Procesamiento de Lenguaje Natural
Tarea 3

Rayo Mosquera, Jhon Stewar
j.rayom@uniandes.edu.co

De La Rosa Peredo, Carlos Raul
c.delarosap@uniandes.edu.co

Mario Garrido Córdoba
m.garrido10@uniandes.edu.co

Octubre 2024

1 Naive Bayes - Regresión Logística

El objetivo es construir dos clasificadores de texto para identificar 20 categorías de noticias: uno usando Naive Bayes y otro regresión logística. Para ambos casos, se vectorizó el texto utilizando conteo de términos (*tf*) y *tf-idf*.

Se preprocesó el texto convirtiéndolo a minúsculas y eliminando espacios en blanco, entre otras técnicas de normalización. Se utilizaron dos técnicas de vectorización principales: *CountVectorizer* (tf) y *TfidfVectorizer* (tf-idf). *CountVectorizer* convierte el texto en una matriz de frecuencia de términos, lo cual es útil para capturar la ocurrencia bruta de palabras y obtener una visión general de la frecuencia de ciertos términos. Por otro lado, *TfidfVectorizer* ajusta estas frecuencias según la importancia de cada palabra en el corpus completo, reduciendo así la influencia de palabras muy comunes y destacando aquellas con mayor relevancia discriminatoria. Esto permite una mejor representación del contenido de cada documento.

Para evitar *overfitting* y encontrar el mejor modelo, se utilizó validación cruzada *k-fold* con $k = 10$ y *Grid Search* para la búsqueda de hiperparámetros [1]. La validación cruzada es una técnica que se usa para evaluar la capacidad de generalización de un modelo de aprendizaje automático. En lugar de entrenar y evaluar el modelo una sola vez con un conjunto fijo de datos, la validación cruzada divide los datos en varias particiones, o *folds*. En el caso de 10-fold cross-validation, el conjunto de datos se divide en 10 partes iguales y se entrena el modelo 10 veces, cada vez utilizando 9 *folds* para el entrenamiento y el restante para la validación. Esto asegura que el modelo funcione bien con datos no vistos y ayuda a reducir problemas de sobreajuste, proporcionando una evaluación más robusta del rendimiento.

El mejor modelo fue la regresión logística con *tf-idf*, alcanzando una precisión global de 0.9. Se realizó una búsqueda de hiperparámetros para el valor de C, que controla el inverso de la fuerza de regularización. Los otros modelos también presentaron buenos resultados, especialmente al usar *tf-idf*.

A continuación, se comparan los resultados obtenidos para los modelos con diferentes representaciones, evaluando métricas de precisión, *recall* y F1 con promedios **macro** y **micro**:

- **TF-IDF**: La regresión logística con *tf-idf* presentó resultados superiores, con una precisión promedio de aproximadamente 0.9. Esto se debe a que la representación *tf-idf* reduce el impacto de términos comunes, lo cual permite que el modelo capte eficazmente la importancia relativa de palabras menos frecuentes pero más informativas. Esta técnica también ayuda a reducir el ruido causado por las palabras que son muy comunes en los documentos.

- **TF**: Los resultados con *tf* fueron ligeramente inferiores, con una precisión promedio alrededor de 0.85. Como este método no ajusta la frecuencia según el contexto del corpus, los términos más comunes pueden tener un mayor peso, lo cual afecta negativamente la capacidad del modelo para identificar términos clave que distinguen las categorías.

En conclusión, el análisis demuestra que la representación *tf-idf* resulta más efectiva para este problema de clasificación de noticias. Las técnicas de preprocesamiento y la correcta elección de la vectorización tuvieron un impacto significativo en el rendimiento final, siendo la regresión logística con *tf-idf* la combinación más eficaz para identificar las 20 categorías del conjunto de datos.

2 Análisis de Sentimientos

El objetivo es construir clasificadores para identificar reseñas positivas y negativas en cuatro categorías (Books, DVD, Electronics, Kitchen), utilizando Naive Bayes y regresión logística con *tf*, *tf-idf*, y características basadas en léxicos (*SentiWordNet*).

2.1 Clasificadores por Categoría

Se construyó un clasificador para cada categoría utilizando tres tipos de representaciones del texto: *tf*, *tf-idf*, y características derivadas de un léxico.

Observamos que la representación *tf-idf* ofrece mejores resultados en comparación con *tf* y léxicos, especialmente en las categorías "electronics" y "dvd". Esto se debe a que *tf-idf* reduce el impacto de palabras muy comunes y resalta aquellas que son más informativas para la clasificación. La representación *tf*, por otro lado, muestra un rendimiento ligeramente inferior, ya que las palabras más frecuentes afectan la capacidad del modelo para diferenciar entre reseñas positivas y negativas.

