

Análisis de sentimientos

Shaira Pérez
Escuela de Computación
Licenciatura en Computación
Universidad Central de Venezuela
Caracas, Venezuela
Email: shairaperez.sp@gmail.com

Manuel Araujo
Escuela de Computación
Licenciatura en Computación
Universidad Central de Venezuela
Caracas, Venezuela
Email: manuel.araujo1511@gmail.com

Abstract—El análisis de sentimientos o minería opinin es el campo de estudio que analiza las opiniones, sentimientos, valoraciones, actitudes y emociones de la gente en lenguaje escrito. Es un área donde el procesamiento del lenguaje natural, la miner 'ia de datos, minería web y minería de texto se encuentran. Por ello, presentamos algunos conceptos necesarios para la comprensión de las características e implementación de estos mtodos para la concepcin de un sistema de un sistema de análisis de sentimientos.

1. Introducción

En los últimos aos, ha habido un enorme crecimiento en el uso de medios sociales tales como blogs, microblogs , Twitter y las redes sociales en general para realizar críticas, calificaciones, recomendaciones y otras formas de expresión en Inea. Lo que ha hecho que la opinión en Inea se convierta en una especie de divisa virtual para negocios que buscan comercializar sus productos, identificar nuevas oportunidades y manejar sus reputaciones. Es por ese crecimiento, que investigan cada vez más nuevas formas de abrir una mina de información en estos sitios, para conocer además de lo anterior, lo que las personas piensan y sienten acerca de sus productos y servicios.

Es por todo esto que surge el análisis de sentimientos, como una manera de conseguir esta información tan determinante para las empresas y organizaciones. Este de manera básica consiste en CITA determinar la actitud de un interlocutor o un escritor con respecto a algún tema o, la polaridad contextual general de un documento. Los propósitos de este artículo son:

- Explicar el funcionamiento de los sistemas de análisis de sentimientos.
- Presentar las características de los principales enfoques de solución para los sistemas de análisis de sentimiento.
- Mostrar los múltiples usos del análisis de sentimientos.
- Exhibir las predicciones para esta área.

May 7, 2016



Figura 1. Mr Tuit: Sistema de Análisis de Sentimientos. Por [InfoMrTuit](#).

2. Conceptos generales

Convengamos en una definición para cada término referido posteriormente:

- **Lingüística Computacional:** Es un campo interdisciplinario entre el modelado basado en reglas y el modelado estadístico del lenguaje natural desde una perspectiva computacional. Dicho modelado no se centra en ninguna de las áreas de la lingüística en particular, sino que es un campo en el que participan lingüistas e informáticos especializados en inteligencia artificial, psicólogos cognoscitivos y expertos en lógica, entre otros.CITA

- **Procesamiento de Lenguaje Natural:** Es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales. De manera simple, es el conjunto de instrucciones que una computadora recibe en un lenguaje de programación dado (formal), que le permitirán comunicarse con un humano en su propio lenguaje, (inglés, francés, español, etc). CITA
- **Minería de datos:** Es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. Siendo el objetivo general la extracción de la información y transformarla en una estructura comprensible para su uso posterior. En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo" CITA
- **Minería de texto:** Consiste en descubrir, a partir de grandes cantidades de texto, el conocimiento que no está literalmente escrito en cualquiera de los documentos. Esto incluye buscar tendencias, promedios, desviaciones, dependencias, etc. La información que se descubre debe ser desconocida incluso por los autores de los documentos que se hayan tomado como punto de partida del proceso. También tiene entre sus objetivos la extracción y visualización de la información procedente de grandes corpus textuales en un formato que facilite su análisis y la deducción de nuevas conclusiones. CITA

3. Proceso básico de un sistema de análisis de sentimientos

Mediante el análisis del sentimiento, se aspira entender, en primer lugar, con qué guarda relación el texto que se analiza. Por ejemplo, interesa saber si la frase en cuestión trata sobre la marca o sobre el concepto que interesa analizar (p. ej. "hoteles"). Idealmente, se querrá saber esto pero con mayor nivel de granularidad: a qué se refiere exactamente la frase? A la marca, en general, o a un aspecto del producto o servicio (envase, precio, la habitación, la atención recibida, entre otros)?

Una vez que el sistema tiene claro sobre qué es la opinión, el segundo objetivo será conocer el sentido de esa opinión (positiva o negativa, básicamente).

