Clasificación sentimental de reseñas de libros en Goodreads como positivas o negativas usando técnicas de aprendizaje de máquinas

- Andrea Salcedo
- Reinaldo Verdugo



















Clasificación sentimental de reseñas de libros en Goodreads como positivas o negativas usando técnicas de aprendizaje de máquinas

- Andrea Salcedo
- Reinaldo Verdugo

Contenido

1. Planteamiento del Problema
2. Marco Teórico
3. Diseño de la Solución
    - 3.1 Descripción del Clasificador
    - 3.2 Algoritmos
4. Implementación
    - 4.1 Lenguajes y Plataforma
    - 4.2 Obtención y Preprocesamiento de Datos
    - 4.3 Implementación de los Algoritmos
5. Resultados
6. Conclusiones

1. Planteamiento del Problema

En este trabajo buscamos clasificar las reseñas de libros realizadas por usuarios de la comunidad de *Goodreads*.

Para dicha clasificación se tomarán en cuenta los sentimientos plasmados en la reseña, a fin de conocer si dicha crítica es positiva por contener sentimientos positivos (que hacen alusión a que el lector disfrutó de la lectura), o si por el contrario es negativa por contener sentimientos negativos (y por lo tanto mostrando descontento por parte del lector).

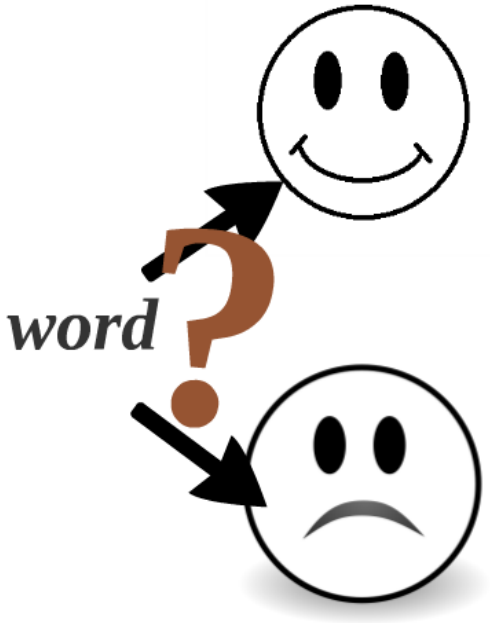
2. Marco Teórico

Goodreads es el sitio para lectores y recomendaciones de libros más grande del mundo.

En él se pueden calificar libros mediante **reseñas** que pueden ser positivas o negativas (dependiendo de lo que el crítico analice) y que pueden estar acompañadas de una calificación.

2. Marco Teórico

El **Análisis de Sentimientos** se refiere al proceso por el que determinamos si una frase o acto de habla contiene una opinión, *positiva* o *negativa*, sobre una entidad concreta o sobre un concepto.



En nuestro caso:

Pretty good! I'm in love with the series! So overwhelming

Just no. Absolutely not. I could NOT continue this book. This was a waste of time.

2. Marco Teórico

Stemming es un método usado para reducir una palabra a su stem, o lema.

consignment, consigned -> consign

liking, like, liked -> lik

Stopwords es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. que son filtradas antes o después del procesamiento de datos en lenguaje natural

a
as
about
and
any
because
the
this
with

3. Diseño de la Solución

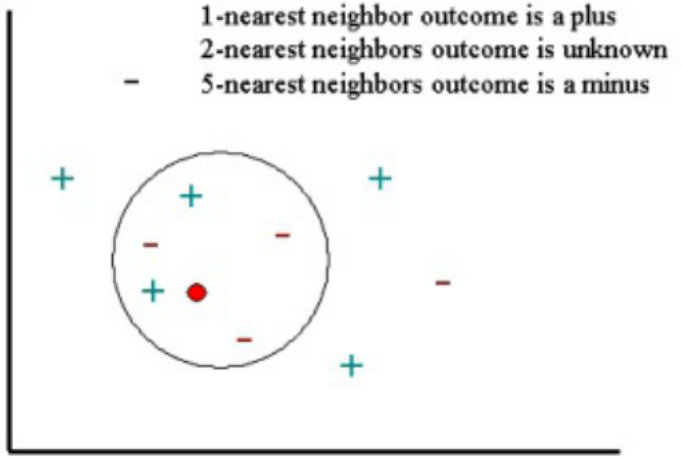
- **Tarea:** Clasificar reseñas de libros en Goodreads como positivas o negativas.
- **Métrica de Performance**: Comparar la clasificación con la calificación o el número de estrellas otorgadas de la reseña.
- **Experiencia**: Reseñas obtenidas del API de Goodreads, etiquetadas como positivas, negativas o neutras utilizando el número de estrellas otorgadas.

