Análisis de sentimientos

Artículo #1

"Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio"*[1]

1. Resumen del artículo

Los autores utilizaron un sistema de análisis de sentimientos llamado Sentitext, este sistema está basado especialmente en conocimiento lingüístico que es independiente del dominio. Los resultados obtenidos fueron alentadores, dada a la alta tasa de acierto en cuanto la polaridad.

2. Problema que se está resolviendo

Cuantificar la proporción de hasta qué punto condicionan los marcadores afectivos específicos del dominio el resultado de la valoración global del texto.

3. Base de datos utilizada

Utilizaron Tripadvisor, un sitio Web destinado a recoger información turística, sobre todo en lo que respecta a hoteles. Tomaron 100 valoraciones aleatorias. Con los siguientes criterios:

- **Similar extensión**: el algoritmo de valoración global que emplea Sentitext tiene en cuenta no sólo el número de unidades léxicas con carga afectiva, sino también el número total de palabras.
- **Original en español:** algunas valoraciones de usuarios incluidas en tripadvisor.es son traducciones automáticas de valoraciones en otras lenguas a través de la herramienta Language Weaver.
- **Homogeneidad en el objeto de crítica:** los hoteles valorados deberían pertenecer a la misma área. Optamos por tomar Londres por ser una de las ciudades con mayor número de valoraciones.
- Distribución proporcional en cuanto al número de críticas para las valoraciones numéricas.

4. Tipo de caracterización usada para los textos

Sentitext se nutre de tres fuentes de datos fundamentales: (i) el léxico de palabras individuales, (ii) el léxico de frases y (iii) las reglas de contexto.

Los textos fueron corregidos ortográficamente utilizando un corrector ortográfico, ya que debe ser ortográficamente correcto, porque si no el lematizador falla y las asignaciones de valencia no son correctas

El proceso de análisis está compuesto de cuatro partes fundamentales[2]:

1. **Lematización** y etiquetado morfológico del texto
2. **Asignación de valencias:** se recorre la lista de unidades léxicas obtenidas y usando como referencia el lema, se busca la valencia de cada una de las unidades en las bases de datos
3. **Aplicación de las reglas de contexto** consiste en recorrer una vez más las lista de unidades léxicas, y en caso de encontrar un modificador que cumpla las restricciones indicadas para una regla de dada (posición, cercanía, y naturaleza del elemento a modificar), se transforma apropiadamente la valencia de la unidad modificada.

   Esta parte es la más delicada, ya que un mismo modificador puede tener asociadas varias reglas de contexto y un mismo elemento modificado puede ser objetivo de varios modificadores, de forma que es necesario especificar un orden jerárquico o unas prioridades en su aplicación.

4. Finalmente, en la **fase de extracción** de datos se obtiene información derivada de los análisis anteriores, como el índice afectivo (cantidad de palabras con carga afectiva en relación con el número total de palabras) o el índice global, que intenta dar una idea aproximada de la positividad o negatividad del texto.

5. Metodología de validación implementada

No es mencionada en las fuentes bibliográficas.

6. Resultados obtenidos

El análisis de los resultados obtenidos lo realizaron desde dos puntos de vista.

1. Comparado la valoración del usuario con la obtenida por el analizador

Dato	Valor
Coincidencia exacta	37%
Diferencia de 1 estrella	52%
Coincidencia en polaridad	89%
Diferencia de 2 estrellas	11%
Diferencia de más de 2 estrellas	0%

2. Estudio pormenorizado del resultado de los análisis en términos de recuperación de información

La herramienta Sentitext les permite saber exactamente qué segmentos textuales han sido etiquetados con qué valencia afectiva. A partir de un valor global expresado en porcentaje de afectividad al que los autores denominaron *gValue*, en el que los valores cercanos al 0% serían negativos y los cercanos al 100% serían positivos.

Hallando resultados de texto a texto, se dieron cuenta que hubo 11 casos en que habían una diferencia de 2 estrellas, por lo tanto estaban erróneos. Pudieron comprobar que la mayoría de estos casos (72,7%) se refieren a valoraciones negativas del usuario, lo que vendría a implicar que ***Sentitext obtiene mejores resultados con textos positivos***

Utilizando otras técnicas como segmento valorativo pudieron confirmar lo anteriormente mencionado y como conclusión se dieron cuenta que Sentitext da una valoración más positiva del texto de la que da el usuario. Típicamente el usuario ha valorado el hotel con una estrella, “pésimo”, y Sentitext lo evalúa con dos estrellas, “malo”.

7. Bibliografía

[1] Moreno, A., Pineda, F., y Hidalgo, R. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext*: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento del Lenguaje Natural*, *45,* 31-39.

[2] Moreno, A., Pérez, Á., y Torres, S. (2010). Sentitext®: sistema de análisis de sentimiento para el español. *Procesamiento del Lenguaje Natural*, *45,* 297-298.