Los sistemas más sencillos se limitan a leer una frase y buscar en ella palabras que tienen registradas en su diccionario como buenas o como malas. Si aparece una palabra buena (p. ej. "estupendo"), es una opinión positiva. Si aparece una palabra mala (p. ej. "decepcionada"), es una opinión negativa. Si no encuentran ninguna palabra que,

según su diccionario, exprese algo negativo o positivo, estos sistemas más sencillos suelen hablar de "opiniones neutras", aunque sería preferible entender que lo que ocurre es que no hay opinión como, por ejemplo, si un tweet contiene la frase "Estoy esperando a mi primo en el McDonalds". Paradójicamente, muchos de esos sistemas simples también dirán que una opinión es neutra si han encontrado una palabra negativa y otra positiva.

A partir de ese mínimo, los distintos sistemas de análisis del sentimiento que hay en el mercado van incrementando su refinamiento y su capacidad de proporcionar información realmente útil.

4. Enfoques en la implementación de los sistemas

En este punto, es importante recordar que existen diversos **campos científicos** implicados en esta área por lo que la selección de un método determinará en gran medida el enfoque a utilizar en el desarrollo del sistema. Los enfoques existentes se pueden clasificar de la siguiente manera:

4.1. Enfoque semántico

se caracterizan por el uso de diccionarios de términos, **lexicons**, con orientación semántica de polaridad u opinión, es decir, los términos están clasificados como positivos o negativos básicamente. En principio los sistemas pre-procesan el texto donde, dividen en palabras, eliminan las palabras de parada que sean conveniente (ya que las mismas agregan información que puede ser importante para el análisis, por lo que no pueden eliminarse por completo) y una normalización lingüística por stemming¹ o lematización², y luego comprueban la aparición de los términos en el lexicon para asignar el valor de polaridad del texto mediante la suma de los valores de polaridad de los términos. Típicamente los sistemas además incluyen un tratamiento más o menos avanzado de:

- Términos modificadores (como muy, poco, demasiado) que aumentan o reducen la polaridad de los términos a los que acompañan.
- Términos inversores o negadores (como no, tampoco), que invierten la polaridad de los términos a los que afectan.

La ventaja de este enfoque es que los errores son relativamente sencillos de corregir, añadiendo cuantos términos fueran necesarios, y se podría obtener una precisión tan alta como se quisiera, simplemente invirtiendo más tiempo en la construcción del *lexicon*. Sin embargo, el esfuerzo para construir un *lexicon*, empezando de cero, es muy elevado,

1. Es un método para reducir una palabra a su raíz

2. Es un proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente

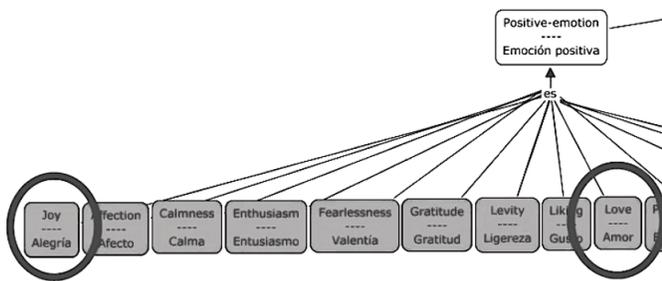


Figura 2. Ejemplo de un lexicon de palabras positivas.CITA

porque se basa en mucho trabajo manual, así que en general son menos viables.

4.2. Enfoque basado en aprendizaje computacional

Consiste en entrenar un clasificador usando un algoritmo de **aprendizaje supervisado** a partir de una colección de textos anotados, donde cada texto habitualmente se representa con un vector de palabras, *bag of words*³, *n-gramas*⁴ o *skip-grams*⁵, en combinación con otro tipo de características semánticas que intentan modelar la estructura sintáctica de las frases, la intensificación, la negación, la subjetividad o la ironía. Los sistemas utilizan diversas técnicas, aunque las más populares son los clasificadores basados en SVM⁶, Naive Bayes⁷ y KNN K-Nearest Neighbor, K-vecinos más cercanos. En las investigaciones más recientes se han empezado a utilizar otras técnicas más avanzadas, como LSA⁸ e incluso Deep Learning⁹.

La ventaja de los enfoques basados en aprendizaje automático es que cuesta muy poco construir un analizador de sentimientos a partir de la colección de textos etiquetados¹⁰, ya que la tarea de modelado reside en el algoritmo. Por ello es relativamente fácil construir clasificadores adaptados a un dominio determinado. No obstante, en general necesitan un reentrenamiento completo para portarlos a un dominio diferente, además suelen ser una caja negra en la que corregir errores o añadir nuevo conocimiento es más complicado, y muchas veces sólo es posible ampliando la colección de textos y volviendo a entrenar el modelo.

