Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles

Elías Rubilar, Francisco Soto, Sofía Leal, Tamara Ruz, Vania Saez

Carlos Henríquez Miranda

Profesor de Ingeniería de Sistemas en la Universidad Autónoma del Caribe.
Algunos proyectos e investigaciones:

- *Modelo de extracción de información desde recursos web para aplicaciones de la planificación automática (2012).*

- *Sistema de control de acceso basado en Java Cards y Hardware libre (2010).*

- *Análisis de sentimientos a nivel de aspecto usando ontologías y aprendizaje automático (2017).*

Jaime Guzmán Luna

Profesor titular de la Facultad de Minas en la Universidad Nacional de Colombia, Bogotá.

Algunos de sus proyectos e investigaciones son:

- *Metodologías y métodos para la construcción de ontologías (2012).*

- *IA planning for automatic generation of customized virtual courses (2005).*

- *Desarrollo de una ontología en el contexto de la web semántica a partir de un tesauro documental tradicional (2006).*

Dixon Salcedo

Investigador Junior y Profesor Adjunto de la Corporación Universidad de la Costa CUC, especializado en la tecnología y la ingeniería.

Algunos de sus proyectos e investigaciones son:

- *“WIFI”, una solución para cerrar la Brecha Digital (2013).*
- *¿” Inclusión o Exclusión digital”? En las Instituciones educativas oficiales del distrito de Barranquilla (2010).*
- *Overhead in available Bandwidth estimation tools: Evaluation and Analysis (2017).*

1. Introducción

- El masivo uso de internet genera que actualmente existan en la web una cantidad enorme de datos generados por los usuarios en las diversas páginas, plataformas y redes sociales.
- Dificultad para analizar todos los textos producidos de manera manual.
- Surgen sistemas de inteligencia artificial de procesamiento de lenguaje.
- Técnica de Minería de opiniones (MO) o también llamada análisis de sentimientos.

¿Qué es la minería de opiniones?

⇨ Sistema de extracción de información subjetiva a partir de contenidos generados por los usuarios.

✓ comentarios

✓publicaciones en blogs

✓calificación de productos o servicios

⇨ Analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia entidades como productos, servicios, organizaciones, individuos, problemas, sucesos, temas dentro de la web.

Objetivo del texto

- La mayoría de investigaciones que trabajan con la minería de opiniones están en inglés.
- Este texto busca trabajar con una construcción de MO en español a partir de la adaptación al español de las normas afectivas para las palabras en inglés (ANEW).
- Extraer y analizar los comentarios que dejan en internet los clientes de diferentes hoteles.

2. Antecedentes y trabajos relacionados

ANEW: The Affective Norms for English Words (ANEW) is being developed to provide a set of normative emotional ratings for a large number of words in the English language. The goal is to develop a set of verbal materials that have been rated in terms of pleasure, arousal, and dominance to complement the existing International Affective Picture System.

The International Affective Picture System (IAPS): is a database of pictures designed to provide a standardized set of pictures for studying emotion and attention that has been widely used in psychological research. The IAPS was developed by the National Institute of Mental Health Center for Emotion and Attention at the University of Florida.

OASIS: http://www.benedekkurdi.com/

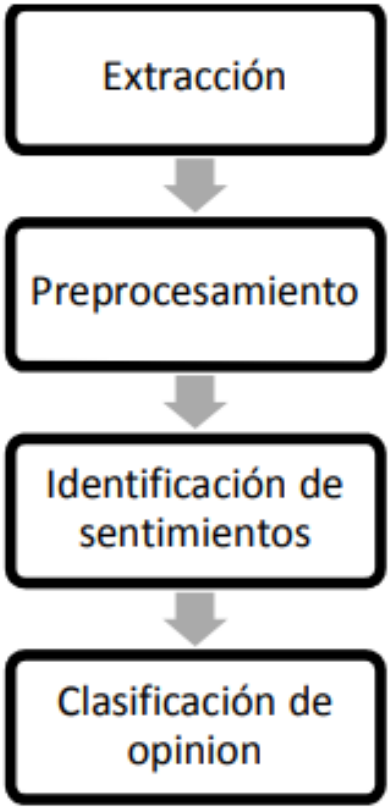
Cambridge Analytica

Cambridge Analytica es una empresa con sede en Londres que usa el análisis de datos para desarrollar campañas para marcas y políticos que buscan "cambiar el comportamiento de la audiencia", según indica su sitio web.
La obtención de perfiles de 50 millones de usuarios de Facebook no fue obra de Cambridge Analytica, sino que se atribuye al profesor de la Universidad de Cambridge Aleksandr Kogan.
A modo de proyecto personal, Kogan desarrolló en 2013 un test de personalidad en formato de aplicación de Facebook.

