Taller 3: Clasificación de reseñas de películas usando Scaling Linear Discriminant Analisys

Franco Cornejo Gonzalez (4 horas)
Departamento de Ingeniería Informática
Universidad de Santiago de Chile, Santiago, Chile
franco.cornejo.g@usach.cl

***Resumen*—En este trabajo se aborda la clasificación automática de reseñas de películas provenientes de Rotten Tomatoes en categorías 'fresh' (positivas) y 'rotten' (negativas), utilizando los algoritmos Scaling Linear Discriminant Analysis (SLDA) y Random Forest (RF). A través de un preprocesamiento detallado del texto, se generaron modelos que alcanzaron cerca del 70 % de efectividad en términos de precisión, recall y F1-score. También se exploró el uso de kernels de palabras claves para mejorar el rendimiento, logrando una mayor parsimonia sin pérdidas significativas en la clasificación.**

***Palabras claves*—Clasificación de texto, Random Forest, Scaling Linear Discriminant Analysis, Minería de texto, Reseñas de películas, Rotten Tomatoes.**

1. Introducción

La masificación de internet y el auge de las redes sociales han permitido la acumulación de una vasta cantidad de comentarios sobre diversos temas. A primera vista, esta información puede no parecer de utilidad, pero en verdad esconde un gran valor. Con un análisis adecuado, se puede extraer conocimiento que permita a las empresas o los estados tomar decisiones más informadas.

En el ámbito del análisis de sentimientos, la minería de texto ha demostrado ser una herramienta poderosa para comprender las opiniones de los usuarios sobre productos, servicios o personajes mediáticos. Un ejemplo particularmente relevante son las reseñas de películas, que no solo expresan la satisfacción o insatisfacción de los espectadores, sino que también pueden influir en el éxito comercial de una película.

Rotten Tomatoes es una popular página dedicada a la crítica de películas, en la cual se puede calificar una película usando el sistema de "Tomatometro", es decir, una escala que permite puntuar la calidad de la película. Ademas esta puntuación puede ir acompañada de una breve reseña donde puede expresarse la razón de la puntuación. A diferencia de otras plataformas, Rotten Tomatoes contempla tanto críticas de expertos, como de aficionados al cine en general.

Este trabajo se centra en la clasificación automática de reseñas de películas utilizando técnicas de aprendizaje supervisado. Para ello, se analizará un corpus de reseñas de Rotten Tomatoes, clasificándolas en las categorías 'fresh' (positivas) y 'rotten' (negativas). Se emplearán dos enfoques principales: Scaling Linear Discriminant Analysis (SLDA) y Random Forest (RF).

Además de facilitar el procesamiento de grandes volúmenes de datos, el análisis automatizado de reseñas puede proporcionar información valiosa para la industria cinematográfica. Al entender mejor las preferencias del público, las productoras podrían adaptar sus estrategias de producción y promoción, identificando qué tipo de contenido atrae más a la audiencia y qué aspectos podrían generar desinterés. Este conocimiento podría contribuir a la creación de películas más alineadas con las expectativas del público, optimizando las decisiones de inversión y marketing.

2. Metodo y datos

2.1. Descripción del corpus

Para el presente estudio se uso como copus *Rotten Tomatoes Reviews Text Dataset* recuperado de Kaggle. Este es un archivo .csv cuenta con 480000 reseñas de películas y 2 columnas una llamada "Freshness" y otra llamada "Review". Como su nombre indica Freshness es la categoria y tiene dos posibles valores "Fresh" para una critica favorable y "Rotten" para una critica negativa. Review es un comentario sobre la película echo por un crítico especializado o por un entusiasta del cine.

Cantidad de reseñas por clase:

- 240000 Fresh
- 240000 Rotten

Notar que los datos se encuentran perfectamente balanceados lo que facilita el preprocesamiento de los mismos.

