SPANISH SENTISTRENGTH AS A TOOL FOR OPINION MINING PERUVIAN FACEBOOK AND TWITTER

Roque López, Javier Tejada, MikeThelwall

***Abstract**: SentiStrength is a well-known tool for opinion mining texts circulated in social networks. There are English, Spanish, Russian versions of SentiStrength, which demonstrated their advantages in numerous examples. In the paper we present our classifier of opinions for analysis of comments from the Peruvian Facebook and Twitter. We corrected and enriched SentiStrength Spanish vocabularies having in view the properties of Peruvian Spanish. We also slightly modified the SentiStrength algorithm rules. We demonstrate the results we have reached with our classifier and SentiStrength. The experimental results are competitive with traditional approaches developed in this area, and the total accuracy proved to be 72%.*

***Keywords**: SentiStrength, Opinion Mining, Facebook and Twitter*

Introduction

Facebook and Twitter have become huge repositories of information. Around 400 million tweets [CNET, http] and more than 2.7 billion comments on Facebook [CNBC, http] are generated every day. Due to the popularity of these social networks, more and more users leave comments expressing their experiences with products or commercial services. This information is of much interest to business, because they would know their strengths, and, most importantly, their weaknesses.

Due to the large volume of data, a manual analysis of this information is an almost impossible problem. For this reason, in recent years, several methods of NLP have been developed to extract and identify subjective information in texts. This area is known as sentiment analysis or opinion mining. Opinion mining aims to extract attributes and components of a text to determine whether comments are positive, negative or neutral [Pang, 2008].

SentiStrength is a tool that has demonstrated good results in opinion mining for the social web [Thelwall, 2010]. SentiStrength estimates the strength of positive and negative sentiments in short texts, even for informal language. SentiStrength uses a dictionary of sentiment words, each one associated with a weight, which is its sentiment strength. In addition, this method uses some rules for non-standard grammar.

This research presents an opinion mining tool based on SentiStrength for analysis of comments from the Peruvian Facebook and Twitter (Spanish). The rest of the paper is organized as follows. In section 2 we explain the SentiStrength algorithm. In section 3 we describe the proposed method for opinions classification. Section 4 contains the test data used in this work, as well as, the results obtained in experiments. Finally in section 5 we present our conclusions.

SentiStrength

SentiStrength is a method oriented to detect sentiment in informal texts [Thelwall, 2010]. For this reason, in addition to a sentimental word dictionary it takes into account the most common spelling styles in social networks (e.g. "h8" and "hate" have the same meaning). This algorithm uses two scales, from 1 to 5 and from -1 to -5, to

classify a text. SentiStrength evaluates the contribution of positive and negative sentiments separately and makes a decision based on their values.

The main resources used in this algorithm are [Thelwall, 2012]:

- A sentimental word list, this is a collection of 298 positive terms and 465 negative terms. Each word has a value from 2 to 5, if positive, or -2 to -5, if negative.
- A spelling correction algorithm for English language.
- A booster word list is used to strengthen or weaken the emotion of a list of sentiment words.
- An idiom list is used to identify the sentiment of a few common phrases.
- A negating word list is used to invert emotion words (skipping any intervening booster words).
- An emoticon list, which has an associated sentimental weight.
- Sentences with exclamation marks have a minimum weight.
- Stemmers are not used

The proposed classifier

The proposed classifier uses a lexicon-based approach [Taboada, 2011]. We consider 3 categories: positive, negative and neutral. This algorithm uses a scale from 0 to 1 to classify a text. Our classifier uses the SentiStrength resources described in the previous section. However, unlike SentiStrength, we consider some linguistic particularities used in Peruvian comments on Facebook and Twitter. We consider slangs, abbreviations and Peruvian combinations

Also, we use a parameterization module to identify: sentimental words, booster words, negating words, emoticons, exclamation marks, idiom words and Peruvian slang. To classify a comment as positive, negative or neutral we take into account 3 types of contributions: words, combination of words and emoticons contributions. The classifier evaluates the contribution of positive and negative sentiments and makes a decision based on their values.. Figure 1 shows the contents of classifier.

