Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos II (Gpo 101)

Tarea 3

Clasificación de sentimientos de películas -Baseline

Profesor: Dr. Alfredo Esquivel Jaramillo

Erika Martínez Meneses

A01028621

8 de Julio del 2021

Features basados en Bag-of-Words (BOW)

Accuracy				
Models/Parameters	Logistic Regression	Support Vector Machine	Random Forest	Gradient Boosting
Nsamp = 1000, maxtokens = 50, maxtokenlen = 20	66%	65.16%	67.16%	63.83%
Nsamp = 1000, maxtokens = 100, maxtokenlen = 100	75.5%	75.33%	75.5%	69%
Nsamp = 1000, maxtokens = 200, maxtokenlen = 100	80%	79.16%	80%	77.5%

Features basados en TD-IDF (TFIDVectorizer() de scikit-learn)

Accuracy				
Models/Parameters	Logistic Regression	Support Vector Machine	Random Forest	Gradient Boosting
Nsamp = 1000, maxtokens = 50, maxtokenlen = 20	69.5%	70.33%	66.83%	66.83%
Nsamp = 1000, maxtokens = 100, maxtokenlen = 100	72.16%	72.66%	68%	68.66%
Nsamp = 1000, maxtokens = 200, maxtokenlen = 100	84%	85.66%	78.16%	76.83%

Conclusión

- Hay una mayor precisión con TF-IDF: los resultados muestran que el enfoque basado en TF-IDF generalmente ofrece mejores niveles de precisión que el enfoque basado en Bag-of-Words en todos los modelos con diferentes parámetros. Por ejemplo, la precisión con Regresión Logística mejora de 80% (BOW) a 84% (TF-IDF). Sin embargo, notamos que para la combinación con maxtokens = 100 BOW tiene mejores resultados.
- El mejor modelo basado en TF-IDF fue SVM con Nsamp = 1000, maxtokens = 200, maxtokenlen = 100 obteniendo 85.66% y para el mejor modelo

basado en BOW fue un empate entre LR y RF con 1000, maxtokens = 200, maxtokenlen = 100 obteniendo 80% de accuracy.

- El peor modelo para BOW fue GBM con Nsamp = 1000, maxtokens = 50, maxtokenlen = 20 obteniendo 63.83% y en el caso de TD-IDF tanto RF como GBM con Nsamp = 1000, maxtokens = 50, maxtokenlen = 20 obtuvieron 66.83% de accuracy siendo los peores modelos.

En general, los modelos entrenados con TF-IDF son más efectivos, lo que sugiere que este método captura mejor la importancia de las palabras en las tareas de clasificación de texto para este conjunto de datos.