How do You Feel?
Investigating lexical-syntactic patterns in sentiment expression

Magali Sanches Duran
magali.duran@uol.com.br
Center of Computational Linguistics (NILC), ICMC, University of São Paulo,Brazil

Carlos Ramisch
ceramisch@inf.ufrgs.br
Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
GETALP - LIG, University of Grenoble, France

Abstract. This study investigates how sentiments are expressed in Brazilian Portuguese. Sentiment verbs like *temer*, (fear), *odiar* (hate) and *invejar* (envy) are examples of lexical units specifically used to express the respective feelings. The same meaning may be conveyed through other verbs associated to sentiment nouns. This study firstly identifies seven recurrent patterns of sentiment expression without sentiment verbs and then employs these patterns to identify sentiment nouns associated to them. Analysis of the patterns shows that six of them focus on the sentiment experiencer and one focuses on the sentiment cause. Combining sentiment nouns with the seven patterns may be useful to automatically identify sentiment expression and additionally know who is feeling and who or what is causing the feeling.

Keywords: sentiment lexicon, sentiment analysis, light verb constructions, multiword expressions, corpus-based analysis, Brazilian Portuguese.

1. Introduction

Sentiment analysis and opinion mining are a very growing topic of interest in the last few years due to the large amount of texts produced through web facilities, like social networking, blogs, e-mail and chats. These texts are full of information about what people think and feel, valuable information for marketing and political decisions. However, it is humanly impossible to deal with such increasing amount of data. In order to facilitate human analysis or even substitute it, computer based techniques were required and for this reason sentiment analysis became a challenge to the Natural Language Processing community.

Whatever the strategy used, it is essential to count on a sentiment lexicon. However, even when they contain sentiment words, some utterances are not instances of sentiment expression. In the following sentence, for example, the sentiment noun *fear* is a topic of discourse:

Overcoming fear is a skill that anyone can learn.

In this example, there is nobody that may be identified as feeling fear, as well as nothing that may be identified as causing fear. There is a simple way to avoid this kind of utterances: it is enough to associate morpho-syntactic features to the sentiment lexicon and select only sentiment verbs to search sentiment expression. But this is not a complete solution. Although sentiment verbs are lexical items specifically used to express feelings, they are not the only way to do this.

In Portuguese, it is possible and frequent to express feelings using other verbs associated to sentiment nouns. For example:

João tem inveja de você. (lit. João has envy of you = João envies you)

In this example, the sentiment expressed is "inveja" (envy), "João" is the one who feels envy and "você" is the cause (or stimulus) for "João" feeling envy.

It would be interesting, indeed, that a Portuguese sentiment lexicon includes collocations like "ter inveja", which corresponds to the verb "invejar" (to envy). As well, it is relevant for sentiment data mining to know how to determine who is feeling the expressed sentiment and what is causing the expressed sentiment. Hence, this study aims to explore recurrent patterns used to express feelings in Portuguese, using verbs other than sentiment verbs, in order to provide new lexical syntactic inputs for sentiment analysis.

2. Related Works

A comprehensive review of sentiment analysis and opinion mining as a research field for NLP is presented in [1]. The review provides guidance for those interested in developing opinion mining search engines. The authors address the problem of deciding where to mine opinion and sentiment expression, how to gather information and how to present the information gathered.

Due to the role played by the lexicon in sentiment analysis systems, the NLP related tasks are highly language dependent. An ontological approach, as proposed by [2] and [3] may benefit the semantic description of the sentiment lexicon and pave the way for multilingual approaches.

Besides the identification of sentiment words, there are studies dedicated to enrich the description of these words, aggregating features that enable clustering the gathered information. Up to this date, features regarding sentiment words are almost always related to their polarity, as may be seen in [4], in SentiWordNet [5] and in SentiLexPT[1] (this latter being a lexical resource of Portuguese).

In Portuguese, there are few reported studies related to sentiment analysis [6, 7]. Due to their role in political and marketing decisions, sentiment analysis and opinion mining systems constitute a competitive advantage. This fact encourages private financial support for developing new resources that remain undisclosed.