2.2. Análisis exploratorio

Al ser los datos texto perfectamente legible la primera exploración fue una lectura rápida de varias reseñas. Con esta lectura pudimos notar los siguientes puntos:

- Uso de lenguaje informal.
- Un vocabulario extenso.
- Gran variedad de adjetivos para expresar una misma idea.
- Las palabras usadas en
- Las criticas negativas suelen tener mas vocabulario y las palabras usados son mas "fuertes".
- Los adjetivos usados en las criticas positivas y negativas son muy diferentes.

A nuestro entender este ultimo punto es la clave para que el computador logre distinguir entre reseñas positivas y

negativas. Para extraer este conjunto de palabras se procedió de la siguiente forma:

1. Se eliminan las StopWords, se transformar todo en minúsculas y se reducen las palabras a su raíz.
2. Se separan las reseñas en dos conjuntos, uno con las reseñas positivas y otro con las negativas.
3. Se reducen los conjuntos a sus 1000 palabras mas frecuentes.
4. Se calcula la interacciona entre los conjuntos y luego se resta a cada uno.

El resultado de este proceso es un conjunto con las palabras más frecuentes que solo se encuentran en las reseñas positivas y otro conjunto análogo pero con las negativas (notar que la suma de estos es muy similar a la idea de diferencia simétrica de conjuntos), a estos conjuntos los llamamos kernels. El kernel de las reseñas frescas contiene 65 palabras y el de las podridas 66.

- Kernel Fresh: good, captures, quiet, gripping, intimate, superb, engrossing, delightful, poignant, thoughtful, wonderfully, affecting, riveting, haunting, friendship, quietly, vivid, tense, absorbing, refreshing, allows, strength, achievement, thought-provoking, exhilarating, finest, mature, gentle, incredible, triumph, captivating, try, brilliantly, bleak, reminder, timely, universal, visceral, chilling, heartbreaking, funniest, gem, inspiring, unsettling, unusual, sensitive, understated, astonishing, nuanced, melancholy, outstanding, suspenseful, doc, carries, exquisite, tender, blast, challenging, meditation, nonetheless, poetic, vital, brave, colorful y exploration.
- Kernel Rotten: unfortunately, flat, tedious, bland, sadly, unfunny, lazy, mediocre, tired, waste, pointless, uninspired, lame, bore, repetitive, poorly, sandler, badly, hollow, pretentious, annoying, offensive, superficial, bloated, clumsy, unnecessary, sitcom, sloppy, wasted, convoluted, incoherent, tiresome, excuse, unless, stale, fail, misfire, underwhelming, clunky, confused, confusing, crude, muddled, parody, inert, trouble, desperately, soap, unpleasant, sitting, disjointed, powerful, lifeless, depressing, misguided, stereotypes, misses, simplistic, apparently, appear, terribly, dreary y horrible.

2.3. Preprocesamiento de datos

Como se contaba con una gran cantidad de reseñas y trabajar con todas no es factible con la capacidad computacional de la que se dispone. Se redujo el conjunto a 24000 reseñas.

- Remover números.
- Remover StopWords, que son palabras comunes como conectores. Algunos ejemplos son the, and y of.
- Transformar todas las letras en minúsculas.
- Eliminar palabras que aparecen en menos del 0.3 de los documentos.

Con este procesado se logro reducir la cantidad de palabras a 989.

2.4. Métodos usado

2.4.1. Arboles de decisión: Los árboles de decisión son métodos de aprendizaje automático que realizan clasificaciones de datos a través de la selección recursiva de variables. Su estructura es fácil de entender y permite derivar reglas, lo facilita su interpretación. Sin embargo, tienden a sobreajustarse a los datos de entrenamiento y al centrarse en una sola variable a la vez, pueden tener una capacidad limitada para clasificar y separar en comparación con enfoques más sofisticados, como las redes neuronales o las máquinas de vectores de soporte (SVM). Por estas razones, a menudo se les considera clasificadores menos robustos (Chacón and R., 2017).

