PARAMETERIZATION OF COMMENTS FROM PERUVIAN FACEBOOK AND TWITTER: LEXICAL RESOURCES AND ALGORITHM

Angels Catena, Mikhail Alexandrov, Roque López

Abstract: *Millions of Facebook and Twitter users send their comments all over the world about products and services, political and economical events, etc. (almost 3 billions each day) The principal problem of opinion mining such information is text parameterization, and in the paper we describe our experience in solution of this problem with Peruvian Facebook and Twitter. We use enriched vocabularies of Spanish SentiStrength and propose a simple algorithm for evaluation of sentiment contribution. The work was completed in the framework of the project WAYRA (Telefonica-Peru). The results proved to be promised: opinion analysis of parameterized texts showed the accuracy about 75% with elementary classifier*

Keywords: *opinion mining, SentiStrength, Facebook and Twitter*

ACM Classification Keywords: *I.2.7 Natural Language Processing*

Introduction

Social Networks became an important source of information for very different applications. Around 400 million tweets [CNET, http] and more than 2.7 billion comments on Facebook [CNBC, http] are generated every day. People want to share their opinions about products and services, economical situation and political events and this feedback proves to be very useful for those who offer these products or who are responsible for these events.

The first NLP tools for Opinion Mining dealt with so-called 'ordinary' documents. Speaking 'ordinary' we mean not very short documents with usual normative lexis. As an example we can mention here the well-known program SO CAL. This program was applied for processing messages on forums, where the quality of various products and services were discussed. It provided the average accuracy 80% for binary classification of opinions [Taboada, 2006]. The other example is the program Opininer-UAB. We used this program for processing analytical texts in electronic publications concerning economical crisis in Europe. Opininer-UAB showed the accuracy 75%-80% for binary classification and about 60% for more detail classification with four categories [Catena, 2010].

But we meet the absolutely other situation when we have to work with comments from Facebook and Twitter. They are super-short (usually 2-5 phrases) and they use non-normative lexis (abbreviations, slang, emotional punctuation, etc.). For this reason it is difficult to reach the accuracy we pointed above. The SentiStrength program is one of the best tools we know for Opinion Mining texts from Facebook and Twitter [Thelwall, 2010; Thelwall, 2012]. By the moment there are versions of SentiStrength in English, German, French, Italian, Spanish, and some other languages. The usual Spanish SentiStrength version provides the accuracy 65% and its last modification especially oriented on the Peruvian Facebook and Twitter provides the accuracy 74% [Lopez, 2012]

This paper is the continuation of the previous work. It was completed in the framework of the ambitious project WAYRA by the company Telefonica-Peru [WAYRA-a, http; WAYRA-b, http]. In this work we concern only the process of text parameterization, which have the decisive role in Opinion Mining texts from social networks. The program is written on Python. The document collection includes 200 comments.

In the section 2 we describe the linguistic resources. Section 3 demonstrates step by step the work of algorithm. The results of experiments are shortly presented in section 4. Section 5 contains the conclusions.

Linguistic resources

Beta version of our program includes 6 vocabularies. The examples below present English translation of Spanish words and word combinations. The values in brackets mean weights or marks. We use 4 values for the scale of sentiment weights {-2, -1, 1, 2}. Here are these vocabularies:

1) Vocabulary-W (words). It is stop words: pronouns, prepositions and articles
2) Vocabulary-P (palabras = words, Spanish). It is single words presented in shorten flexible form with their weights. The vocabulary contains normative lexis and slang. Some examples are: excellen# [2], better[1], problem# [-1], terrible [-2]. Here '#' means possible letters.
3) Vocabulary-C (combinations). It is fixed word combinations with their weights. The vocabulary contains normative lexis and slang. Some examples are: nothing bad [1], it is worth [1], never one can [-1], nothing good [-1]
4) Vocabulary-N (negations). It is patterns with marks. Some examples are: nobody [2], no [1]. Here mark 2 means that a word related with this pattern changes its polarity and its absolute weight doubles; mark 1 means that a word related with this pattern only changes its polarity.
5) Vocabulary-R (realces = emphatic constructions, Spanish). It is patterns with marks. Some examples are: super [2], ultra [2], more [1], less [-1]. Here: marks mean increasing or decreasing the value of an adjacent word respectively.
6) Vocabulary-S (signs). It is patterns with weights. The vocabulary contains emoticons and punctuation. Some examples are: :-)[1], :-([-1], !![0], ?[-1], ??[-2]. Here the elements with the weight '0' increase the value of previous sentiment over its basic value. Let we have two sentiments: good [1] and terrible [-2]. According this rule the total weight of 'good!!' is equal (1+1)=2, and the total weight of 'terrible!!' is equal (-2-1)=-3.

