Análisis de sentimientos

Artículo #2

"eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico"[1]

1. Resumen del artículo

Los autores se centraron en la clasificación de polaridad de opiniones en español y presentan un nuevo recurso léxico adaptado al dominio turístico, llamado eSOLHotel. Este lexicón usa el enfoque basado en corpus.

Han realizado varios experimentos usando una aproximación no supervisada para la clasificación de polaridad de las opiniones en la categoría de hoteles del corpus SFU. Los resultados que obtuvieron con el nuevo lexicón eSOLHotel superan los resultados obtenidos con otro lexicón de propósito general.

2. Problema que se está resolviendo

Tener un recurso lexicón en español adaptado al dominio turístico para superar los resultados obtenidos con lexicones de propósito general.

3. Base de datos utilizada

Utilizaron un corpus con opiniones extraídas de TripAdvisor para diferentes hoteles de Andalucía.

El corpus generado consiste en una colección de opiniones escritas por usuarios no necesariamente profesionales. Este hecho incrementa la dificultad de la tarea, porque los textos pueden no ser gramaticalmente correctos, incluso contener palabras mal escritas o expresiones informales. Se han seleccionado solo hoteles andaluces. Por cada provincia de Andalucía (Almería, Cádiz, Córdoba, Granada, Jaén, Huelva, Málaga and Sevilla), se han elegido 10 hoteles, siendo 5 de ellos de valoración muy alta y los otros 5 con las peores valoraciones, para obtener las mínimas opiniones neutras en el corpus. Todos los hoteles seleccionados deben tener al menos 20 opiniones escritas en español en los últimos años. Finalmente, se han obtenido 1.816 opiniones.

4. Tipo de caracterización usada para los textos

Aplicaron a los documentos un algoritmo de normalización morfológica basado en la eliminación de prefijos y sufijos (stemmer). El algoritmo de stemming empleado fue el de Porter para español.

Luego, los documentos fueron representados como vectores de unigramas ponderados por el índice de relevancia TF-IDF.

Por tanto, las características que recibía como entrada el algoritmo de aprendizaje automático eran únicamente el valor TF-IDF de los unigramas de los documentos. Por último,

5. Metodología de validación implementada

Se realizó una validación cruzada con el algoritmo SVM

6. Resultados obtenidos

En la Tabla se muestran los resultados obtenidos en la categoría de hoteles del corpus SFU en español usando los lexicones iSOL (independiente del dominio) y eSOLHotel (adaptado al dominio de hoteles).

Lexicón	Precisión	Valor F1	Exactitud
iSOL	77,41%	73,52%	70,0%
eSOLHotel	84,72%	81,22%	78,0%

Lo que permite asegurar que la inclusión de información del dominio en una lista de palabras de opinión genérica mejora los resultados de la clasificación de la polaridad

7. Bibliografía

[1] Molina, M. D., Martínez, E., Martín, M. T., y Jiménez, S. M. (2015). eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico. *Procesamiento del Lenguaje Natural*, *54,* 21-28.