2. Methodology

This study had five steps. The first one was to identify recurrent lexical-syntactic patterns to express feelings using sentiment nouns instead of sentiment verbs. The second step was to use the patterns identified as search arguments to identify sentiment expression. The third step was the human analysis of the candidate lists resulting from step two. The task was to say whether the noun collocated at the right of each pattern was or not a sentiment noun. In the fourth step we analysed the validated candidates and assigned them some features. In the fifth and last step, we combined the patterns of step one with the sentiment nouns identified in step three and searched

[1] http://dmir.inesc-id.pt/reaction/SentiLex-PT_01

the combinations in the web. The five steps are presented and analysed separately in the following sections.

3. Identifying lexical-syntactic patterns of sentiment expression (STEP 1)

This was the very beginning of our work and was incidentally made during another work that investigated complex predicates [8]. When we analysed light verb constructions candidates, we noticed seven recurrent constructions with sentiment nouns[2]:

1. Sentir * de (to feel * of)
2. Sentir * por (to feel * for);
3. Ter * de (to have * of);
4. Ter * por (to have * for);
5. Ficar com * de (to become with * of);
6. Estar com * de (to be with * of)
7. Dar * em (to give * in).

The analysis of these patterns showed us that all of them have three variables associated with semantic roles [9, 10]:

1) somebody that feels, who will be referred here as "experiencer",
2) the feeling itself, which is referred as "sentiment noun" and
3) the cause or stimulus that causes somebody to feel the feeling, which is referred as "cause".

In patterns 1 to 6, the experiencer takes the subject position and the cause is a verbal complement, as may be observed in the following example:

Eu tenho medo de avião = I have fear of airplanes (lit.)

In pattern 7 we observed the contrary, that is, the subject position is occupied by the cause and the experiencer is a verbal complement:

Avião não dá medo em crianças = Airplanes do not give fear in children (lit.)

These two ways of expressing feelings, focusing the experiencer or the cause, denote a change of point-of-view.

The relevance of these findings for sentiment analysis motivated us to further investigate the seven patterns and verify how generic they are to express feelings.

4. Using the patterns to gather sentiment nouns (STEP 2)

Our aim in this step was to survey sentiment nouns which are expressed through the patterns identified in first step. For this, we used the PLN-BR-FULL corpus (http://www.nilc.icmc.usp.br/plnbr/), consisting of news texts from Folha de São Paulo from 1994 to 2005, with 29.014.089 tokens automatically lemmatised and POS-tagged. Then, the

[2] The character "*" is used in place of the sentiment noun.

patterns were fed into the mwetoolkit [11], a computational system for language-independent identification of multiword expressions in corpora.

The result consisted of seven lists, one for each pattern, with the collocated nouns and their respective frequency in the corpus. The 1.774 candidates are distributed as follows:

PATTERN	CANDIDATES
1. Sentir * de	49
2. Sentir * por	18
3. Ter * de	1218
4. Ter * por	131
5. Ficar com * de	51
6. Estar com * de	92
7. Dar * em	215

Table 1. Candidates per pattern

5. Analysing candidate lists (STEP 3)

The noisy occurrence lists have been carefully analysed by human annotators in order to distinguish nouns denoting sentiments from other nouns, for example "ter ódio de" vs. "ter camisa de" (lit. to have hate of vs. to have shirt of).

The analysis of these lists identified 173 combinations of sentiment nouns into the patterns, distributed as follows:

PATTERN	VALIDATED CANDIDATES
1. Sentir * de	22
2. Sentir * por	13
3. Ter * de	69
4. Ter * por	29
5. Ficar com * de	14
6. Estar com * de	16
7. Dar * em	10

Table 2. Validated candidates per pattern

PATTERN	PRECISION
1. Sentir * de	44.90%
2. Sentir * por	72.22%
3. Ter * de	5.67%
4. Ter * por	22.14%
5. Ficar com * de	27.45%
6. Estar com * de	17.39%
7. Dar * em	4.65%

Table 3. Percentage of validated candidates per pattern

Comparing the quantity of candidates analysed (Table 1) with the quantity of candidates validated (Table 2), we found the precision of each pattern (Table 3). This measure indicates how

much a pattern is associated with sentiment nouns or, in other words, how specific is a pattern to express feelings.

