NaturalOpinions: NLP-based opinion extraction in user-generated content

NaturalOpinions: extracción de opinión basada en PLN para contenidos generados por usuarios

Antonio S. Valderrábanos
BITEXT
Cólquide 6, Las Rozas, Madrid, Spain
asv@bitext.com

Enrique Torrejón
BITEXT
Cólquide 6, Las Rozas, Madrid, Spain
etorrejon@bitext.com

Resumen: Cada vez más personas acceden a Internet y cada vez es mayor el contenido generado por los mismos. La necesidad de saber lo que escriben estas personas en blogs, foros y medios sociales en general sobre marcas y productos específicos tiene una importancia estratégica para grandes empresas en todo el mundo. Presentaremos una demostración de una aplicación, NaturalOpinions, que sigue un enfoque de PLN basado en reglas para analizar opiniones en Twitter en español. La aplicación puede detectar el tema del que se opina y extraer el sentimiento, ya sea positivo o negativo, sobre nombres de marcas o características de productos concretos. De este modo, se puede implementar rápidamente con la tecnología lingüística de Bitext soluciones de análisis de medios sociales.

Palabras clave: Extracción de información, análisis de opiniones, clasificación de sentimientos, contenidos generados por usuarios, procesamiento del lenguaje natural, inteligencia de marcas.

Abstract: More and more people have access to Internet and the content they produce keeps growing. The need to know what people are writing in blogs, forums and social media in general about specific brands and products has become strategically important for large corporations all around the globe. We will present the demo of an application, NaturalOpinions, which follows a rule-based NLP approach to parsing opinion in Twitter in Spanish. The application is able to detect the topic and extract the sentiment, either positive or negative, about particular product features or brand names. Social media intelligence solutions can thus be implemented rapidly with Bitext's language technologies.

Keywords: Information extraction, opinion analysis, sentiment classification, user-generated content, Twitter, NLP, brand intelligence.

1 Introduction

Brand intelligence tools are becoming a strategic asset for companies interested in keeping track of users' opinions and sentiments in user-generated content, most significantly in blogs, forums, Twitter, Facebook, etc. Those tools have traditionally adopted statistics-based approaches along the lines of standard text mining techniques (Popescu and Etzioni, 2005).

We claim that the tasks of opinion extraction and polarity determination require a more sophisticated approach where mainly rule-based NLP software, furnished with syntactic and semantic processing, can provide the fine-grained analysis needed for reliable reports on opinion and sentiment classification. Bitext's focus therefore continues the path of using dependencies grammars and full semantic representation (Dini and Mazzini, 2002) to tackle the challenge of opinion extraction. We will demo an application of NaturalOpinions which analyzes opinions about main brands in Twitter in Spanish.

2 User-generated content

User-generated content continues to grow at an amazing pace. We will consider three types of user-generated content: blogs, Facebook, and Twitter. Regarding blogs, in August 2010 the number of blogs doubles every 5 months and there was a new blog published every second. In early 2009, there were approximately 200 million blogs in English. In August 2010, the number of active blogs in Spanish was 400 million. If we take into account other languages, there were 1 billion active blogs.

If we consider Facebook, in February 2010, there were 400 million active users; on July 21st, 2010, there were 500 million active users. Every user writes an average of 25 comments per month; more than 35 million users update their status every day and more than 60 million updates are made every day. Just to keep track of all the status update and detect opinions in them is a considerable task for which search engines such as Booshaka (www.booshaka.com) can be very useful.

If we focus on Twitter, there are more than 105 million registered users and 300,000 new users register every day. There are approximately 80 million tweets every day. In August 2010, Twitter has accumulated 20 billion tweets. In early September, it surpassed the barrier of 23 billion tweets. And it is estimated that by the end of 2010 Twitter will have accumulated 30 billion tweets. Just to process daily tweets in search for opinions about brands is a considerable task.

