Entity-centric Sentiment Analysis on Twitter data for the Potuguese Language

Marlo Souza[1], **Renata Vieira**[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul - UFRGS
Porto Alegre – RS – Brazil

[2]Faculdade de Informática – Pontifífica Universidade Católica do Rio Grande do Sul
Porto Alegre – RS – Brazil

marlo.souza@inf.ufrgs.br, renata.vieira@pucrs.br

Abstract. *Twitter is a popular microblogging platform which is commonly used to express opinions about entities of the world. The solutions provided to perform Sentiment Analysis in such a media, however, relies on classifying an entire sentence regarding the opinion it express, rather than the content and reference of the opinion expressed in the text. We propose and evaluate a Entity-centric Sentiment Analysis method over Twitter data for the Portuguese language.*

1. Introduction

Twitter is a popular microblogging platform released in 2006 and in wide-spread use. Sentiment analysis in Twitter data has been used in many commercial tools for Social Media Monitoring and Competitive Intelligence. In our opinion, the depth of analysis performed is, however, inadequate for the task, since most tools focuses on a sentence level.

We propose and evaluate a modular entity-centric Sentiment Analysis (ESA) method over Twitter data for the Portuguese language. The current paper is structured as follows: we present the most influential work on entity-based sentiment analysis and opinion mining on Twitter microtexts in Section 2. On Section 3, we present our proposal, which combines multiple techniques already developed in the literature to perform entity-centric sentiment analysis. We, then, evaluate our methods (Section 4).

2. Related Work

While multiple solutions have been proposed for identification of opinionated expressions in text, work on entity-centric sentiment analysis, i.e. to associate opinions with its referent, fall over three major approaches: those which use the context of an entity - as a fixed window of words around the entity or its syntactic context - to identify an opinion about the entity [Grefenstette et al. 2004, Hu and Liu 2004]; those which use pre-defined rules and linguistics resources - such as FrameNet - to identify the opinion reference as [Ding et al. 2008, Kim and Hovy 2006, Wu et al. 2009]; and those which relies on machine learning techniques as [Popescu and Etzioni 2005, Kobayashi et al. 2007, Ding and Liu 2010].

More related to our work, however, are the work of Jansen et al. [Jansen et al. 2009] and Silva et al. [Silva and TEAM 2011]. Jansen et al. use out-of-the-box commercial tool - no longer available - to perform Entity-centric subsentential sentiment analysis on Twitter. They apply their strategy on brand names for word-of-mouth detection.

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 173–177,
Fortaleza, CE, Brazil, October 21–23, 2013.

Silva et al. [Silva and TEAM 2011] describe the construction of the Twittómetro - a tool for subsentential sentiment analysis on Twitter for the political domain. They explore a dictionary-based approach combined with lexico-syntactic rules to identify and compose opinions and to attribute reference to them.

We believe that, while these work address a problem similar to ours, their strategies are not adequate for our case since they do not perform the analysis on a sufficiently grained fashion, as in [Jansen et al. 2009], or they rely too strongly on the structure of the domain, as in [Silva and TEAM 2011].

3. Entity-centric sentiment analysis on Twitter data

Given the difficulty of working with Twitter data - an extremely noisy channel - and the inexistence of Twitter specific linguistic processors for the Portuguese language, we opted to use only shallow linguistic information, such as lexical and morphological information. We perform the necessary steps to perform entity-centric sentiment analysis, such as entity identification and opinion expression identification separately and integrate the partial results with a opinion reference resolver.

3.1. Subsentential sentiment analysis

In the opinion identification and polarity classification, we rely on a dictionary-based approach, similar to [Souza and Vieira 2012]. The method may be summarized by searching the opinion expression of the lexicon in the tweet. Since many words on the lexicon are in the canonical form, we apply a Stemmer and search for the words in the tweet if there is a polarized word with the same stem in the lexicon. The polarity is determined by the lexicon and the presence of a negation particle in the vicinity of the opinion expression. As opinion lexicon, we employ the OpLexicon [Souza et al. 2011] which already contains polarized emoticon and hastags - Twitter user-generated metadata.

3.2. Named entity recognition

Following the work of Liu et al. [Liu et al. 2011] for NER on Twitter data and the Ratinov and Roth [Ratinov and Roth 2009], we developed a NER system based on a Conditional Random Fields tagger. As features for the NER system, we provide Ratinov and Roth's [Ratinov and Roth 2009] lexical and morphological features and external information features - based on Repentino name gazetteer [Sarmento et al. 2006].

3.3. Opinion reference resolution

To identify which opinion-bearing expressions reference which named entity, we apply a opinion reference resolution method. In this phase, those opinion expressions that do not refer to a mentioned entity will be discarded. The results of this phase is then the annotated text. We implement a linear Support Vector Machine (SVM) classifier with the features:

- Positional features: location of the opinon expression (OE) and entity in the sentence, distance between the OE and the entity, centrality of the OE and entity in the sentence;
- Number of identified entities in the sentence;
- Length of the sentence;

- Number concordance of expression and entity.

Once established the methods employed, we will now discuss the implementation of the prototype and its usage to validate our method for entity-centric sentiment analysis in Twitter data.