We are not sure whether it is worth to use our scale of weights {-2, -1, 1, 2} instead of the rude one {-1,1}. Such a question needs the additional experiments. In particularly, the paper [Kaurova, 2010] demonstrates examples where we lose only 2%-4% of accuracy with the scale {-1,1} instead of the detail scale {-5,-4,…4,5}.

Algorithm

3.1 Preprocessing

On this stage we introduce four arrays: **Comment, Text, Mark-1, Mark-2.** To demonstrate how the algorithm works we consider an example:

- no noa no me gustaa este producto muya malooo :-(:-(vale la pena olvidarlo (incorrected text in Spanish)

- no I don't like this very bad product it is worth to forget it (text translated to English after correction)

1) **Comment**. It is a comment presented in the lineal form without the division on phrases. Each cell contains one word or one sign before correction

No	noa	no	me	gustaa	este	producta	muya	maloo	;-(	;-(	vale	la
pena	olvidarlo											

2) **Text**. It is a comment presented in the lineal form without the division on phrases. Each cell contains one word or one sign after correction

N	no	n	m	g	est	pro	m	m	;-	;-	v	L
o		o	e	usta	e	ducto	uy	alo	(	(	ale	a
P	olv											
ena	idarlo											

Having eliminated the repeated elements we have a new list of words and signs

n	m	g	e	pro	m	m	;-	v	l	p	olv
o	e	usta	ste	ducto	uy	alo	(	ale	a	ena	idarlo

Having eliminated stop-words we have the other list. But stop-words being the part of fixed word combinations are not considered. See here: vale **la** pena (= it is worth).

	gu	pro	m	m	;	v	l	p	olv
no	sta	ducto	uy	alo	-(	ale	a	ena	idarlo

3) **Mark-1.** It is the set of denominations, which reflect the type of elements. We use the following list of denominations:

- P is a word from the Vocabulary-P
- C is a word combination from the Vocabulary-C
- R is an emphatic construction from the Vocabulary-R
- N is a negation from the Vocabulary-N
- S is a sign from the Vocabulary-S
- Z means that an element has no any weight

Therefore for our example we have:

N	P	Z	R	P	S	C	Z

4) **Mark-2.** This array is directly related with the array **Mark-1**. Namely, here each cell contains the marks and weight of a correspondent element from the **Mark-1.**

1	1	0	1	-	-	1	0
				1	1		

These data are taken from the vocabularies.

3.2 Processing

On this stage we use two arrays: **Indicator** and **Weight**

5) **Indicator.** It is a set of indicators, which show each moment whether the correspondent element of text has been evaluated or no. The program evaluates a given text from left to right revealing patterns according the list of rules (we mention them later). The elements are marked with '1' step by step.

6) **Weight.** It is the array of points, each pattern of a comment contributes to the total assessment. Here the array **Mark-2** is used**.** The array **Weight** changes its contents having revealed each pattern. The points are written to the first cell this pattern occupies

Initially we have:

Mark-1

N	P	Z	R	P	S	C	Z

Mark-2

1	1	0	1	-1	-1	1	0

Indicator

0	0	0	0	0	0	0	0

Weight

0	0	0	0	0	0	0	0

After the first step:

Mark-1

N	P	Z	R	P	S	C	Z

Mark-2

1	1	0	1	-1	-1	1	0

Indicator

1	1	1	0	0	0	0	0

Weight

-1	0	0	0	0	0	0	0

After the second step:

Mark-1

N	P	Z	R	P	S	C	Z

Mark-2

1	1	0	1	-1	-1	1	0

Indicator

1	1	1	1	1	0	0	0

Weight

-1	0	0	-2	0	0	0	0

After the third step:

Mark-1

N	P	Z	R	P	S	C	Z

Mark-2

1	1	0	1	-1	-1	1	0

Indicator

1	1	1	1	1	1	0	0

Weight

-1	0	0	-2	0	-1	0	0

After the fourth step:

Mark-1

N	P	Z	R	P	S	C	Z

Mark-2

1	1	0	1	-1	-1	1	0

Indicator

1	1	1	1	1	1	1	1

Weight

-1	0	0	-2	0	-1	1	0

Then all positive and negative scores are summed. The Table 1 shows the final result of text processing

Table 1. Results of comment processing

Positive scores	***Number of positive sentiments***	***Negative scores***	***Number of negative sentiments***
1	1	-4	3

In this example we indirectly used some rules for processing sequences NPZ, RPZ, etc. Here the symbols ‘N’, “P’, ‘Z’ mean the types of elements in a phrase under consideration (see section 3.1). This moment the program includes 6 different rules for processing phrases.