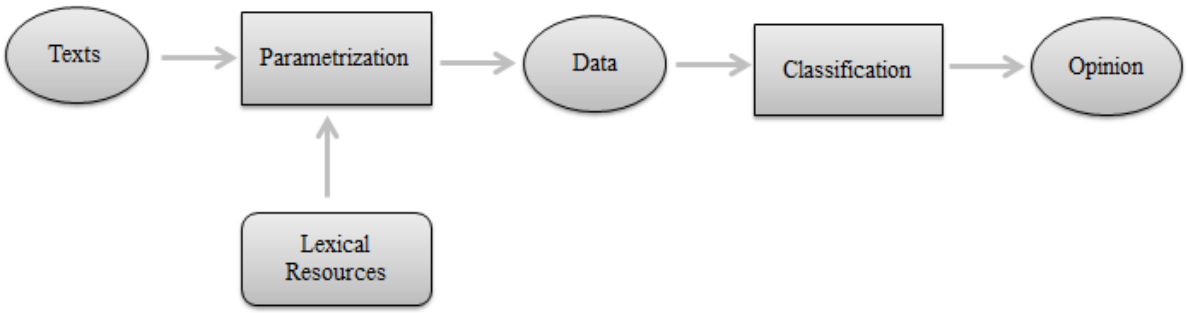

Fig. 1. Contents of opinion classifier

Experiments

Document set

We realize experiments to evaluate the quality of classifier performance. We also compared it with SentiStrength Itself. For the experiments we used a collection of 282 comments from Facebook and Twitter, which were collected from various Peruvian Internet pages. The collection was created manually and every comment was

classified as positive, negative or neutral by 3 experts. The coincidence of classification of the 282 comments by experts was 82.5%. Table 1 shows the composition of this collection.

Table 1. Composition of comments collection

Category	Facebook	Twitter
Telecommunication	21	21
Fast food chain	22	22
Cinema	19	19
Bank	20	20
Chocolate shops	21	--
Filling station	20	--
Gastronomic Festival	11	--
Public personalities	16	--
Life Insurance	15	--
Soda	17	--
Airlines	18	--
Total	**200**	**82**

Results

Table 2 shows positive, negative, neutral and total accuracy obtained by SentiStrength and our proposed classifier for Facebook and Twitter. In both cases the classifier overcomes standard SentiStrength. The reason is: the classifier is specially adjusted for Peruvian texts. The results for comments on Facebook are better than on Twitter. The reason is: comments on Twitter are shorter (at most 140 characters). The best neutral accuracy proves to be reached with Twitter comments, because the vast majority of comments in Twitter do not express sentiment.

Table 2. Results on Facebook and Twitter

	Facebook		Twitter	
Accuracy	**SentiStrength**	**Classifier**	**SentiStrength**	**Classifier**
Positive Accuracy	85.37%	86.90%	52.17%	34.78%
Negative Accuracy	62.34%	83.33%	51.43%	62.86%
Neutral Accuracy	26.83%	15.79%	29.17%	62.50%
Total Accuracy	**64.5%**	**72.0%**	**45.12%**	**54.88%**

Conclusion

The main results of the paper are:

- We proposed a classifier which uses 3 types of contributions: words, word combinations and emoticons. This classifier uses lexical resources of standard Spanish SentiStrength enriched by Peruvian lexis.
- The experiments demonstrate the essentially better results than the standard SentiStrength provides.

In the future we suppose:

- to use more categories of classification, such as very positive and very negative
- to improve the grammatical rules used

Bibliography

[Balahur, 2009] A. Balahur, Z. Kozareva, A. Montoyo. Determining the polarity and source of opinions expressed in political debates Lecture Notes in Computer Science, 5449, 468-480 (2009).

[CNBC, http] http://www.cnbc.com/id/45582325/The_World_s_Most_Liked_Brands

[CNET, http] http://news.cnet.com/8301-1023_3-57448388-93/twitter-hits-400-million-tweets-per-day-mostly-mobile/

[Pang, 2008] B. Pang, L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 1 (1-2), 1-135 (2008).

[Taboada, 2011] M.Taboada, J. Brooke, M.Tofiloski, K. Voll, M. Stede. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267-307 (2011).

[Thelwall, 2010] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai. Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology. 61, 2544–2558 (2010).

[Thelwall, 2012] M. Thelwall, K. Buckley, G. Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology. 63, 163–173 (2012).

Authors' Information

Roque López *– Student of System Engeenering at San Agustin National University, calle Santa Catalina Nº 117 Arequipa, Peru; e-mail: rlopezc27@gmai.lcom*

Major Fields of Scientific Research: natural language processing, text mining, social network analysis

Javier Tejada Cárcamo *– Professor of Computer Science Department, San Pablo Catholic University; Research and Software Development Center of San Agustin National University (Cátedra Concytec); e-mail:* jawitejada@hotmail.com

Major Fields of Scientific Research: natural language processing, word space models, business intelligence

Mike Thelwall *– Professor, Department of Information Studies-Abo Akademi University. Research Associate, Oxford Internet Institute-Oxford University; e-mail:* m.thelwall@wlv.ac.uk

Major Fields of Scientific Research: Cybermetrics, scientometrics, information science and social metadata (tagging, folksonomy)