3. Diseño de la Solución

3.1 Algoritmos

- **K-Nearest Neighbor (k-NN)**

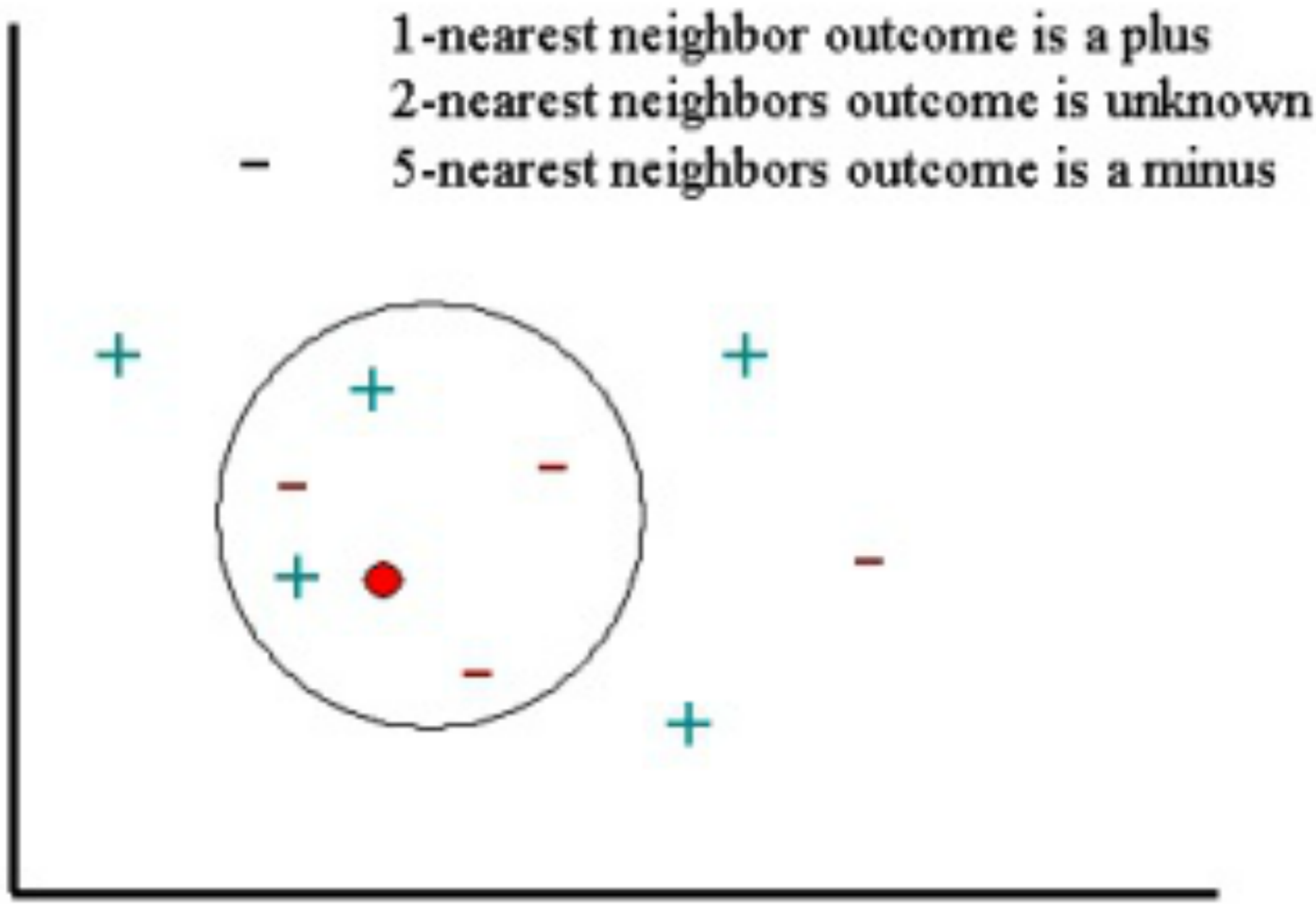
Aprendizaje supervisado basado en la clasificación de objetos, realizando un entrenamiento mediante objetos cercanos en el espacio de los elementos.



