IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 147–158
RECEIVED 15/12/12 ACCEPTED 11/01/13 FINAL 27/03/13

Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis

LARISSA A. DE FREITAS AND RENATA VIEIRA

PUCRS, Brazil

ABSTRACT

In this paper we evaluate different lexicons in feature level opinion mining on Brazilian Portuguese movie reviews. Research in this field often considers English data, while other languages are less explored. So we discuss and compare available resources and techniques that can be applied to Portuguese for dealing with this task. We found better results when using SentiLex adjetives. The results indicate a F-score of 0.73 for positive polarity recognition and 0.76 for negative polarity recognition.

KEYWORDS: *Opinion Mining; Sentiment Analysis; Portuguese Online Reviews; Movie Reviews*

1 INTRODUCTION

Studies about "opinion mining", also called "sentiment analysis"(SA) have been developed more intensively in the last decade. In general, research in this area focuses in detecting the holder's sentiment about a topic in a review. Opinions are important because whenever we need to make a decision, we want to know other points of view.

Nowadays, opinion mining has been investigated mainly in three levels of granularity (document, sentence or feature). According to Liu [1], both the document level and sentence level analyses do not discover what exactly people liked or not. However, feature level opinion mining, required for that, is extremely challenging.

In feature-based opinion mining, features related to an object are analysed. This technique comprises the following steps: identifying the features about the object in review, deciding whether the review is positive or negative and summarizing the information [2]. Overall, the output is a tuple containing the feature and the polarity of objects. The model of feature-based opinion mining is proposed by many researchers, such as Hu and Liu [3] and Popescu and Etzioni [4].

In the literature, recent works about ontology-based opinion mining in feature level are Zhao and Li [5] and Peñalver-Martínez et al. [6]. Both have been applied on English movie reviews, presenting high quality results.

In this context, we address the issue of feature-based opinion mining but applied on Brazilian Portuguese movie reviews. We used part-of-speech (POS) tags, movie ontology concepts and two available Portuguese opinion lexicons.

This paper is organized as follows: works about feature-based opinion mining are discussed in Section 2. Our approach is introduced in Section 3. Tests are discussed in Section 4. Finally, conclusion and future works are presented in Section 5.

2 FEATURE-BASED OPINION MINING

The works by Hu and Li [3] and Popescu and Etzinoni [4] are the most representative ones in this area of study. Hu and Li [3] use association rule mining while Popescu and Etzinoni [4] use the Pointwise Mutual Information (PMI) for feature extraction. According to Hu and Li [3] implicit features occur much less frequently than explicit ones. This paper focuses on features that appear explicitly in the reviews.

Most of the existing work on review mining and summarization is focused on product reviews [3, 4]. When people write a movie reviews, they probably comment not only on movie elements (e.g., music, vision effectcs, award, genre), but also on movie-related people (e.g., director, actor, writer, producer). Therefore, the commented features in movie reviews are much richer and more challenging than other domain, such as: hotel, restaurant and product. Zhuang et al. [7] have done a pioneer work on classifying and summarizing movie reviews by extracting high frequent opinion keywords. Feature-opinion pairs were identified by using a dependency grammar graph.

Binali et al. [8] present an overview about feature-based opinion mining. The following tasks are identified: the extraction of objects (entities

mentioned in reviews e.g., movie); the extraction of object features (components and attributes e.g., title); the detection of sentiment about object features (e.g., good title); the detection of sentiment about objects (the global sentiment expressed in relation to an entity e.g., recommended or not recommended); the comparison of two entities (e.g., movie A and movie B); the comparison of features of two entities (e.g., actors movie A and actors movie B). In our study, we intend to extract object features and detect sentiment about object features.

Feature-based opinion mining that uses ontologies, in the English language are [6, 9, 5, 10]. The literature shows that there are different levels of knowledge representation: authors using complex structures [6, 9, 10]—even if they do not use all the knowledge available—and authors using simple structures [5] for feature identification. A common point is the use of IMDb data. Unfortunately, the ontologies cited in [9, 5, 10] are not available. The only ontology we found was the Movie Ontology (MO[1]).