4. Evaluation

We implemented a prototype of the previously discussed methods in the Python programming language using the NLTK[1] language toolkit for linguistic processing and the Mallet[2] and SciKit toolkits for the Machine Learning techniques. Since we perform the extraction of referenced opinion in three different processes, namely opinion mining, named-entity extraction and opinion reference resolution, the evaluation will be performed individually for each task. We perform an intrinsic evaluation of each method using a common manually annotated resource created for this purpose [Souza 2012].

4.1. Subsentential sentiment analysis

Since in the corpus only those opinionated expressions which referred to a entity explicitly mentioned in the tweet were annotated, we chose to evaluate the each annotated opinionated expression in the corpus would be classified by the sentiment analysis method previously discussed. Note that we do understand that applying our method directly to the text would generate more - non-evaluated - expressions, but they should be discarded in the opinion reference identification step.

The results of the evaluation over the 130 opinions annotated in the corpus may be seen in the Table 1.

Table 1. Sentiment Analysis method evaluation

	Method			Metrics		
Anotation	Pos	Neutral or Non-opinion	Neg	Prec	Rec	F-measure
Pos	27	26	2	0.73	0.49	0.59
Neg	10	30	31	0.94	0.44	0.60

4.2. Named entity recognition

To evaluate the results of our method for NER in Twitter for Portuguese language, we implemented the method in Python using the NLTK and the Mallet toolkit as an implementation of the CRF tagger.

The cases in which a polylexical name has been identified as multiple entities have been counted as one partial correctly identified entity and the first entity of the set has been used to compute the error factor of [Santos et al. 2006]. Table 2 presents the results of the evaluation, along with the HAREM evaluation score - which are use to compute the Precision, Recall and F-measure, according to the definitions for the HAREM evaluation [Santos et al. 2006].

[1]http://nltk.org/

[2]http://mallet.cs.umass.edu

Table 2. NER evaluation

	Correct	Partial	Faulty	Spurious	Prec	Rec	F-measure
Number of occ	975	418	472	101	-	-	-
HAREM score	975.00	218.26	472.00	101.00	0.92	0.64	0.75

4.3. Opinion reference resolution

To evaluate the reference resolution method, as implemented in Python using the SciKits implementation of linear SVM, we performed a 10-fold cross validation on the evaluation corpus. To exclude influence of the errors of the previous phases, we used the entities and opinions as annotated by the human judges. The accumulated confusion matrix may be seen in Table 3.

Table 3. Accumulated confusion matrix of the opinion reference resolution method over 10-fold cross validation

	Method		Metrics		
Anotation	Refer	Don't Refer	Prec	Rec	F-Measure
Refer	94	39	0,69	0,71	0,70
Don't Refer	43	85	0,69	0,66	0,67

5. Discussion

Regarding the Sentiment Analysis method, we observed that many errors resulted from the low coverage of the OpLexicon. The use of morphosyntactical rules may help to extrapolate the data of the lexicon and identify patterns of opinions on text. The entity identification method achieved good results, specially for single word entities. Many errors occurred because of the system bias to classify multi-word entities as multiple simple entities. For reference identification, the main problem is that the system gives much importance to the distance between the entity and the opinion expression. Overall, however, the results achieved for such a hard task with such a simple method are very satisfying.

In the future, we plan to explore more reliable methods for opinion identification, such as a model for opinion composition or linguistic-inspired opinion expression patterns. Also, an hybridization of the opinion reference method with reference identification rules and patterns may be useful to improve the performance of the system.

References

Ding, X. and Liu, B. (2010). Resolving object and attribute coreference in opinion mining. In *23rd International Conference on Computational Linguistics*, COLING '10, pages 268–276, Stroudsburg, EUA. Association for Computational Linguistics.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *1st International Conference on Web search and web data mining*, pages 231–240, New York, EUA. ACM.

Grefenstette, G., Qu, Y., Shanahan, J. G., and Evans, D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In Fluhr, C., Grefenstette, G., and Croft, W. B., editors, *7th International Conference on Computer-Assisted Information Retrieval*, pages 186–194. CID.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *10th International Conference on Knowledge Discovery and Data mining*, pages 168–177, New York, EUA. ACM.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Kobayashi, N., Inui, K., and Matsumoto, Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In *12th Conference on Empirical Methods in Natural Language Processing*.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011). Recognizing named entities in tweets. In *ACL*, pages 359–367. The Association for Computer Linguistics.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *10th Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Morristown, EUA. Association for Computational Linguistics.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *13th Conference on Computational Natural Language Learning*, pages 147–155, Morristown, EUA. Association for Computational Linguistics.

Santos, D., Cardoso, N., and Seco, N. (2006). Avaliação no harem: Métodos e medidas. Technical report, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

Sarmento, L., Pinto, A. S., and Cabral, L. (2006). REPENTINO–a wide-scope gazetteer for entity recognition in portuguese. In *Computational Processing of the Portuguese Language*, pages 31–40. Springer.

Silva, M. J. and TEAM, R. (2011). Notas sobre a realização e qualidade do twitómetro. Technical report.

Souza, M. (2012). Mineração de opiniões aplicada a mídias sociais. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

Souza, M. and Vieira, R. (2012). Sentiment analysis on twitter data for portuguese language. *Computational Processing of the Portuguese Language*, pages 241–247.

Souza, M., Vieira, R., Busetti, D., 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*, Cuiabá, Brazil.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *14th Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541, Singapore. Association for Computational Linguistics.