4.3. Enfoque basado en aspectos

Se refiere a determinar las opiniones o sentimientos expresados sobre diferentes rasgos o características de en-

3. Es un método que se utiliza en el procesamiento del lenguaje para representar documentos ignorando el orden de las palabras

4. Es una subsecuencia de n elementos de una secuencia dada.

5. Son una generalización de n-gramas donde pueden haber huecos por omisión. Proporcionan una forma de superar el problema de escasez de datos

6. Support Vector Machines, Maquinas de Soporte Vectorial

7. Bayes ingenuo

8. Latent Semantic Analysis, Análisis Semántico Latente

9. Aprendizaje profundo

10. Las palabras del texto están etiquetadas con su categoría gramatical(sustantivo, adjetivo, verbo, etc)

tidades. Un rasgo puede ser una entidad, un concepto, una etiqueta temática, o, en general, cualquier dimensión de análisis que resulte de interés. Por ejemplo, de un teléfono celular, una cámara digital, o un banco. Este problema involucra varios sub-problemas, por ejemplo, identificar entidades relevantes, extraer sus rasgos/ características, y determinar si una opinión expresada sobre cada rasgo / característica es positiva, negativa o neutral.

Es posible ampliar el sistema con diccionarios de usuario, incluyendo definiciones de entidades, conceptos y aspectos propios de cada dominio concreto, se puede dar respuesta a cualquier necesidad de análisis que se plantee.

5. Evaluación del sistema

La precisión de un sistema de análisis de sentimiento es, en principio, cuánto coincide con juicios humanos. Esto es usualmente medido mediante precisión y recuperación. Sin embargo, de acuerdo a investigaciones evaluadores humanos normalmente están de acuerdo en un 79% de las veces. De este modo, un programa acertado un 70% de las veces se comporta casi tan bien como los humanos, a pesar de que tal precisión pueda no sonar impresionante. Si un programa estuviera "bien" el 100% de las veces, los humanos aún estarían en desacuerdo con él alrededor de un 20% de las veces, puesto que están en desacuerdo lo mismo con cualquier respuesta. Es posible aplicar mediciones más sofisticadas, pero la evaluación de sistemas de análisis de sentimientos sigue siendo un asunto complejo. Para las tareas de análisis de sentimiento que devuelven una escala en lugar de juicios binarios, la correlación es una mejor medida que la precisión porque tiene en cuenta cuán cercano es el valor predicho del valor esperado.

Sin embargo, en función del contexto, las cifras de acierto pueden andar docenas de puntos por debajo de la perfección. Los motivos más habituales son:

- **La complejidad y riqueza de los idiomas** y de los lenguajes permite hacer frases tremendamente complicadas, jugar con conocimiento implícito que no se refleja de forma explícita en los actos de habla o, sencillamente, confiar en el contexto para que se entienda el significado real de lo que queremos decir, contexto que no tiene por qué estar disponible para el motor de análisis. Se pueden detectar las ironías más evidentes, pero no las más sutiles, por ejemplo.
- En ocasiones, las distintas variedades dialectales¹¹ o el argot de los diversos grupos sociales puede impedir el correcto análisis de algunos textos, aunque esto es algo que se puede amortiguar si el proveedor del sistema es suficientemente meticuloso.

6. Usos del análisis de sentimientos

Los beneficios del análisis de sentimiento son numerosos e importantes. La habilidad de extraer información de datos

11. Hace referencia a las posibles variedades de una lengua

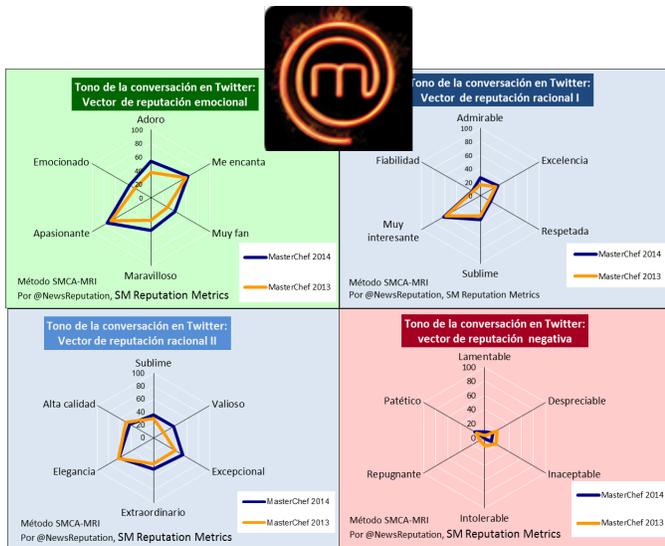


Figura 3. Ejemplo de análisis de sentimientos en Twitter sobre MasterChef

de las redes sociales es una práctica que ya están adoptando organizaciones a nivel mundial.