En nuestra aplicación, los vecinos cercanos corresponden a los textos con la mayor cantidad de palabras similares.

*It's rather **like** a lifetime special--**pleasant**, sweet, and forgettable*

*I **liked** this book, made for a **pleasant** evening*

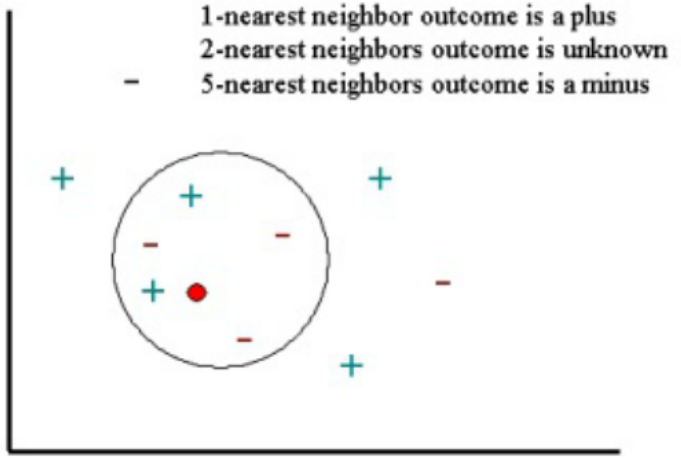

1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus

3. Diseño de la Solución

3.1 Algoritmos

- **K-Nearest Neighbor (k-NN)**

Aprendizaje supervisado basado en la clasificación de objetos, realizando un entrenamiento mediante objetos cercanos en el espacio de los elementos.



En nuestra aplicación, los vecinos cercanos corresponden a los textos con la mayor cantidad de palabras similares.

*It's rather **like** a lifetime special--**pleasant**, sweet, and forgettable*

*I **liked** this book, made for a **pleasant** evening*

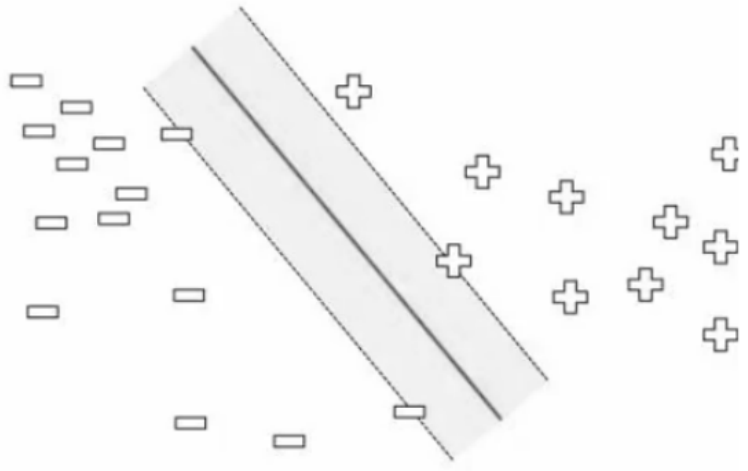
3. Diseño de la Solución

3.1 Algoritmos

- **Support Vector Machine (SVM)**

  Se centra en dividir los datos (reseñas) en 2 clases: la clase de reseñas positivas, y la clase de negativas. Entre ambas clases se busca un **vector frontera** (o de pesos).

  El **vector frontera** consiste en la suma de las reseñas positivas por un peso particular menos la suma de las reseñas negativas, también multiplicadas por un peso particular.



3. Diseño de la Solución

3.1 Algoritmos

- **Maximun Entropy Classifier (Maxent)**

  Solución particular del problema de clasificación que asume que una combinación lineal de los rasgos observados y algunos parámetros específicos del problema pueden usarse para determinar la probabilidad de cada posible valor de la variable dependiente

  Dado un evento x se tiene:

  - La probabilidad de que el evento ocurra. ***px***
  - La "sorpresa" de que ocurra el evento, definida como: ***log(1/px)***

