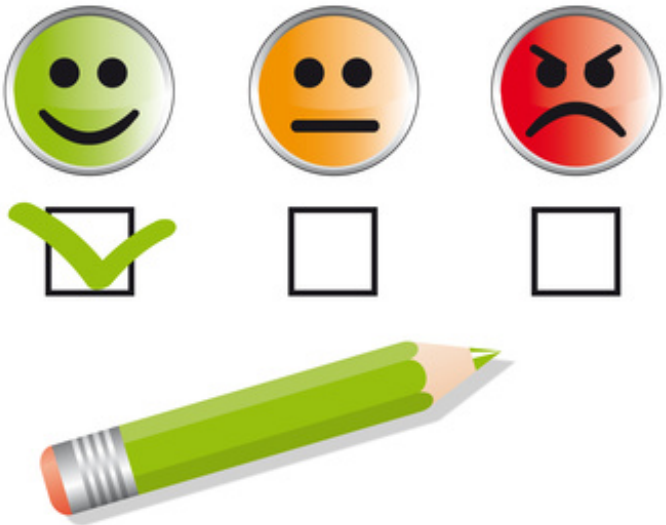
Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias

Emanuel García Pérez

October 31, 2014

- En los últimos años el alcancen de los blogs, foros y redes sociales han hecho que millones de usuarios los usen para expresar sus opiniones sobre diversos temas.
- Esta gran variedad de opiniones y críticas que abundan en la web son de gran utilidad para que vendedores y fabricantes conozcan el impacto de sus productos en los consumidores.



Debido a las ventajas que presenta toda esa información y la complejidad que conlleva su análisis es necesario la búsqueda de soluciones eficientes capaces de monitorizar este flujo de información. La minería de opinión(MO) se enfoca en el tratamiento automático de información con opinión, lo que permite extraer la polaridad(positiva, negativa, neutra, mixta) de un texto.

Polaridad en textos

La clasificación de la polaridad en textos es un problema de gran relevancia en la MO, este ha sido abordado desde dos enfoques:

- Concebir esta tarea como un proceso genérico de clasificación, y a partir de un conjunto de datos de entrenamiento, donde los textos son anotados con su respectiva polaridad, se construye un clasificador mediante aprendizaje automático(AA).
- Apoyarse en la orientación semántica(OS) de las palabras, cada término que expresa opinión es anotado con un valor que representa su polaridad.

Sistemas de MO

La mayoría de los sistemas de MO existentes se enfocan en tratar únicamente textos en ingles. Para textos en español el, quizás, sistema más relevante es "The Spanish SO Calculator", desarrollado en la universidad Simon Fraser de Canadá.

Spanish SO CA

- Resuelve la OS almacenada a nivel individual en: adjetivos, sustantivos, verbos y adverbios.
- Trata modificadores de la polaridad, como la negación o intensificadores("muy", "poco", "bastante").
- Detecta y excluye el sentimiento reflejado en el contenido no fáctico del texto(expresiones condicionales o subjuntivas.)
  La forma más común de tratar estas construcciones lingüisticas es a nivel léxico, y "The Spanish SO CA" no es la excepción.

Negación

Respecto a la negación tenemos lo siguiente:

- Taboada, 2011: utiliza información morfológica para identificar el alcance dela negación.
- Yang, 2008: considera el alcance de la negación como los términos a la derecha de la negación.
- Fernández Anta, 2012: se emplea una heurística que asume que los tres elementos a continuación de una negación son los que deben cambiar su polaridad.

Intensificación

Para los intensificadores tenemos:

- Fernández Anta, 2012: considera de nuevo que los tres términos a la derecha son los que deben variar su polaridad.
- Taboada, 2011: además de los intensificadores propiamente dichos, trata como tales aspectos del discurso como la conjunción "pero" o las mayúsculas.

La propuesta utilizada se basa en obtener la estructura sintáctica del texto para tratar las construcciones lingüisticas e identificar los elementos de la frase que están implicados en ellas.
Trabajos anteriores(Jiu, yuMeng, 2009) han mostrado los beneficios de utilizar la estructura sintáctica de la frase en textos con ocurrencias de términos negativos.