Experiments

4.1 Elementary classifier

To demonstrate the effectiveness of proposed algorithm we used 200 comments from the Peruvian Facebook and Twitter. All comments were evaluated by 3 experts using the scale: positive, neutral, negative. Then these comments were parameterized: the part of texts (150) was used for constructing classifier, and the other part (50) for testing.

The elementary threshold-based classifier is formed by the following way

- All parameterized comments are presented in the scale

  r = (*PosScore* + *NegScore*) / (*PosScore*+| *NegScore* |). Here: *PosScore* and *NegScore* are contributions of positive and negative sentiments respectively. Obviously, $|r| \leq 1$
- Histogram of the learning set (150 comments in our case) is constructed on the basis of variable r
- Thresholds for decision-making are manually adjusted to provide the maximum accuracy

Of course, we do not pretend here to have any sophisticated classifier. The modern approaches for opinion mining are known and these approaches have been already described in detail in the publications [Pang, 2008; Taboada, 2011]. The only we want is to show that our algorithm of parameterization allows to obtain good results even with the simplest classifier.

4.2 Results of classification

The Table 2 shows some results with the classification of testing set (50 comments)

Table 2. Results of classification

Categories and the number of objects	***Thresholds***	***Rules***	***Accuracy***
Positive (42%) Neutral (39%) Negative (19%)	Up = 0,05 Un = -0,05	*I* ≥ Up Positive *I* < Un Negative The other ones are Neutral	72%
Positive (42%) Indefined - Negative (19%) The neutral category is not considered	Up = 0,2 Un = -0,2	*I* ≥ Up Positive *I* < Un Negative The other ones are Indefined	77%

Notes.

1) We have to introduce the so-called undefined category to improve the quality of results: sometimes is better to say “I do not know” instead of any erroneous answer
2) We suppose the results will be essentially better if to use any more advanced classifier instead of the elementary one

Conclusions

The main results of the paper are:

- We proposed method (linguistic resources and algorithm) for parameterization of comments from Facebook and Twitter.
- Experiments with real Peruvian Facebook and Twitter showed the promised results
- The method with modifications can be used for processing comments on other languages

In the future we suppose:

- to extend the list of grammatical rules
- to test the binary scale for sentiment classification

Bibliography

[Catena, 2010] A. Catena, M. Alexandrov, N. Ponomareva. Opinion Analysis of Publications on Economics with Limited Vocabulary of Sentiments the Intern. Journal on Social Media MMM: Monitoring, Measurement, and Mining, Brno, Czech Rep., N_1, 2010, Publ. House 'Konvoj' , pp. 20-31

[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/

[Kaurova, 2010] O. Kaurova, M. Alexandrov, N. Ponomareva. The Study of sentiment word granularity for Opinion Analysis (a comparison with Maite Taboada works). Intern. Journal on Social Media MMM: Monitoring, Measurement, and Mining, Brno, Czech Rep.,N_1, 2010, Publ. House 'Konvoj' , pp.45-57

[Lopez, 2012] R. Lopez, J. Tejada, M.Thelwall. Spanish Sentistrength as a tool for opinion mining Peruvian Facebook and Twitter. (this Proceedings)

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

[Taboada, 2006] M. Taboada, C. Anthony, K. Voll. Opinion mining and sentiment analysis. Foundations and trends in Information Retrieval, 1(1-2), 1-135 (2008). Creating semantic orientation dictionaries. *Proc. of 5th Intern. Conf. on Language Resources and Evaluation (LREC)*. Italy, 2006, p. 427-432.

[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 Inform. Science and Technology. 63, 163–173 (2012).

[WAYRA-a, http] WAYRA: http://innovar.org/?p=772

[WAYRA-b, http] WAYRA: http://www.youtube.com/watch?v=Kc1ho6GLQJU

Authors' Information

Angels Catena – *Professor, Department of French and Romance Philology, fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: Angels.Catena@gmail.com

Major Fields of Scientific Research: sentiment analysis, French lexis and semantics sentiment analysis,

Mikhail Alexandrov – *Professor of the Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; research fellow of the fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: MAlexandrov@mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

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