In this paper, we conducted the adaptation of the algorithm Polarity Recognizer in Portuguese (PIRPO) [11] applied to Brazilian Portuguese movie reviews and using MO concepts (Figure 1). PIRPO receives as input a set of reviews which are pre-processed in order to extract their sentences and detect which reviews are split into positive and negative segments. The system output is a list of sentences with polarity that reflects the polarity of the words characterising the concepts of the ontology in the reviews [11].

3 APPROACH

This approach is composed of two main steps: preprocessing and semantic orientation recogniser. These steps are described in detail below.

3.1 *Preprocessing*

The main objective of this step is to obtain the grammatical categories. For this task we used Portuguese TreeTagger[2]. The TreeTagger is a tool for annotating text with POS and lemma information. For example, the sentence “Um dos melhores filmes que já vi!” [“One of the best movies I have watched!”] and “É simplesmente o PIOR filme que vi nos últimos

[1] http://www.movieontology.org/

[2] http://gramatica.usc.es/ gamallo/tagger.htm

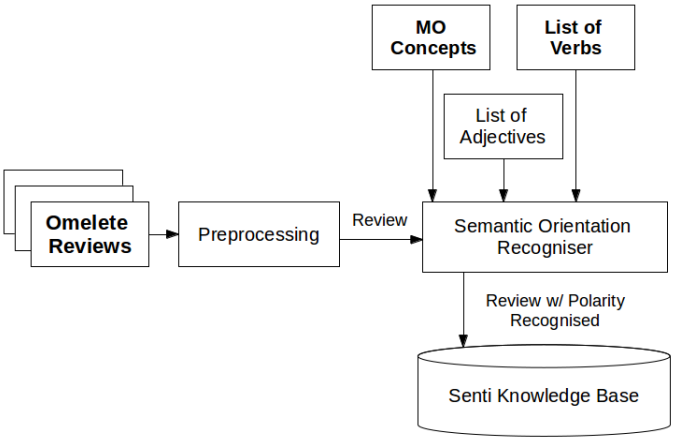

Fig. 1. PIRPO Information Architecture. Adapted from [11].

tempos." ["It is simply the WORST movie I have watched lately."] obtains the following lemmatized words accompanied by their grammatical categories:

Um DET um dos PRP+DET de melhores ADJ melhor filmes NOM filme que PR que já V <unknown> vi V ver ! SENT !

É V <unknown> simplesmente ADV simplesmente o DET o PIOR NOM pior filme NOM filme que PR que vi V ver nos P nos últimos V <unknown> tempos NOM tempo . SENT .

3.2 *Semantic Orientation Recogniser*

In this step, external resource was used, such as: ontology concepts and opinion lexicons.

The main idea is to use the opinion words around each movie concept in a review sentence to determine the opinion orientation. Still, the orientation of an opinion on a feature indicates whether the opinion is positive, negative or neutral.

In our work, features are represented by concepts the MO of ontology. Firstly, concepts are identified and extracted of pre-processed reviews.

For example:

Um DET um dos PRP+DET de melhores ADJ melhor **filmes NOM filme** que PR que já V <unknown> vi V ver ! SENT !

After, we used opinion lexicons, i.e., adjectives or verbs contained in SentiLex and OpLexicon for polarity identification. The adjectives or verbs around each movie feature identified are analysed.

For example, when we use the list of adjectives:

Um DET um dos PRP+DET de **melhores ADJ melhor filmes NOM filme** que PR que ja V <unknown> vi V ver ! SENT !

We identified the adjective "melhores" ["best"] near the word "filme" ["movie"]. In SentiLex this adjective is neutral and in OpLexicon is positive.

For example, when we use the list of verbs:

Um DET um dos PRP+DET de melhores ADJ melhor **filmes NOM filme** que PR que já V <unknown> **vi V ver** ! SENT !

We identified the verb "ver" ["watch"] around "movie" ["filme"]. In OpLexicon this verb is positive. SentiLex did not have this verb.

Finally, the output, a tuple containing the feature and polarity of objects, is stored in a database.

For example, tuple: (movie, positive).

