An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis

Pedro BALAGE FILHO, Sandra ALUÍSIO, Thiago PARDO

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
Sao Carlos - SP, Brazil

{balage,taspardo,sandra}@icmc.usp.br

Abstract

This work presents an evaluation of the Brazilian Portuguese LIWC dictionary for Sentiment Analysis. This evaluation is conducted by comparison against two other sentiment resources for Portuguese language: Opinion Lexicon and SentiLex. We conducted an intrinsic and an extrinsic evaluations and show how LIWC dictionary could be used in sentiment analysis projects.

Motivation

Linguistic Inquiry and Word Count (LIWC) is a text analysis software that calculates the degree of use for different categories of words across a wide array of texts (Pennebaker et al., 2001).

The core of this program is a lexicon resource, best known as LIWC dictionary, which recently has been made available for Portuguese Language . The resource was kindly provided by the researchers: Profa. Rove Chishman (Unisinos), Profa. Sandra Maria Aluísio (ICMC-USP) and Rosângela Lopes Toledo Checchia (Checon Pesquisa).

Table1. Brazilian Portuguese LIWC categories and number of entries

Category	Number of entries	Category	Number of entries	Category	Number of entries
achieve	9865	future	268	preps	69
adverb	139	health	7003	present	4715
affect	28475	hear	3045	pronoun	128
anger	6867	home	2019	quant	622
anx	3012	humans	22258	relativ	24965
article	10	i	7	relig	2066
assent	58	incl	3071	sad	3864
auxverb	1445	ingest	11085	see	4634
bio	17858	inhib	13031	sexual	1819
body	4766	insight	18683	shehe	16
cause	11770	ipron	88	social	13632
certain	3428	leisure	6331	space	5313
cogmech	46307	money	5352	swear	14041
conj	27	motion	13641	tentat	5719
death	2429	negate	21	they	11
discrep	2943	**negemo**	15115	time	7324
excl	483	nonfl	14	verb	23873
family	96	number	83	we	8
feel	7727	past	7684	work	7735
filler	12	percept	17607	you	25
friend	679	**posemo**	12878		
funct	5512	ppron	54		

Sentiment analysis, or opinion mining, is a relatively new topic of research in natural language processing that has gained lots of attention due to the growth of the social web. The sentiment analysis approach based on lexicon uses it to provide the polarity, or semantic orientation, for each word or phrase in the text.

This work aims to evaluate the new LIWC lexicon against two other sentiment resources available for Portuguese: the Opinion Lexicon (Souza et al., 2011) and the SentiLex (Silva et al., 2012).

Lexicons and Data Normalization

Table2. Examples from the lexicons

Word	LIWC	Opinion Lexicon	SentiLex
admirar (to admire)	-	admirar,vb,1	admirar.PoS=V;POL:N0=0;POL:N1=1
alegre (joyful)	posemo	alegre,adj,1	alegre.PoS=Adj;POL:N0=1
alto (high)	-	alto,adj,0	alto. PoS=Adj;POL:N0=0
encorajar (to encourage)	posemo	encorajar,vb,0	encorajar.PoS=V;POL:N0=0;POL:N1=1
famoso (famous)	-	famoso,adj,1	famoso.PoS=AdjPOL:N0=1
inimigo (enemy)	negemo	inimigo,adj,1	inimigo.PoS=Adj;POL:N0=-1
quebrar (to break)	-	quebrar,vb,-1	quebrar.PoS=V;POL:N0=-1

Table3. Lexicons normalized for comparison

Word	LIWC	Opinion Lexicon	SentiLex
admirar (to admire)	neutral	positive	neutral
alegre (joyful)	positive	positive	positive
alto (high)	neutral	neutral	neutral
encorajar (to encourage)	positive	neutral	neutral
famoso (famous)	neutral	positive	positive
inimigo (enemy)	negative	positive	negative
quebrar (to break)	neutral	negative	negative

Intrinsic Evaluation - Lexical Agreement

Table 4. Lexical agreement

Agreement	LIWC	Opinion Lexicon	SentiLex
LIWC dictionary	x	35.07% (of 9,810 entries)	33.03% (of 20,282 entries)
Opinion Lexicon	x	x	93.68% (of 17,087 entries)
SentiLex	x	x	x

Table 5. Lexical agreement for polar words

Agreement	LIWC	Opinion Lexicon	SentiLex
LIWC	x	80.17% (of 1,871 entries)	74.83% (of 7,310 entries)
Opinion Lexicon	x	x	97.04% (of 13,880 entries)
SentiLex	x	x	x

Extrinsic Evaluation - Sentiment Classification

For conducting this evaluation, we choose the ReLi (Freitas et al., 2012), a corpus from a Brazilian social network of book reviews.

The corpus is composed by 2,056 reviews from 13 different books (approximately 200 reviews each) and has 300,000 words. The sentiment annotation is present in the opinion and sentence levels. The corpus has 4,210 positive opinion spans and 1,024 negative opinion spans. In the level of sentence the corpus has 2,883 positive sentences and 596 negative ones.

The algorithm adopted for this task is similar to the SO-CAL described in Taboada et al. (2011).

Table 6. Results for Opinion Classification

Lexicon	Class	Precision	Recall	F-measure	Accuracy
LIWC	Positive	88.93%	58.22%	70.37%	52.02%
	Negative	65.80%	34.51%	45.28%	
Opinion Lexicon	Positive	86.87%	55.42%	67.66%	50.53%
	Negative	58.18%	36.72%	45.02%	
SentiLex	Positive	95.74%	53.85%	68.93%	53.35%
	Negative	71.73%	51.95%	60.25%	

Table 7. Results for Sentence Classification

Lexicon	Class	Precision	Recall	F-measure	Accuracy
LIWC	Positive	86.42%	65.43%	74.48%	57.33%
	Negative	40.06%	22.66%	28.95%	
Opinion Lexicon	Positive	87.85%	50.95%	64.49%	47.42%
	Negative	35.96%	32.35%	34.06%	
SentiLex	Positive	91.67%	43.22%	58.74%	44.17%
	Negative	46.34%	48.26%	47.28%	

Conclusions

This evaluation aims to guide future works in lexicon-based sentiment analysis. We conducted two evaluations: an intrinsic evaluation, by measuring the agreement compared with two other lexicons; and an extrinsic evaluation, by measuring the lexicon impact in a sentiment classification task. All programming code used by this evaluation are available for reproduction of the results at: https://github.com/pedrobalage

References

Freitas, C., Motta, E., Milidiú, R., & Cesar, J. (2012). Vampiro que brilha... rá! Desafios na anotação de opinião em um corpus de resenhas de livros. *Proceedings do XI Encontro de Linguística de Corpus (XI ELC)*. São Carlos - SP.

Silva, M., Carvalho, P. and Sarmento, L. (2012), "Building a Sentiment Lexicon for Social Judgement Mining", *Computational Processing of the Portuguese Language*, Springer, pp. 218–228.

Souza, M., Vieira, R., Chishman, R. and Alves, I. M. (2011), Construction of a Portuguese Opinion Lexicon from multiple resources, in *8th Brazilian Symposium in Information and Human Language Technology – STIL*, Mato Grosso, Brazil.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based Methods for Sentiment Analysis. Computational Linguistics, 37(2):267–307.0

IX STIL

IX Brazilian Symposium in Information and Human Language Technology – STIL – 21 a 22/10/2013
Joint event with 2nd Brazilian Conference on Intelligent Systems (BRACIS-13) - October, 20-24
Fortaleza/Ceará