Análisis de la subjetividad

Análisis de sentimientos (AS)
Minería de opiniones (MO):
-Extracción de la opinión: datos, que van desde redes sociales, hasta sitios web donde abundan opiniones y comentarios en línea.
-Clasificación del sentimiento: Para la clasificación se usan técnicas basadas en *Machine Learning* y en Léxico (LEX): Dependen de diccionarios de sentimientos.

3. Metodología: modelo para minería de opiniones



extrae el texto que contiene las opiniones.

Aplicación de diferentes técnicas para limpiar el texto

con el texto ya normalizado, se etiqueta morfológicamente cada palabra con su categoría gramatical

se le asigna una polaridad a las palabras

Diccionario de sentimientos: 71 adjetivos positivos y 52 negativos, que se complementan con sinónimos y antónimos hasta llegar a 600 términos en total

- manejo básico de negación: se analiza la inclinación de la opinión y luego se afecta el promedio calculando que afecta a la polaridad

- 1ra parte: se basa en el tratamiento propuesto por Zafra donde se parte de partículas de corte negativo: no, tampoco, nadie, jamás, ni, sin, nada, nunca y ninguno + no, mal y malo

- se calcula el n° de veces que aparecen, si es un n° considerable se recalcula la polaridad dándole un peso de 50% al promedio y 50% al valor más negativo.

4. Experimentos y resultados

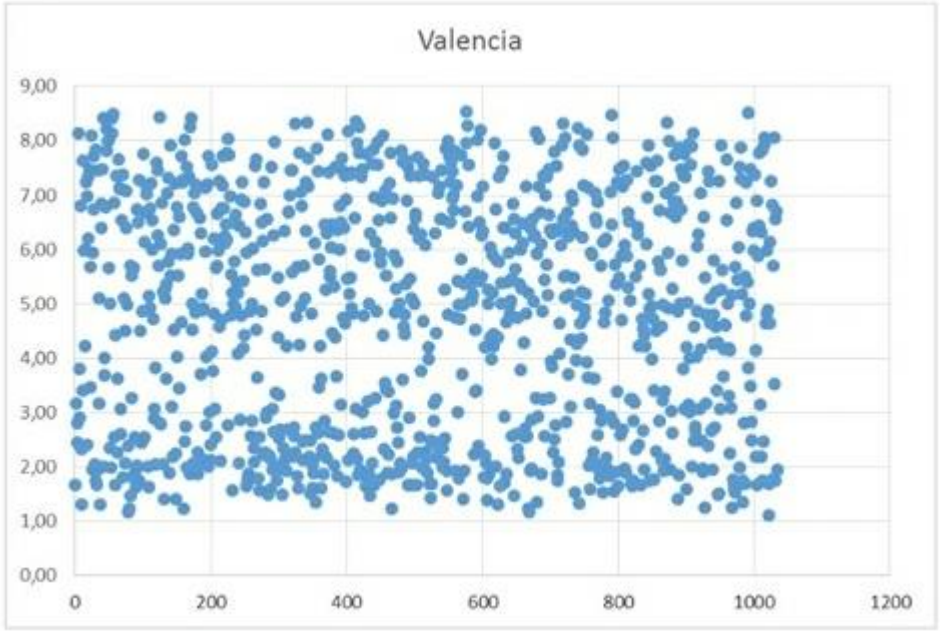
- **Herramientas usadas:** Java 7.0, JDOM, FREELING
- Para probar el sistema se escoge el dominio de turismo (opiniones acerca de hoteles). Los datos son tomados de TripAdvisor, más específicamente de un corpus trabajado por Molina-González, M. et al.

Fase de preprocesamiento

- **Corpus:** 1816 opiniones, 5 niveles de opinión (del 1, negativo, al 5, positivo)
- **Fase de preprocesamiento:** Fase de normalización (se eliminan símbolos y palabras sin sentido, se cambian las letras mayúsculas a minúsculas y se lleva a cabo un proceso de *lematización* -buscar el lema de una palabra, tal como se encuentra en un discurso textual-)
- **Identificación de sentimientos:** se utiliza la técnica de etiquetado morfológico (asignar a cada palabra su categoría gramatical). Hecho el etiquetado, se toman en cuenta solo los verbos, adverbios y adjetivos.