El enfoque basado en léxicos (*SentiWordNet*) mostró un rendimiento significativamente menor, con puntajes F1 alrededor de 0.60-0.64 en comparación con *tf-idf*, debido a la incapacidad del léxico para capturar adecuadamente el contexto específico de las reseñas.

Resultados de F1 por Categoría:

- **Books:** La representación *tf-idf* combinada con regresión logística alcanzó un puntaje F1 de 0.82, mientras que el enfoque basado en léxicos alcanzó un F1 de solo 0.61.
- **DVD:** Similarmente, *tf-idf* tuvo un rendimiento superior con F1 de 0.83, mientras que el enfoque léxico se quedó en 0.60.
- **Electronics y Kitchen:** Ambas categorías mostraron patrones similares, con *tf-idf* logrando un F1 superior a 0.85 y los léxicos alcanzando un máximo de 0.64.

2.2 Análisis de Características Importantes

Utilizando los coeficientes del modelo de regresión logística, identificamos las características más relevantes para la clasificación de cada categoría:

- **Books:** Las palabras más importantes para clasificar como positivo incluyen *excellent*, *great*, *love*, *best*, y *wonderful*, lo cual indica que los términos asociados con recomendaciones y emociones positivas son claves para esta categoría.

- **DVD:** Las palabras más influyentes fueron *great*, *best*, *love*, *excellent*, y *wonderful*. Estos términos sugieren que las reseñas sobre DVDs están fuertemente influenciadas por la experiencia emocional y el disfrute, especialmente relacionado con contenido familiar y entretenimiento.
- **Electronics:** En esta categoría, las palabras *great*, *price*, *excellent*, *perfect*, y *works* fueron las más importantes. Esto muestra que los aspectos técnicos y el valor del producto (como *price* y *works*) son fundamentales para la valoración positiva de artículos electrónicos.
- **Kitchen:** Las palabras *great*, *easy*, *love*, *excellent*, y *perfect* tuvieron el mayor peso positivo. La facilidad de uso (*easy*) y la efectividad (*works*, *perfect*) son aspectos esenciales que los consumidores valoran en productos de cocina.

Para el **clasificador unificado**, observamos que las características más importantes para clasificar una reseña como positiva fueron *great*, *excellent*, *best*, *easy*, y *love*. Esto sugiere que estas palabras representan sentimientos positivos de manera consistente a lo largo de todas las categorías.

Por otro lado, las características más importantes para clasificar una reseña como negativa incluyeron *waste*, *bad*, *disappointed*, *worst*, y *poor*. Estos términos reflejan insatisfacción general, calidad baja, y experiencias negativas que contribuyen significativamente a una clasificación negativa.

Este análisis confirma que las palabras relacionadas con emociones positivas y facilidad de uso tienden a ser las más influyentes para clasificar una reseña como positiva, mientras que los términos que expresan frustración o decepción son los principales indicadores de una valoración negativa.

2.3 Clasificador Unificado

Se construyó un clasificador unificado que abarca todas las categorías. Para cada representación (*tf*, *tf-idf*, léxicos), se utilizó regresión logística y Naive Bayes.

Tanto la regresión logística como Naive Bayes mostraron un rendimiento sólido al usar *tf-idf* y *tf*, con un F1 promedio de 0.85 para la mayoría de los casos. Sin embargo, el enfoque basado en léxicos continuó mostrando limitaciones, con un F1 alrededor de 0.61.

Comparación con los Clasificadores Específicos: El clasificador unificado simplifica la infraestructura y el mantenimiento, pero presenta una ligera pérdida en precisión al compararlo con los clasificadores específicos. Los modelos específicos capturan mejor las características particulares de cada categoría, mientras que el clasificador unificado tiende a ser más general y puede pasar por alto detalles relevantes.

2.4 Conclusión General

Los resultados muestran que la combinación de regresión logística con *tf-idf* es la mejor opción para el análisis de sentimientos en cada categoría de productos, proporcionando una precisión y F1 superiores a 0.85 en la mayoría de los casos. Los clasificadores específicos por categoría tienden a rendir mejor que el clasificador unificado, lo cual sugiere que cada categoría tiene características particulares que son más difíciles de capturar con un único modelo general.

El enfoque basado en léxicos (*SentiWordNet*) presentó dificultades, especialmente en categorías con terminología específica, como "electronics" y "dvd". Esto sugiere que la polaridad de términos debe ser contextualizada para obtener mejores resultados.

References

[1] Scikit-learn. "Cross-validation: Evaluating estimator performance." (), [Online]. Available: `https://scikit-learn.org/stable/modules/cross_validation.html` (visited on 10/03/2024).