3. Diseño de la Solución

3.1 Algoritmos

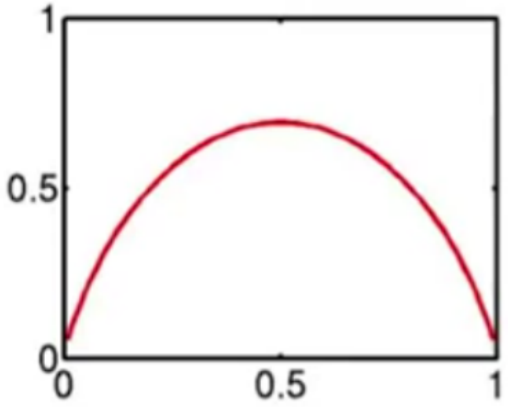
- **Maximun Entropy Classifier (Maxent)**

Dado un evento x se tiene:

- La **entropía**, que se define como el valor esperado de la sorpresa con respecto a p.

$$\mathrm{H}(p) = E_p\left[\log_2 \frac{1}{p_x}\right] = -\sum_x p_x \log_2 p_x$$

Ejemplo: lanzar una moneda



En el caso de este proyecto, las **restricciones** serían que la sorpresa esperada de cada palabra sea lo más cercana a la ocurrencia verdadera de la misma

4. Implementación

4.1 Lenguajes y Plataformas de Desarrollo

- **Python**

Para obtener, preprocesar y almacenar los datos necesarios para el proyecto.

- **R**

Se realizó la implementación de los tres algoritmos.

4. Implementación

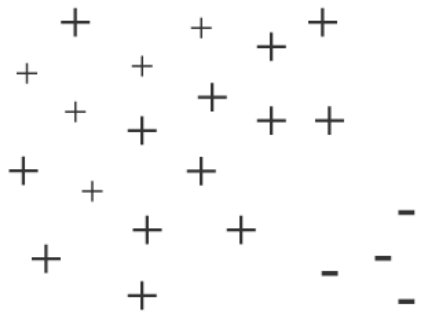
4.2 Obtención de Datos

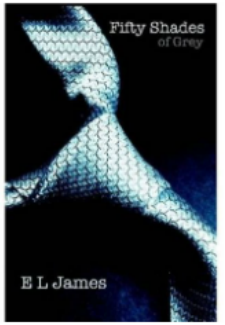
API

recent reviews

review

- cada 10 min





4. Implementación

4.2 Preprocesamiento de Datos

"Watch review "
"Real good lov

4. Implementación

4.3 Implementación de los Algoritmos

Algoritmo	Librería	Tiempo
KNN	class	13m22s
SVM	e1071	1m53s
Maxent	maxent	1m58s

5. Experimentos y Resultados

Un total de 1782 reseñas.

- 1154 positivas
- 350 negativas
- 278 neutras

El entrenamiento se realizó con el 70% de los datos y la clasificación con el 30% restante.

Se realizaron 30 corridas por algoritmo.

5. Experimentos y Resultados

Medida	KNN(%)	SVM (%)	MAXENT (%)
Exactitud	67.75%	77.65%	82.60%
Precisión	84.98%	77.45%	89.50%
Sensibilidad	70.39%	99.97%	88.04%
Especificidad	59.02%	4.23%	63.29%
Error	32.25%	22.35%	17.40%

Reviews Neutros	
Positivo	Negativo
257	21

5. Experimentos y Resultados

KNN

Predicción/Actual	Positivo	Negativo
Positivo	308	67
Negativo	38	38

Medida	Porcentaje
Exactitud	76.72%
Precisión	82.13%
Sensibilidad	89.02%
Especificidad	36.19%
Error	23.28%

SVM

Predicción/Actual	Positivo	Negativo
Positivo	329	93
Negativo	0	7

Medida	Porcentaje
Exactitud	78.32%
Precisión	77.96%
Sensibilidad	100%
Especificidad	7%
Error	21.68%

Maxent

Predicción/Actual	Positivo	Negativo
Positivo	311	31
Negativo	35	74

Medida	Porcentaje
Exactitud	85.37%
Precisión	90.94%
Sensibilidad	89.88%
Especificidad	70.48%
Error	14.63%

5. Experimentos y Resultados

5.2 Limitaciones

- Límite de solicitudes al API de Goodreads.
- Desequilibrio en el conjunto de datos.
- Uso de sarcasmo, referencias culturales, modismos y emoticonos.

:) :D :(xD

This book is SOOOOO **amazing**.
LUV IT!!!!

LOL

Heartbreaking.

6. Conclusiones

- Maximum Entropy fue el mejor clasificador con un accuracy de 82.60%.
- Experimentar con la cantidad de reseñas positivas y negativas utilizadas.
- Considerar la inclusión de la clase neutra durante el entrenamiento.

¡Gracias!