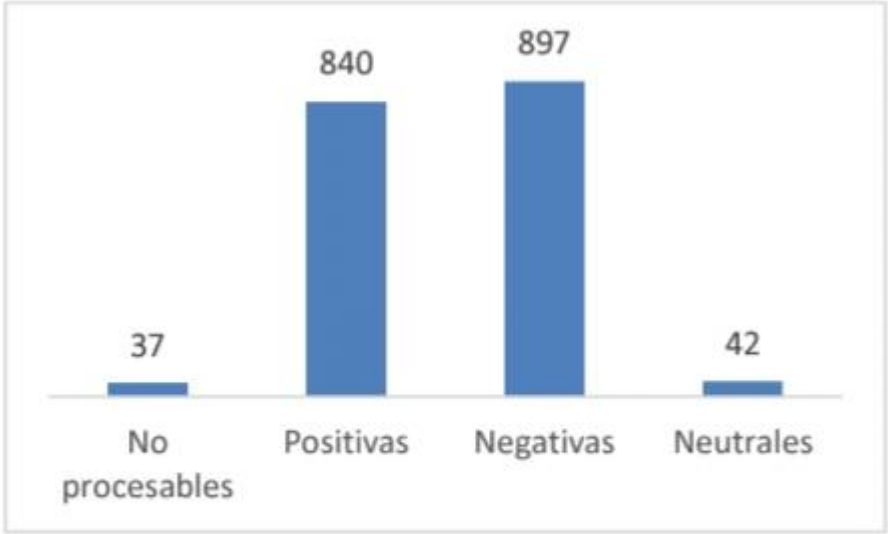
Fase de clasificación de la opinion

- Se toman las palabras de la fase anterior, se buscan en ANEW y se procesa la valencia



Prueba del sistema

- Resultados de las 1816 opiniones del corpus
- **Se hacen 4 experimentos:**
  - Se analizan las palabras solo con ANEW y sin manejo de negación
  - Se analizan las palabras solo con ANEW y con manejo de negación
  - Se analizan las palabras con ANEW, con un diccionario de sentimientos hecho manualmente y sin manejo de negación
  - Se analizan las palabras con ANEW, con un diccionario de sentimientos hecho manualmente y con manejo de negación



Validación de los experimentos

- Los experimentos se validan a través de una **medida de precisión** (los ejemplos positivos clasificados correctamente se dividen los ejemplos etiquetados por el sistema como positivos)
- Usando solo ANEW se obtiene una precisión máxima de 92%
- Usando ANEW y el diccionario de sentimientos, la precisión aumenta a 94% (esto porque el diccionario incluye palabras que no están en ANEW)
- Se compara el sistema con trabajos similares anteriores (que tengan enfoque léxico de sentimientos y dominio en hoteles)

Experimento	Precisión
1	89
2	92
3	93
4	**94**

Trabajo	Opiniones	Léxico	Precisión
(Moreno, Castillo, y García, 2010)	100	*Sentitext*	84.8
(García, Gaines, y Linaza, 2012)	994	*Léxico propio*	80.0
(González, Cámara, y Valdivia, 2015)	1816	*eSOLHotel*	84.7
Propuesta	**1816**	***Anew***	**94.4**

5. Conclusiones

Conclusiones de los Autores:

- Mencionan que, con lo trabajado, se ha logrado un sistema más que aceptable y de una altísima precisión para el dominio del turismo, utilizando el recurso lingüístico independiente ANEW.
- El uso de varios algoritmos ayuda a complementar el cálculo de la polaridad.
- El uso de herramientas de PLN potentes ayuda a construir sistemas de opiniones de minería más robustos.
- La calidad de recursos ANEW permitirá realizar nuevos experimentos de análisis de sentimientos enfocados a diferentes dominios empleando la metodología que se ha propuesto aquí.

5. Conclusiones

Conclusiones Personales:

La MO es un proceso arduo, pero no necesariamente complejo: La clasificación de las palabras (“etiquetado”) es una labor mecánica y de baja dificultad que puede realizarse con la ayuda de un diccionario. La principal dificultad de este procedimiento es, en realidad, la amplia envergadura de los corpus a clasificar.

Por otro lado, cabe destacar que es altamente cuestionable que estos procedimientos analicen las emociones realmente. Más bien se trata de la asignación de valores convencionales a palabras de uso general. Esto pudimos verlo con el ejemplo del proyecto Oasis. Queda la duda de si los valores asignados son realmente objetivos o al menos extrapolables a una parte significativa de la población.

5. Conclusiones

Conclusiones Personales:

Estos trabajos e investigaciones desde hace algunos años, demuestran el gran interés que se debe de tomar respecto a este tema, pues hoy en día, se trata de algo que viene a ser muy dinámico, pero también muy amplio, respecto a los diferentes informes y enfoques que varios investigadores, además de los ya mencionados aquí, proponen a la hora de trabajar y analizar los corpus.

Finalmente, cabe destacar lo curioso que resulta que investigaciones de este corte se piensen desde y para el marketing empresarial. ¿Cuál es el aporte que esta clase de sistematizaciones agrega al estudio del español? Los parámetros de clasificación, como vimos, no suelen ser más de 3 en ninguno de los estudios antecedentes, ni en el presente. ¿La gama de valores asignables a las palabras del español son tan sencillamente reductibles? ¿Qué nos dicen estas investigaciones sobre la “originalidad” de la subjetividad sensible humana? Quedan muchas e interesantes cuestiones que pensar en torno al tema y sus consecuencias. ¿Qué opinan ustedes?