The pattern “ter * de” returned the largest amount of validated candidates, but, at the same time, it is the one that presented one of the largest amounts of noise. This is most probably due to the high polysemy of the verb "ter" (to have). In this sense, the patterns “sentir * de” and “sentir * por” are much less ambiguous and their precision ranges from 44.9% to 72.22%, respectively. Patterns 5 and 6 have a similar profile; both are responsible for 8 and 9% of the final list, with a precision between 17.39% (estar) and 27.45% (ficar). Pattern 7 presents the lowest precision, 4,5%, what is expected as the verb “dar” is highly polysemous in Portuguese.

In spite of “ter”, “ficar”, “estar” and “dar” being very polysemous verbs, every time they integrate a collocation with a sentiment, they will have an unambiguous sense, that is`

ter=sentir (to feel),
ficar=começar a sentir (start to feel),
estar=sentir temporariamente (feel temporarily).
dar=provocar (make to feel),

This observation proves Yarowsky’s intuition about “one sense per collocation” [12].

When annotating the candidates, we also noticed that most of the expressions were actually expressing negative emotions. We have two hypotheses to explain this fact: either this is a bias from our newspaper corpus (there are often more bad news than good news in general newspapers) or Brazilian Portuguese native speakers prefer to use the identified patterns instead of sentiment verbs because they somehow diminish/blur the impact of the negative emotion expressed.

6. Analysing sentiment nouns expressed by the patterns (STEP 4)

The 173 validated candidates present, evidently, some repetitions of nouns. Eliminating the redundancies, we obtained a list of 98 sentiment nouns. We observed some features associated to these sentiment nouns that could be used to further annotate them.

For example, we annotated the polarity [4, 5, 7], associated to each sentiment noun. This was double annotated, as it involves subjectivity. The result is shown in Table 4.

POLARITY	N	EXAMPLE
negative	45	hate, contempt, grudge
positive	29	love, tenderness, compassion
neutral	15	Interest, impression, curiosity
context dependent	9	pride, ambition, anxiety

Table 4. Distribution of sentiment nouns according to their polarity

Another feature we observed is the “source” of the feeling expressed by the sentiment noun. This made it possible to distinguish physical sensations, expressed through the same patterns, from more psychological feelings. As well, we separated rational feelings from emotional feelings, as shown in Table 5.

SOURCE	QUANTITY	EXAMPLE
psychological-emotional	67	jealousy, sympathy, anger
psychological-rational	18	confidence, respect, concern
physical	13	cold, thirst, hunger, pain

Table 5. Distribution of sentiment nouns according to their source

7. Patterns and sentiment nouns in the web (STEP 5)

In this step we merged the 98 sentiment nouns identified in the third step with the seven patterns identified in the first step, thus artificially generating 686 collocations that were automatically looked up in the web. Additionally, as Portuguese has verb inflections and in web we can not search for lemmas, for each collocation we generated three inflected forms. For instance, the candidate "ter medo de" (to have fear of) became "ter|tem|teve|tinha medo de" (to have|has|had|was having fear of), where the vertical bar | denotes the alternative. That is, this query retrieves any sequence containing one of the forms of verb "ter" in infinitive, present, past perfect or imperfect followed by the target sentiment noun and the corresponding preposition.

Results showed some collocations with zero occurrences. This may be due to the inexistence of the combination or due to limitations of our search arguments, which should be refined. For example, we realized that the pattern "dar * em" is almost always presented with a personal pronoun taking the place of the experiencer, avoiding the preposition "em" and preceding the verb: "*Isso me dá medo*" (lit. This gives me fear). The same pattern may be used without the experiencer, in utterances like "*Dá medo pensar nisso*" (lit. Give fear thinking about this = Thinking about this causes fear).

Aiming to evidence whether the preferred way to express feelings varies according to the feeling expressed, we built Table 6. In this table we show how many sentiment nouns take each pattern as preferred pattern. This table evidences the pattern "ter * de" as the preferred one for expressing 61 of a total of 98 sentiment nouns. Therefore, this pattern is extensively used to express feelings. However, all the patterns are preferred by, at least, two sentiment nouns, as is the case of "estar com * de".

