidas de la página *Rotten Tomatoes*, con el fin de obtener sólo las palabras importantes. Luego, se genera una clasificación de los datos utilizando el modelo de máxima entropía en el *software* estadístico R.

2 Estado del arte

En [4], se utiliza un clasificador de aprendizaje profundo llamado Redes Neuronales Convulcionales, con esto se clasifican las críticas de la página *Rotten Tomatoes* en conjunto con otros *datasets* de texto para definir si el texto tiene una connotación positiva o negativa. Este clasificador es programado en C++ para dispositivos móviles, entregando resultados de la clasificación en tiempo real, obteniendo un máximo de precisión de 71.5%.

En [5] se realiza un análisis sentimental de un *dataset* con información recolectada de los sitios Box Office Mojo y Rotten Tomatoes, donde se incluyen todos las películas estrenadas entre 2004 y 2015 disponibles. Se encontró evidencia de que las opiniones de los consumidores influían en el tiempo en el que las películas se mantenían en el cine, pero que los comentaristas expertos tienen muchísimo peso en el mercado de las películas.

3 Datos

El *dataset* utilizado es un dataset público llamado *Rotten Tomatoes*, el cual consta de 480.000 registros con dos columnas, *Freshness* y *Review*, donde la primera puede tomar los valores *rotten* o *fresh*, es decir "podrido" y "fresco" respectivamente, ambas con 240.000 registros. La segunda columna *Review*, contiene un texto largo con la crítica escrita en la página *Rotten Tomatoes*.

4 Metodología

El clasificador de máxima entropía es un algoritmo de aprendizaje de maquina que pertenece a la clase de modelos exponenciales. A diferencia de otros clasificadores como *Naive Bayes*, el clasificador de máxima entropía no asume que las características son condicionalmente independientes las unas de las otras. Se basa en el principio de la máxima entropía y de todos los modelos que ajustan los datos de entrenamiento, selecciona los que tengan una mayor entropía [6].

En primer lugar, se eliminan 8.916 registros que no pueden ser codificados de manera correcta, principalmente debido a algunos caracteres especiales de algunos idiomas como el portugués y frances. Luego, se toma una muestra de los datos de 40.000 registros, ya que al utilizar los 480.000 registros originales del *dataset*, el algoritmo toma mucho tiempo en ser ejecutado. Para el preprocesamiento de los datos, se realizan varias transformaciones al campo *"Review"* para obtener sólo la información que es de interés, eliminando aquellas palabras o partes de palabras que puedan generar ruido en el modelo. A continuación se enumeran los procesamientos que se llevan a cabo sobre los datos de una manera cronológica:

- 1. Se remueven las puntuaciones.
- 2. Se remueven las palabras ".

- 3. Se remueven las *stopwords* (conjunto de palabras de cada idioma que se descartan del cuerpo del texto con el fin de disminuir la cantidad de palabras que no agregan información).
- 4. Se transforman a minúscula
- 5. Se remueven las *stopwords*.
- 6. Se mantiene solo la raíz de las palabras (la parte de las palabras que agrega valor y que puede ser compartida por otras palabras de similar significado).
- 7. Separación con espacios.
- 8. Se remueven números.
- 9. Se transforma a formato *MatrixDocs*, legible para la librería *MaxEnt*.

Se realizaron 2 experimentos para el análisis sentimental utilizando la función *maxent* de la librería con el mismo nombre. Para definir los parámetros del modelo, se utiliza la función *tune*, el cual aplica el modelo para cada una de las combinaciones de parámetros, entregando la precisión de cada uno de esos experimentos, pudiendo seleccionar la mejor configuración. Los experimentos a realizar son los siguientes:

- Experimento 1: Con 40.000 datos.
- Experimento 2: Con 40.000 datos, pero eliminando las palabras "film" y "movi"

5 Resultados

En esta sección se presentan los resultados obtenidos de los diferentes experimentos planteados y explicados en la sección 4. Estos resultados serán discutidos en la siguiente sección.

A continuación se muestra la nube de palabras obtenida luego de preprocesar el *dataset* con 40.000 registros, y fue lo que gatillo principalmente la definición del segundo experimento.

Fig. 1: Nube de palabras para experimento 1

Por otro lado, podemos ver los mejores resultados obtenidos por la función *tune*, para los dos experimentos, así como también sus matrices de confusión, desde donde se derivan las métricas de calidad *Precisión*, *Kappa*, *Recall* y *F1*.

Experimento	**Parámetros**		
	l1_regularizer	**l2_regularizer**	**use_sgd**
1	0.0	1.0	0
2	0.0	1.0	0

Table 1: Tabla con parámetros de máxima entropía para ambos experimentos

	Referencia	
Predicción	**fresh**	**rotten**
fresh	3800	1275
rotten	1211	3714

Table 2: Matriz de confusión Exp. 1

	Referencia	
Predicción	**fresh**	**rotten**
fresh	3896	1300
rotten	1242	3662

Table 3: Matriz de confusión Exp. 2

	Métrica			
Experimento	**Accuracy**	**Kappa**	**Recall**	**F1**
1	0.75140	0.5028	0.758	0.755
2	0.738	0.5115	0.7422	0.7399

Table 4: Métricas de calidad para experimento 1 y 2

6 Discusión y conclusiones

Ambos experimentos tuvieron muy buenos resultados, esperados dentro de este tipo de análisis y clasificación, congruente sobre todo con el estado del arte, específicamente con [4] donde se utiliza redes neuronales convulcionales, donde cual obtiene una precisión levemente menor a la obtenida con máxima entropía en el presente trabajo.

Si bien las métricas estudiadas muestran valores razonables, no se pueden observar diferencias notorias en cuanto a ambos experimentos, probablemente debido a que la extracción de dataset de palabras como "movi" o "film", no significaron un gran cambio en el modelo, es decir que dichas palabras no eran significativas para la clasificación de "fresco" o "podrido", por lo que incluso el resultado del segundo experimento fue levemente menor. Solo el índice Kappa tuvo un mejor rendimiento, pero se sigue manteniendo en el rango "débil" de esta métrica.

Si bien el algoritmo funcionó de manera correcta, tuvo problemas para manejar una gran cantidad de datos, al inicio de la actividad, se intentó correr el algoritmo con los 480.000 registros del *dataset*, pero el tiempo de ejecución fue demasiado alto, por lo que se decidió trabajar con una muestra de los datos, de manera de poder abordar el problema en tiempos razonables. Es posible que otras técnicas, librerías o incluso lenguajes de programación sean más aptos para el procesamiento de grandes cantidades de información, sobre todo para textos largos como lo son las críticas de películas.

También se recomienda probar con otros tipos de clasificación para comparar resultados, aun que según lo visto en la literatura, máxima entropía en conjunto con redes neuronales son los modelos más eficientes para este tipo de análisis.

References

1. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.
2. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.
3. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.
4. Sankar, H., Subramaniyaswamy, V., Vijayakumar, V., Arun Kumar, S., Logesh, R., & Umamakeswari, A. J. S. P. (2020). Intelligent sentiment analysis approach using edge computing-based deep learning technique. Software: Practice and Experience, 50(5), 645-657.

5. Souza, T. L., Nishijima, M., & Fava, A. C. (2019). Do consumer and expert reviews affect the length of time a film is kept on screens in the USA?. Journal of Cultural Economics, 43(1), 145-171.
6. "https://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/". Rescatado el 10 de Mayo de 2020.