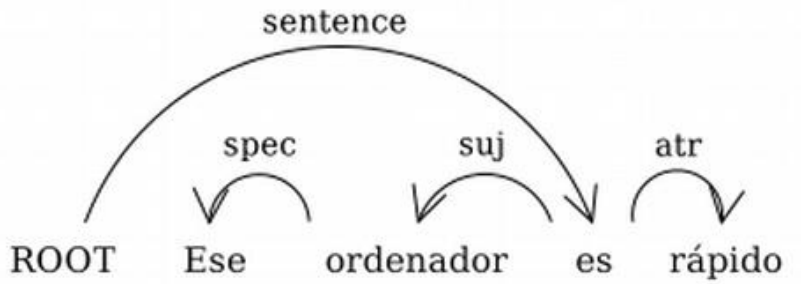
Otro problema que enfrentan los sitemas de MO es la calidad ortográfica de los textos. Si estos provienen de la web, se debe tener en cuenta que a menudo sus autores omiten acentos, letras, vocablos, etc., o utilizan abreviaturas no reconocidas y oraciones agramáticales.
La solución más utilizada es usar patrones heurísticos para adaptar el texto.

En contraste a las propuestas léxicas dominantes, se propone utilizar la estructura sintáctica de la frase para obtener la OS de un texto. Inicialmente se debe hacer un preprocesamiento de los textos que contemple los siguientes aspectos:

- Unificación de expresiones compuestas que actuan como una sola unidad de significado("a menos que").
- Normalización de los signos de puntuación.

Posteriormente se debe segmentar el texto en oraciones y tokenizar cada una, para después hacer el etiquetado morfosintáctico de cada palabra del texto.

El siguiente paso es realizar el análisis sintáctico de dependencias para identificar las relaciones binarias "padre/dependiente" entre los términos de una oración. Se incluye un elemento inicial(ROOT) para facilitar las definiciones formales e implementaciones. Cada vinculo binario constituye una dependencia, que se anota con la función sintáctica que relaciona los dos términos(Árbol de dependencias). Se utilizo el corpus Ancora(Taulé, Martí, Recasens, 2008) como referencia para definir las relaciones de dependencia.



Finalmente para realizar el análisis semántico, se utiliza SODcitionariesV1.11Spa(Brooke, Tofiloski, Taboada, 2009), un conjunto de diccionarios de polaridad para adjetivos, sustantivos, verbos, adverbios e intensificadores. Cada término tiene asociado un valor entre -5 y 5, siendo -5 lo más negativo y 5 lo más positivo. El valor asignado a cada palabra corresponde con una OS genérica, independientemente del dominio o contexto donde se use.

Ej: rápido(adj):= 2, recomendar(verb):= 2

Diccionario	Nº términos
adjetivos	2,049
sustantivos	1,324
verbos	739
adverbios	548
intensificadores	157

Los valores numericos asociados a los intensificadores tiene un significado distinto, representan el porcentaje(positivo o negativo) por el que modifican el sentimiento de la expresión a la que afectan.

Se determina la polaridad de un texto a partir de la combinación de la OS de sustantivos, adjetivos, verbos, y adverbios, sin considerar la estructura sintáctica del texto(construcciones lingüisticas complejas).

Los intensificadores son términos o expresiones que modifican la polaridad de ciertas palabras. Podemos clasificarlos en dos tipos:

- Amplificadores: permiten aumentar la polaridad; ("muy", "bastante").
- Decrementadores: permiten disminuir la polaridad; ("poco", "en absoluto").

Para modelar esta construcción, se asocia a cada intensificador un factor de ponderación. Se utiliza el árbol de dependencias para determinar la parte de la frase que se ve afectada por tal modificación, considerando las dependencias anotadas en Ancora.

Si hay varios intensificadores presentes, se combinan sus porcentajes de intensificación antes de que actuen sobre el término afectado.
Ej: "muy" := 0.25, "en absoluto" := -1
"muy rápido"—OS: 2*(1+0.25) = 2.5
"en absoluto muy rápido"—OS: 2*(1+(-1+0.25)) = 0.5

Los nexos adversativos contraponene hechos expresados en dos oraciones. En MO este tipo de frases se emplean para restringir o excluir opiniones, que puede ser considerado como un caso especial de intensificación.
Un árbol de dependencias permite identificar con precisión la oración subordinada y la subordinante.

El corpus Ancora representa sintácticamente este tipo de oraciones de forma diferente según el nexo concreto, se opto por hacer enfoque en los nexos más relevantes que Ancora representa de manera uniforme; siendo estos divididos en dos grupos:

- Restrictivos: reducen la OS de la oración principal. Destaca la conjunción "pero".
- Excluyentes: eliminan por completo lo expresado en la primera oración; "sino".