Se ha probado que los cambios en el sentimiento de las redes sociales se corresponden a **cambios en la bolsa**. CITA La administración de Obama utilizó el análisis de sentimiento para sondear la opinión pública sobre sus políticas y mensajes de campaña antes de las **elecciones presidenciales** del 2012.

Lo habitual es identificar el análisis del sentimiento con su uso en las redes sociales pero, en realidad, su aplicación va mucho más allá. Analizar tweets es importante, pero muchas marcas no se limitan a eso y monitorizan de forma constante toda la web, principalmente aquellos lugares donde es más probable que se viertan opiniones relevantes, como los blogs de temas específicos.

Las marcas que cuentan con un buen motor de análisis del sentimiento, extraen conclusiones y toman decisiones concretas basadas en la información que les proporcionan: modificar elementos de su offering, lanzar acciones correctivas cuando detectan un problema con la experiencia del cliente o construir sus mensajes de marketing.

7. Caso de Estudio: Expedia Canadá

La habilidad de entender rápidamente actitudes del consumidor y reaccionar adecuadamente es algo que Expedia Canadá aprovechó cuando se dieron cuenta de que había un crecimiento continuo de feedback negativo hacia la música de uno de sus anuncios de televisión.

“El peor anuncio, lo repiten tanto que llega a ser molesto, y probablemente no vuelva a usar Expedia solo porque su anuncio es tan irritante.”

El análisis de sentimiento que la marca llevó a cabo reveló que la música de uno de sus anuncios llegaba a ser muy irritante tras varias escuchas, y los consumidores se estaban dirigiendo a las redes sociales para desahogarse.

Un par de semanas después de que el anuncio se emitiera por primera vez, más de la mitad de los comentarios de la conversación online eran negativos.

En vez de anotarse un fallo con el anuncio, Expedia supo dirigir el sentimiento negativo de una manera divertida y consciente al emitir una nueva versión del anuncio que mostraba cómo rompían el violín que tanto molestaba.

8. Predicciones

Es difícil especular sobre cómo un sistema relativamente inmaduro evolucionará en el futuro, pero hay una creencia general de que el análisis de sentimiento necesita ir más allá de una escala unidimensional positiva-negativa.

Del mismo modo que la política no siempre se puede reducir a una posición en una escala izquierda-derecha, hay algunos tipos de sentimiento que no se pueden colocar en un simple barómetro.

En el futuro, para capturar y entender de verdad el amplio rango de emociones que los humanos expresan por escrito, se necesitará una escala multidimensional más sofisticada. Se puede medir el escepticismo, la esperanza, la ansiedad, la emoción o la falta de ella? Hasta que esto ocurra el análisis del sentimiento es (literalmente) unidimensional.

Las organizaciones estarán más al tanto de las aplicaciones del análisis de sentimiento a su mercado.

Veremos un cambio en la percepción de la fiabilidad del análisis de sentimiento. Los usuarios estarán más abiertos a la idea de que el análisis automático de textos difícilmente igualará al desempeño humano.

La información que se puede extraer de grandes conjuntos de datos (millones de tweets) eclipsará la preocupación sobre la fiabilidad a nivel granular (un solo tweet).

9. Conclusión

En este artículo, hemos abordado la compleja tarea de explicar el, aún naciente, contexto del análisis de sentimientos y algunos conceptos asociados a esta área.

La información constituye, en las condiciones actuales, un recurso económico altamente valorado no sólo por sus propiedades intrínsecas, sino porque permite, además, el perfeccionamiento del empleo del resto de los recursos de las organizaciones. De ahí que la gestión de la información y del conocimiento, como estadio superior, cobre cada día mayor importancia en la elevación de su competitividad.

En este sentido, el análisis de sentimiento es una herramienta poderosa, con la cual se puede obtener una gran cantidad de información valiosa para la empresa u organización. Y como gran parte de sus datos provienen de fuentes como Twitter, Facebook o blogs, los cuales crecen en volúmenes inmanejables la ciencia de datos juega un papel relevante en este proceso, ya que permite el almacenamiento y procesamiento de grandes cantidades de datos.