2.4.2. Scaling Linear Discriminant Analysis: El Scaling Linear Discriminant Analysis (SLDA) es una variante del Análisis Discriminante Lineal (LDA) que se utiliza principalmente en problemas de clasificación supervisada. En términos generales, el LDA busca encontrar las combinaciones lineales de las características de entrada que mejor separen dos o más clases. Sin embargo, el SLDA introduce un paso de escalado (scaling) para mejorar el rendimiento del modelo en escenarios donde las variables pueden tener distintas escalas o distribuciones. Al ser un método lineal no reajusta (Chacón and R., 2024).

3. Resultados

3.1. Aplicación de SLDA y RF

Los modelos se entrenaron usando 989 palabras y 24000 documentos. La cantidad de documentos necesarios se calculo aumentando el numero gradualmente hasta notar que los modelos dejaron de mejorar. Ademas el tiempo de entrenamiento fue de aproximadamente 30 minutos. Ademas se evaluó contra otros 24000 documentos.

Freshness	Precisión	Recall	F1-Score
Fresh	0.69	0.66	0.69
Rotten	0.67	0.70	0.71

Tabla 1: Métricas de rendimiento para las reseñas usando SDLA.

Freshness	Precisión	Recall	F1-Score
Fresh	0.69	0.72	0.70
Rotten	0.71	0.68	0.69

Tabla 2: Métricas de rendimiento para las reseñas usando RF.

El rendimiento de los dos modelos parece bastante equilibrado. Ambos modelos están cercanos al 70 % en todas las métricas. Notando con esto un equilibrio razonable entre precisión y la captura de verdaderos positivos.

3.2. Aplicación de SLDA y RF con ayuda de Kernel

A modo experimento se modifico la matriz de termino de la siguiente forma:

1. Se mantuvo el mismo preprocesamiento anterior salvo que se elimino el parámetro *removeSparseTerms=.997*.
2. Se identificaron las 500 palabras mas frecuentes.

3. Se filtraron todas las palabras que no estuvieran entre las 500 mas frecuentes o no pertenecieran a un Kernel de palabras.

De esta forma se redujo la cantidad de palabras de la matriz a 603. También se tuvo cuidado de no dejar documentos vacíos. Finalmente se usaron 9400 reseñas y 624 palabras para entrenar los modelos y este proceso tardo 5 minutos. Ademas evaluó contra 24000 documentos.

Freshness	**Precisión**	**Recall**	**F1-Score**
Fresh	0.68	0.67	0.67
Rotten	0.67	0.68	0.67

Tabla 3: Métricas de rendimiento para las reseñas usando SDLA y Kernel.

Freshness	**Precisión**	**Recall**	**F1-Score**
Fresh	0.68	0.68	0.68
Rotten	0.68	0.67	0.67

Tabla 4: Métricas de rendimiento para las reseñas usando RF y Kernel.

Los resultados son bastante parecidos a los obtenidos sin usar los kernels. Teniendo una perdida una pequeña perdida cercana al 3 % del F1-score en ambos modelos. A pesar de esto se considera que se gano en parsimonia ya que se redujo la cantidad de variables. Probablemente si se hubiera echo un trabajo volviendo a ponderar los valores de la matriz los resultados serian mejores.

4. Conclusiones

En este trabajo se logro clasificar las reseñas de películas con casi un 70 % de efectividad. Al comparar los modelos de SDLA con los de RF no se aprecia una ventaja significativa de uno por sobre el otro. También se exploro la idea de identificar palabras claves y usarlas a nuestro favor, sin embargo, con la implementación efectuada no se logro ver una gran ventaja comparativa contra el preprocesamiento común, obteniendo una ventaja en parsimonia pero no en rendimiento. Probablemente si esta idea se explora en más profundidad también se podría obtener una mejora en el rendimiento.

Referencias

[1] M. Chacón and F.-A. B. R., "Taller de minería de datos avanzada: Capítulo iii bosques aleatorios," in *Taller de minería de datos avanzada PPT*, 2017.

[2] ——, "Capítulo iii: Scaling linear discriminant analysis y minería de texto," in *Taller de minería de datos avanzada PPT*, 2024.