Preferred Pattern	Sentiment Nouns
Ter * de	61
Sentir * por	14
Ter * por	12
Dar * em	6
Sentir * de	3
Estar com * de	2

Table 6. Distribution of preferred patterns

In Table 7, we present the quantity of sentiment nouns that accept[3] one or more patterns. With these data, we are able to distinguish more flexible constructions from more fixed ones. Lexicalised constructions present zero frequency for all alternative patterns except for the preferred one. This is the case of four sentiment nouns, as may observed in the last line of Table

[3] We say that a noun "accepts" a pattern if the frequency returned by the web search engine is greater than 3 pages, thus avoiding noise probably due to typos and artificial results.

7. Most of the nouns, however, are quite flexible and accept several patterns, although it is not clear whether alternative patterns express the same sentiment with the same connotation and use.

Quantity of Patterns	Sentiment Nouns
7	26
6	17
5	14
4	15
3	13
2	9
1	4

Table 7. Quantity of sentiment nouns vs. quantity of patterns

8. Future work

The growing importance of sentiment analysis encourages further developments of this work. It would be interesting, for instance, to compare, across genres, utterances using sentiment verbs with utterances using the patterns we have identified. For this purpose, one may use the list of sentiment verbs from Brazilian Wordnet [13], provided in Appendix 1, and the sentiment nouns obtained in this study, listed in Appendix 2, associated with the patterns here discussed.

A limitation of our work is using a corpus of news. A corpus of speech or blogs [14] or social networking would more likely present sentiment expression material. Even though, the results obtained here can be fed back into computational systems that try to automatically extract polarity or execute sentiment analysis of textual data. As a by-product, we expect to discover new features for automatic verb clustering [15].

Furthermore, these collocations may be used to improve bilingual dictionaries with information on how to express sentiments from the point-of-view of a Brazilian speaker.

9. Acknowledgments:

Our thanks to FAPESP for the financial support. This research was partly supported by CAMELEON project (CAPES-COFECUB 707-11).

References:

[1] Bo Pang and Lillian Lee (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. Vol. 2, issues 1-2, pp 1-135.

[2] Juan Miguel López, Rosa Gil, Roberto García, Idoia Cearreta, Nestor Garay (2008). Towards an ontology for describing emotions. In: Emerging Technologies and Information Systems for the Knowledge Society, LNCS, 2008, Volume 5288/2008, 96-104. Springer.

[3] Yvette Yannick Mathieu (2005). Annotation of Emotions and Feelings in Texts. Conference on Affective Computing and intelligent Interaction (ACII2005), Springer Notes in Computer Science. Berlin/Heidelberg: Springer. Available at: http://www.aui.computing.edu.au/acii/docs/104.pdf

[4]. Soo-Min Kim; Eduard Hovy (2004). Determining the Sentiment of Opinions. Proceedings of the COLING Conference, Geneva, 2004.

[5] A. Esuli and F. Sebastiani (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, Genova, IT, 2006.

[6] Paula Carvalho, Luís Sarmento, Mário J. Silva, Jorge Teixeira (2011). Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates.9th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HTL) Portland, Oregon, USA, June, 2011.

[7] Mário J. Silva, Paula Carvalho, Luís Sarmento, Eugénio Oliveira, Pedro Magalhães (2009). The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics.New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence p. 565-576, October, 2009. Universidade de Aveiro.

[8] Magali Sanches Duran; Carlos Ramisch; Sandra Maria Aluísio e Aline Villavicencio (2011). Identifying and Analyzing Brazilian Portuguese Complex Predicates. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 74–82, Portland, Oregon, USA, 23 June 2011. Association for Computational Linguistics.

[9] Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato (2002). Seeing arguments through transparent structures. In Proc. of the Third International Conference on Language Resources and Evaluation (LREC 2002). pp 787-791, Las Palmas, Canary Islands, Spain, May.

[10] Beth Levin (1993). English Verb Classes And Alternations: A Preliminary Investigation, The University of Chicago Press.