Spain is among the 10 countries with greater number of registers in Twitter. It represents 1.7% of the users, that is, over 2 million users. According to the study "Uso de Twitter en España", from the Asociación Española para la Economía Digital, 63% of users recommend products in Twitter, 61% share complaints about products or services, and a 94% follow specific companies in Twitter. These percentages mean that Twitter is a source of social opinions which cannot be ignored in brand intelligence.

3 Opinions in Twitter: social media writing style

Tweets usually have their own writing style features which they share with other user-generated content repositories such as Facebook, blogs, forus, or even SMS. Any NLP-based opinion analysis software for Twitter must handle the following style features:

1. 140-character limit: tweets as microblogs have a limit of 140 characters and, therefore, users need to restrain themselves from writing full-fledged opinions and condensed them as much as possible, for example:
   a. La calidad de Spotify mobile desde el Ipod Touch, Ipad o Nexus One es genial. Mucho mejor que desde el MacBook
   b. @hayleytheone no es nada bonita la coca-cola sin curvas xdd parece que es la de marca blanca xdd a mi no me gusta la coca-cola xdd
2. Use of tokens: tweets contain special tokens such as @ for user names, # for trending topics; they also have http links for related content; for example:
   a. @ranablue T recomiendo hablar sobre el iPAD a @fotomaf y @cuasante
   b. Estoy muy contento con mi #Kindle
   c. esta es la direccion si quieren bajar la aplicacion para iPhone, iPod touch, y iPad en el es gratuita http://goo.gl/aia1
3. Lack of punctuation marks: in order to save characters, users write tweets without punctuation marks, which makes parsing even harder without

sentences/phrases separation marks; for example:

   a. @hayleytheone no es nada bonita la coca-cola sin curvas xdd parece que es la de marca blanca xdd a mi no me gusta la coca-cola xdd
   b. Eso nunca se sabe jeje @AntonioPamos @cosechadel66 Pero el IPAD tiene más aplicaciones que la Sra. Obama

4. Relaxation in the use of accented characters (vowels): this makes morphological processing difficult and opinion analysis software may have to include spell-checking; for example:
   a. creo que movistar jode las conexiones a proposito
   b. iphone es muchisimo mas caro y tactil...
   c. dios que miticas las canciones de los juegos de nintendo
   d. los mensajes automatizados de bienvenida en twitter me dan mala sensación
5. Spelling errors: most tweets are written with spelling errors given the careless and colloquial writing style which is preponderant in Twitter. Therefore, including a spell-checker with the opinion analysis software must be considered. For example:
   a. qué tío más pesado el chabal de coca-cola
   b. que esquisito este twitter
6. SMS style: tweets have also adopted the emoticons, abbreviations, slang, etc, which is typical of SMS. Opinion analysis software must consider whether to include them in the parsing process if they add sentiment information to the opinion conveyed in the tweets. For example, emoticons such as ☺ ;-) ^^ ¬¬ xD xDD etc. :
   a. @Shinfu No está mal el iPhone 4 ¿verdad? ;-)
   b. @Maria__Lourdes ^^ xro el ipod tiene un lado malo... es un enganxe OO
   c. @yeyustyle jajaja un blog? pues uno tuyo personal, no va mal =P mi ipod y yo... y X horas de viaje.. ¬¬ como un dia vaya os arrepentireis xD
   d. parece ke co echofon si ke me funciona en el ipod
   e. @miguelrtorija Yo tengo flickr, molaaaaaaa!! XDD yo también pensaba comprarme un blackberry o algo jajaja el nokia no mola para twittear XD
7. Colloquial style: tweets also feature a colloquial style which can be described as "I write the way I speak"; this includes all the phrasing, swear words, chopped words, etc which are typical of spoken language; for example:
   a. @asturking pos es muy fácil vincular el iphone con un macbook
   b. @guarroman: Flipando con el iPhone de @vego" // si es que mi iPhone es el mejor!
   c. Partidita al Parchis en el iPad con mi hermana, sobrina y mi cuñado. Como mola el iPad leches, aquí el Parchis http://yfrog.com/0m5dxdj
   d. Y joder, por la puta ballena pensaba que el iPod iba mal, su puta vida...
   e. @Natychan Entre eso, el iPhone 4 y que la pantalla de mi iPod Nano está cascando de mala manera
   f. el twitter ta mal me esta borrando los tweets
8. Space of creativity and humor: tweets also contain a great deal of humorous expressions that show the creativity user like to indulge in to make their comments more brilliant and, therefore, retweetable. Opinion analysis software has to continuously update their sentiment lexicons in order to be able to evaluate these tweets correctly; for example:
   a. rt @jakarrion a quien buen apple se arrima buen iphone le cobija #variantes
   b. este twitter falla más que una escopeta de feria
   c. el iphone 4 es la cosa más bonita que ha parío mare
   d. @aletshe @ferr_kon las penas con iphone son menos penas xd
   e. Oh dios. Mi ipod está tan jodidamente brillante, y suave, que voy a orgasmizar!
   f. dios la peña esta archienganchada al twitter