4 TESTS

In this section, we evaluate the algorithm using the semantic orientation recogniser. We have conducted tests using the movie corpus, the MO concepts and Portuguese lexicons (SentiLex[3] and OpLexicon[4]). These resources are described below.

4.1 *Movie Corpus*

In order to build the movie corpus, initially we automatically got reviews about 1.160 movies on the website Omelete[5]. In these tests, 150 reviews were randomly selected. The corpus has only 8.999 words and 440 sentences. After that, TreeTagger is used to generate part-of-speech tags.

The manual annotation of the corpus was conducted by two people. The agreement between annotators was measured through the Kappa Statistics. The Kappa Statistics is the metric that evaluate concordance level in classification task. The value was moderate (Kappa 0.58) for agreement about opinion mining and fair (Kappa 0.39) for agreement

[3] http://xldb.fc.ul.pt/wiki/SentiLex-PT01

[4] http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php

[5] http://omelete.uol.com.br/

about feature identification (see Table 1). We believe that the annotation has an acceptable value for the problem proposed in this study. In manual annotation the most frequent concepts were: *movie*, *actor*, *people* and *genre*.

Table 1. Kappa Statistics [12].

Interval	Agreement
< 0.00	Poor
0.00 to 0.20	Slight
0.21 to 0.40	**Fair**
0.41 to 0.60	**Moderate**
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

4.2 *Movie Ontology*

In this study, we used the concepts of MO. MO aims at providing controlled vocabulary to describe semantically related concepts, such as a movie, genre, director, actor and individuals—for example "A Era do Gelo 3" ["Ice Age 3"], "Animação" ["Animation"], "Carlos Sandanha" and "Márcio Garcia", respectively. This ontology was described in OWL and is available in English. MO provides hierarchies of concepts and a set of instances. Only 11 out of 78 concepts (Table 2) were identified in the movie corpus, such as: *action*, *actor*, *director*, *fun*, *genre*, *kids*, *love*, *movie*, *place*, *person*, and *thrilling*.

Table 2. Movie Ontology Metrics.

Metrics	Value
Number of Concepts	78
Number of Object Properties	38
Number of Data Properties	4
Number of Individuals	282

4.3 *Portuguese Lexicons*

In the literature, many papers about opinion mining use SentiWordNet[6] [13]. SentiWordNet 3.0 [14] is a fragment of WordNet 3.0 manually annotated for positivity, negativity and neutrality. Each synset has three numeric values in the interval 0 to 1 for positive, negative and neutral. Both [5] and [6] calculated the polarity of the features using SentiWordNet. This resource has nearly 117.000 words in English.

There are languages in which this type of resource started to be built recently, as is the case of Portuguese. SentiLex and OpLexicon, Portuguese opinion lexicons, appeared in 2010.

SentiLex 2.0 [15] has 7.014 lemmas and 82.347 inflected forms (of nouns, verbs, adjectives and adverbs). SentiLex is useful for opinion mining applications involving European Portuguese, in particular for detecting and classifying sentiments and opinions. In tests we used 16.833 SentiLex adjetives and 28.989 SentiLex verbs (Table 3).

OpLexicon [16] has nearly 30.322 words and was built based on a corpus, thesaurus and translated texts. Three different opinion lexicons generated by each techniques are conjoined to create a large lexicon for Brazilian Portuguese. In tests we used 23.433 OpLexicon adjetives and 6.889 OpLexicon verbs (Table 3).

Table 3. Portuguese Lexicons.

Lexicon	Number of Words
SentiLex Adjectives	16.833
SentiLex Verbs	28.989
OpLexicon Adjectives	23.433
OpLexicon Verbs	6.889

Even though SentiLex or OpLexicon are small and new, we used this lexicon. Both have three numeric values: 1 (positive), -1 (negative) and 0 (neutral).