[11] Carlos Ramisch, Aline Villavicencio, and Christian Boitet (2010) Multiword expressions in the wild? The mwetoolkit comes in handy. In Proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, Aug.

[12] David Yarowsky. (1993) One Sense per Collocation. In: Proceedings of ARPA Human Language Technology Workshop, Princeton.

[13] Bento Carlos Dias da Silva. (2010) Brazilian Portuguese WordNet: A Computational Linguistic Exercise of Encoding Bilingual Relational Lexicons. International Journal of Computational Linguistics and Applications, New Delhi, v.1, n. 1-2, p.137 - 150, 2010.

[14] Alastair J. Gill, Robert M. French, Darren Gergle, Jon Oberlander. 2008. The Language of Emotion in Short Blog Texts. In Proc. of the 2008 ACM Conference on Computer supported cooperative work.

[15] Tim van de Cruys, Thierry Poibeau and Anna Korhonen (2011). "Latent Vector Weighting for Word Meaning in Context ". Proceedings of Empirical Methods in Natural Language Processing (EMNLP). Edinburgh.

Appendix 1. Sentiment Verbs extracted from Brazilian WordNet

abalar
abominar
aborrecer-se
abrandar
acalmar
acalmar-se
acender-se
acovardar-se
adorar
afligir
agitar-se
agradar
alarmar
alarmar-se
alegrar
aliviar
alterar
alucinar
alvoroçar
animar
antipatizar
apiedar
apoquentar
apreciar
arrasar
assanhar
atormentar-se
atraiçoar
atrair
atrapalhar-se
babar-se
cativar
chatear
cobiçar
comover
comover-se
compadecer-se
conciliar
confortar
conquistar
consolar
consolar-se
consumir-se
decepcionar
decepcionar-se
deleitar-se
desadorar1
Desagradar
desagradar-se
desagradecer
desalentar-se
desangustiar
desanimar
desapoquentar
desassossegar
desconfortar
desejar
desemburrar
desemburrar-se
desencabular
desencorajar
desenjoar
desesperar-se
desfazer-se
desiludir
desinteressar
desmotivar
despertar
despreocupar
desprezar
distrair-se
doer
embaraçar
emburrar
encantar
encantar-se
encorajar
enfurecer
enfurecer-se
enlouquecer
enlouquecer-se
enlutar
enlutar-se
entristecer
entristecer-se
entusiasmar
entusiasmar-se
envaidecer-se
envergonhar
espezinhar
estimar
estimular-se
exasperar
exasperar-se
excitar
expectar
expiar
fascinar
frustrar
fustigar
horrorizar
horrorizar-se
humilhar-se
impacientar-se
incomodar
inferiorizar-se
inquietar-se
intimidar
intimidar-se
invejar
irar-se
irritar-se
irromper
lastimar
magoar-se
malucar
nublar
nublar-se
obsequiar
orgulhar-se
penitenciar-se
perrengar
perturbar
perturbar-se
pirraçar
preferir
preocupar-se
rebaixar-se
simpatizar
sossegar
temer
torturar
venerar
zangar

Appendix 2. Sentiment Nouns Identified

admiração	disposição	pavor
adoração	dó	pena
ambição	dor	piedade
amor	dor-de-cabeça	prazer
angústia	dúvida	predileção
ansiedade	esperança	preguiça
antipatia	expectativa	preocupação
apego	fadiga	pudor
apelo	falta	raiva
apreço	fascinação	rancor
asco	fobia	receio
aspiração	fome	rejeição
atração	frio	remorso
bronca	gosto	repugnância
carinho	horror	repulsa
certeza	ímpeto	respeito
cheiro	impressão	responsabilidade
choque	instinto	sabor
ciúme	interesse	saudade
compaixão	inveja	segurança
complexo	irritação	sensação
confiança	mágoa	sentimento
consciência	medo	simpatia
constrangimento	moleza	sintoma
convicção	necessidade	suador
coragem	nojo	suspeita
culpa	nostalgia	tentação
curiosidade	obsessão	tranquilidade
desejo	ódio	trauma
desespero	orgulho	tristeza
desprezo	paciência	vergonha
devoção	paixão	vontade
dificuldade	pânico	