Según la clase del nexo, se pondera el sentimiento acumulado, tanto en la oración subordinante y la subordinada, de forma distinta.
Para homogenizar la estructura sintáctica de otras subordinadas adversativas y simplificar la ponderación de estas oraciones, se opto por reestructurarlas en el árbol de dependencias. Se creó un nuevo tipo de dependencia, art_rel_adversative, para identificar sintácticamente el inicio de una cláusula de este tipo.

Muchos términos o expresiones permiten negar una opinión, la distinción de un negador y un intensificador decrementador suele ser difusa. Se restringio el tratamiento de esto a los términos "no", "nunca" y "sin". Expresiones negadoras como "lo menos" o "en absoluto" fuerón tratadas como intensificadores, se uso la información semántica de SODictionariesV1.11Spa para este tipo de locuciones.
Para resolver el sentimiento de una oración con términos negativos se necesitan dos pasos:

- Identificar el alcance de la negación.
- Modificar la polaridad del fragmento de la oración correspondiente.

En primer lugar se establece un alcance candidato, formado por el padre del negador y sus hermanos, después se corrige ese alcance aplicando las siguientes reglas:

1. Regla del padre subjetivo: si el padre del negador aparece en los diccionarios semántico, entonces solo él constituye el alcance corregido de la negación.
2. Regla del atributo o complementeo directo: si alguno de los hermanos desempeña una de estas funciones sintácticas, entonce dicho hermano forma parte del alcance de la negación.
3. Regla del complemento circunstancial más cercano: si alguna rama al mismo nivel del negador actua como complemento circunstancial, entonces dicha rama forma el alcance corregido. En caso de varios complementos circunstanciales, solo se incorpora el más cercano físicamente al negador.

Si ninguna de las reglas se cumple, entonces se asume el alcance candidato(salvo el nodo padre) como el corregido.

Una vez obtenido el alcance corregido de la negación, se extrae su polaridad y después el valor obtenido es modificado en una cantidad preestablecida de signo contrario.

Se realizo la implementación en Python usando el toolkit NLTK, para la etiquetación se aplico el algoritmo de Brill utilizando el corpus Ancora para el entrenamiento(90% del corpus para entrenar y el 10% restante para evaluar).
Para mejorar el rendimiento practico del etiquetador para textos de la web, el fragmento del corpus para el aprendizaje fue ampliado para que cada oración tuviese un equivalente si acentos gráficos. El análisis sintáctico de dependencias se realizo con el algoritmo Nivre arc-eager(Nivre, 2008) generado con Malt-Parser mediante AA a partir del corpus Ancora.

Un problema a tener en cuenta, fue solventar la mayor importancia que suelen tener las oraciones finales de una opinión. Para modelar esto, se opto por aumentar 75% la OS de las 3 últimas frases de una crítica.
Otro problema, de la tendencia positiva del lenguaje, es expresar una opinión negativa usando negaciones de términos positivos, "no barato", "no bueno", para compensar esta desviación muchas aproximaciones léxicas incrementan la OS de los términos negativos. Se mejoro la presición del sistema aumentando la dispenrsión de la OS un 20%, haciendo que las polaridades sean entre -6 y 6.

Para la evaluación se empleo un corpus formado por 400 documentos; SFU Spanish Review Corpus(Brooke, Tofiloski, Taboada, 2009). Contiene reseñas de productos y servicios de ocho categorías: lavadoras, hoteles, películas, coches, ordenadores, libros, música y moviles.
Cada categoría dispone de 50 documentos, 25 expresan una opinión positiva y 25 una negativa. Se procesa cada texto y se obtiene la OS, si es mayr a 0 es positiva y es negativa en caso contrario.

Las construcciones lingüisticas tratadas han mejorado el rendimiento, el incremento obtenido al incorporar la negación es muy significativo.

Propuesta	Precisión
Base	0,618
+ intensificación	0,660
+ adversativas	0,670
+ negación	0,755*
Final	0,785*

Se realizarón pruebas chi-cuadrado($p < 0.01$), comparando con las polaridades correctas.

Haber utilizado el mismo corpus y los mismos diccionarios semánticos, permite comparar esta alternativa sintáctica con la solución léxica.
Hubo un incremento en el rendimiento respecto a "The Spanish SO-CA", también se creo un clasificador SVM, basado en AA, usando WEKA.

Método	Precisión (%)
Nuestra propuesta	**78,50**
The Spanish SO-CAL	74,25
SVM	72,50