4.4 *Results*

The results are presented in Table 4. In the table, lines 2 and 8 give the results that uses OpLexicon adjectives and MO concepts for positive and

[6] http://sentiwordnet.isti.cnr.it/

negative polarity recognition. The results indicate that precision for negative polarity recognition is poor. Lines 3 and 9 show corresponding results that uses SentiLex adjectives and MO concepts. We can see that the f-measure is the best result. Lines 4 and 10 give the results that uses OpLexicon verbs and MO concepts for positive and negative polarity recognition. The results also indicate that precision for negative polarity recognition is poor. Lines 5 and 11 show corresponding results that uses SentiLex verbs and MO concepts. We can see that the f-measure is the same as positive polarity recognition as negative polarity recognition.

In summary, the best results are obtained when using SentiLex adjectives, the f-measure of 73% for positive polarity recognition and 76% for negative polarity recognition.

Table 4. Results for Feature-Based Opinion Mining.

		Precision	Recall	F-Measure
Positive	OpLexicon(ADJ) + MO(C)	1.0	0.45	0.62
	SentiLex(ADJ) + MO(C)	**0.87**	**0.63**	**0.73**
	OpLexicon(V) + MO(C)	1.0	0.40	0.57
	SentiLex(V) + MO(C)	1.0	0.50	0.66
	OpLexicon(ADJ and V) + MO(C)	1.0	0.43	0.61
	SentiLex(ADJ and V) + MO(C)	0.90	0.57	0.70
Negative	OpLexicon(ADJ) + MO(C)	0.08	1.0	0.15
	SentiLex(ADJ) + MO(C)	**0.66**	**0.88**	**0.76**
	OpLexicon(V) + MO(C)	0.04	1.0	0.08
	SentiLex(V) + MO(C)	0.50	1.0	0.66
	OpLexicon(ADJ and V) + MO(C)	0.11	1.0	0.20
	SentiLex(ADJ and V) + MO(C)	0.63	0.92	0.75

4.5 *Error Analysis*

In the following, we show a few examples to analyse some typical errors. Bold is used to denote feature objects and adjectives or verbs polarity indicates.

Example 1:

Sentence: “incrível o filme, me emocionei em alguns momentos, perfeitos.” [“amazing film, moved in some moments, perfect.”]

Annotated Sentence: (‘incrível’, ‘ADJ’), (‘o’, ‘DET’), (‘**filme,me**’, ‘NOM’), (‘emocionei’, ‘V’), (‘em’, ‘PRP’), (‘alguns’, ‘P’), (‘momentos,perfeito’, ‘V’), (‘.’, ‘SENT’)

Error: filme,me NOM

Expected: filme NOM

Here the word “filme” [“movie”] is grouped with comma and pronoun “me” [“me”]. In fact, there are many writting error in movie reviews. To solve the problem, a heuristic should be build.

Example 2:

Sentence: “... esse filme apesar de ruin causou ...” [“... this movie although bad cause ...”]

Annotated Sentence: ... (‘esse’, ‘DET’), (‘**filme**’, ‘NOM’), (‘apesar’, ‘L’), (‘de’, ‘PRP’), (‘**ruin**’, ‘NOM’), (‘causou’, ‘V’) ...

Error: ruin NOM

Expected: ruim ADJ

The word “ruim” [“bad”] is misspelled. Maybe phonetic algorithm or spellchecker should be used to solve the problem.

Example 3:

Sentence: “Filme excelente, elenco competente, direção fantástica, trilha sonora de Alberto Iglesias no mínimo brilhante ...” [“Excellent movie, competent cast, fantastic direction, trowel Alberto Iglesias score of at least brilliant ...”]

Annotated Sentence: (‘**Filme**’, ‘NOM’), (‘excelente’, ‘ADJ’), (‘,’, ‘VIRG’), (‘**elenco**’, ‘V’), (‘competente’,‘ADJ’), (‘,’, ‘VIRG’), (‘**direção**’, ‘V’), (‘fantástica’, ‘V’), (‘,’, ‘VIRG’), (‘**trilha**’, ‘NOM’), (‘**sonora**’, ‘ADJ’), (‘de’, ‘PRP’), (‘Alberto’, ‘NOM’), (‘Iglesias’, ‘NOM’), (‘no’, ‘PRP+DET’), (‘mínimo’, ‘NOM’), (‘brilhante’, ‘ADJ’) ...