4 Overview of NaturalOpinions

NaturalOpinions for Twitter has been developed using Bitext's proprietary NLP technology. NaturalOpinions consists of three main components:

1. DataSuite NLP engine, which includes:
    a. DataLexica: a component with over 3 million Spanish words morphologically classified and used for POS-tagging.
    b. DataGrammar: a syntactic parser which uses a specifically-designed dependency grammar for opinion analysis. This parser can return both a complete syntactic tree and a shallow parsed tree when the tweet is not grammatically correct. The parser establishes the dependency structure from which brand features and opinions (either positive or negative) can be identified.
2. Semantic extraction component: this component takes care of parsing the syntactic tree and extracting relevant information for opinion analysis, namely
    a. Brand or product name about which the opinion is expressed, for example, "iPad", "iPhone", etc.
    b. Brand component/feature about which the opinion is expressed, for instance, "the screen", "the battery", etc.
    c. Brand component/feature attributes, which allow for a topic classification of opinions. These attributes may be customized according to the domain. Currently, attributes include General, Product, Service, Image, Quality, and Price.
    d. Semantic polarity: whether the opinion is an affirmative or negative statement.
    e. Comparative opinion: whether there is a comparison of two or more brands with detection of topical brand and compared-against brand; likewise, with detection of topical feature and compared-against feature.
    f. Opinion itself: part of the parsed sentence which contains the user's assessment of the brand/feature and includes the words expressing value.
3. Scoring component: this component takes into account the opinion and the semantic polarity returned by the previous component and calculates a score (integer with two decimal values) which measures the strength of the conveyed opinion. The scoring component consists of:
    a. Sentiment dictionaries for the following part of speech: nouns, adjectives, verbs, adverbs, and determiners. Also, there is a sentiment dictionary of features.
    b. Scoring algorithm: the algorithm takes into account the values in the sentiment dictionaries. These values include:
        i. Adding value: values which are summed in the algorithm process, for instance for summing values of adjectives
        ii. Multiplier value: values which are multiplied in the algorithm process, for instance for adverbs which modify adjectives
        iii. Hue value: values which have to do with the intrinsic meaning of the words; for instance, "gorgeous" and "horrible" have an intrinsic absolute value (positive hue and negative hue respectively), whereas "cheap" has a relative value (neutral hue) which may turn out to be positive or negative depending on the context, for instance "cheap price" (positive) versus "cheap material" (negative).

NaturalOpinions for Twitter also includes a graphical dashboard with visualization of tweets according to brands, features, attributes, opinion polarity, time stamps, opinion holders, among others. It can be accessed on Bitext's website www.bitext.com.

References

Dini, I., and Mazzini, G. 2002. Opinion classification Through information extraction. In *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and other fields,* páginas 299-310.

Popescu, A. M. and Mazzini, G. 2005. Extracting Product Features and Opinions from Reviews. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/ EMNLP),* pages 339-346.