Error: (movie, positive)

Expected: (movie, positive), (cast, positive), (direction, positive), and (soundtrack, positive)

This sentence has a (movie, positive), (cast, positive), (direction, positive), and (soundtrack, positive) tuple but the algorithm only detected a (movie, positive) tuple for review.

5 Conclusion and Future Works

In summary, the application of the adaptation of the algorithm proposed in [11] in the movie domain presented good results. In future works we intend to use the complete ontology (*concepts*, *properties*, *instances* and

hierarchies). Furthermore, we intend to redo these tests in other domains, such as: education, politics, and others.

Aiming at improving the results, the preprocessing step might be broadened. We intend to use lemmatizer in preprocessing and properties, instances and hierarchies of ontologies in identification feature.

Also, we intend to add lists of adverbs and list of nouns in polarity identification. At last, we would apply a set of linguistic rules, such as negatives and intensifiers which vary from language to language [17]. In opinion mining, the negation is a more common linguistic construction that affects the polarity. It is not only transmitted by negative words, but also by lexical units, such as diminutives and connectives. The works described in [17–19] were considered pioneers in the negation model in sentiment analysis.

Besides, we intend to study ways of solving problems such as the use of different words (e.g., *filmes* and *filmão*) that refer to the same concept.

6 Acknowledgments

We thank the Brazilian funding agency CAPES/FAPERGS for the scholarship granted.

References

1. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group (2010)
2. Bhuiyan, T., Xu, Y., Josang, A.: State-of-the-art review on opinion mining from online customer's feedback. In: 9th Asia-Pacific Complex Systems Conference. (2009)
3. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: 19th national conference on Artifical intelligence. (2004) 755–760
4. Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. (2005) 339–346
5. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: 3rd International Conference Knowledge, Science, Engineering and Management. (2009)
6. Peñalver Martínez, I., Valencia-Garcia, R., Garcia-Sanchez, F.: Ontology-guided approach to feature-based opinion mining. In: International Conference on Applications of Natural Language to Information Systems. (2011)

7. Zhuang, L., Jing, F., Zhu, X.: Movie review mining and summarization. In: 15th ACM international conference on Information and knowledge management. (2006)
8. Binali, H., Potdar, V., Wu, C.: A state of the art opinion mining and its application domains. In: International Conference on Industrial Technology. (2009)
9. Shein, K.P.P.: Ontology based combined approach for sentiment classification. In: 3rd International Conference on Communications and Information Technology. CIT'09, Stevens Point, Wisconsin, USA, World Scientific and Engineering Academy and Society (2009) 112–115
10. Zhou, L., Chaovalit, P.: Ontology-supported polarity mining. Journal of the American Society for Information Science and Technology **59** (2008) 98–110
11. Chaves, M., Freitas, L., Souza, M., Vieira, R.: Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. In: 17th International Conference on Applications of Natural Language Processing to Information Systems. (2012)
12. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics **33** (1977) 159–174
13. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: 5th International Conference on Language Resources and Evaluation. (2006)
14. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: 7th International Conference on Language Resources and Evaluation. (2010)
15. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: 10th International Conference Computational Processing of the Portuguese Language. (2012)
16. Souza, M., Vieiras, R., Busetti, D., Chishman, R., Alves, I.M.: Construction of a portuguese opinion lexicon from multiple resources. In: 8th Brazilian Symposium in Information and Human Language Technology. (2012)
17. Polanyi, L., Zaenen, A.: Contextual valence shifters. AAAI Spring Symposium on Attitude **20** (2004) 1–10
18. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence **22** (2006) 110–125
19. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics **35** (2009) 399–433

LARISSA A. DE FREITAS
FACULDADE DE INFORMÁTICA,
PUCRS,
PORTO ALEGRE, BRAZIL
E-MAIL: <LARISSA.FREITAS@ACAD.PUCRS.BR>

RENATA VIEIRA
FACULDADE DE INFORMÁTICA,
PUCRS,
PORTO ALEGRE, BRAZIL
E-MAIL: <RENATA.VIEIRA@PUCRS.BR>