Sentiment Analysis on Twitter Data for Portuguese Language

Marlo Souza and Renata Vieira

Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre RS, Brasil
marlo.souza@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. This work presents an study on Sentiment Analysis on Twitter data for the Portuguese language. It evaluates the impact of different preprocessing techniques, Portuguese polarity lexicons and negation models showing low impact of preprocessing and negation modelling in classification of tweets.

Keywords: Sentiment Analysis, Twitter, Portuguese language.

1 Introduction

Twitter is a microblog system in which users publish limited length messages. The importance of this system is due to its worldwide and fast growth - reaching 200 millions users and around 150 millions user publishings daily in 2011[1]. Twitter has been shown as an important source of information in a different range of areas - from social sciences [9], marketing and economy [12] and others-, but Twitter data poses difficulties for automatic linguistic analysis - due to a great deal of language variation and slangs.

Sentiment Analysis, or Opinion Mining, corresponds to the problem of identifying or extracting emotions, opinions or points of view expressed in text. This area has received a great deal of attention in the last years due to its potential applicability, according to Wilson et al. [26], among others.

Sentiment analysis techniques have been applied to several tasks in the literature (e.g. [10,18]) improving the treatment of non-factual information in text. More recently, Twitter data has been used in Sentiment Analysis studies [8,12,16,21]. Most of these works, however, focus on a rather shallow content analysis of the text due to the difficulties involved in processing Twitter data.

This work presents an study on sentiment analysis techniques on Twitter messages to improve Sentiment Analysis performance on this medium. We evaluate the impact of different models of negation and different sentiment lexicons for the Portuguese language and the effect of pre-processing techniques for the task. The remainder of this work is structured as follows: Section 2 presents the related work on sentiment analysis for Twitter; Section 3 presents our method for sentiment analysis on Twitter data, discussing some of our decisions and limitations; Section 4 presents the experiment performed using the described method and its results and, in Section 5, we conclude this work.

[1] Source: http://business.twitter.com/basics/what-is-twitter. Statistics of july, 2011

H. Caseli et al. (Eds.): PROPOR 2012, LNAI 7243, pp. 241–247, 2012.

2 Related Work

Work on sentiment analysis is expanding on the last years, applying the proposed solutions to a myriad of problems, such as opinion summarization [10], Information Extraction [18] among others.

Despite most work focus on a document-level analysis to identify and extract sentiment from text, a deeper level of analysis has been explored in the literature. It is the case of works like [17] that explore subjectivity in a sentence-level, or [26,27] that use machine learning techniques to identify phrase-level sentiment, [13,24] exploring statistics and semantic relations to access sentiment in a word-level case.

Works as [3,15] show that in the sub-sentential treatment of sentiment, the semantic structure of the sentence plays an important role in determination of polarity, specially the treatment of negation.

Work on Sentiment Analysis with Twitter data has only recently begin to flourish in literature, due to its potential applicability to market analysis and other areas. Some of these focus on a bag-of-word and n-gram approach to classify tweets according to their subjective content.

Early papers on Twitter processing as [8,16] use n-gram and POS features, and simple models of negation - as a valence shifter with limited scope - to feed different classifiers to predict the polarity of a tweet.

Davidov et al.[4] and Kouloumpis et al. [14] use similar approach based on traditional features for sentiment analysis and Twitter specific feature - such as hastags, links and emoticons - to predict the tweet polarity.

For the Portuguese language, the Twittómetro, a tool for analyzing opinion about candidates for the Portugal's presidential campaign, is discussed in [21]. The authors execute the sentiment analysis using a lexicon-based approach combined with a lexical-syntactic patterns to identify and compose sentiments expressed in tweets.

Other work for sentiment analysis in Twitter setting for Portuguese language are [2] - which relies on an association discovery over lexical features for classifying tweets - and [20] - which relies on detecting bias for a user towards a particular topic.

The automatic construction of training data is also an interesting topic, when dealing with Twitter. Many works rely on information specific to microblogging texts, such as emoticons and hashtags. For example, [8,16] use emoticons to provide polarity information for the tweets - an approach that has proven to be unreliable, by our analysis. [4] use the hashtags as source of polarity, but in their case the hashtags were the sentiment class attributed to the tweet, not a polarity value. Other approach is to use Amazon Mechanical Turk[2] - or similar services - to annotate tweets about polarity [5].

Our work is related to [21], since we intend to perform sentiment analysis in Twitter data for the Portuguese language by a lexicon-based approach. Our work, however, does not rely on a defined domain - and thus on domain-specific resources as that one.

[2] `https://www.mturk.com/mturk/welcome`

3 Sentiment Analysis for Twitter Data

When dealing with micro-blogging data an extra effort must be made to pre-process it, since well-established techniques do not perform well on this new source of data. Gimpel et al. [7], for example, discuss that in this context, problems like tokenization and POS tagging, are much more difficult to deal with than in normal English texts.

This may have a major impact on the method used as, for example, Part-of-Speech features, which are important for Sentiment Analysis in traditional data, but has proven to have little impact on Twitter texts [14].

Most of the work on Sentiment Analysis deal with this problem in a rather ad-hoc way, not paying much attention to it. We propose the use of heuristics to perform a lexical normalization on the data. To define our heuristics, a set of 500 Portuguese Twitter messages were collected and analyzed. Those texts have no intersection with the ones we use in the experiment described further in the work.

The most common case of variation found was the repetition of vowels to denote intensity, which can be easily treated. Another common case of lexical variation in the texts is misspelling of words based on phonetic similarity. To solve this, we applied a variation of the metaphone algorithm for Brazilian Portuguese. More difficult cases, however, are the ad-hoc abbreviations commonly made by the users. A regular abbreviation pattern is the omission of vowels. Thus, words as "saudade" (to miss something/someone) are written as "sdd". It can be corrected by comparison of words by consonants. This heuristic, however, must be applied only in specific cases, as when no vowels are found in the word, since many errors can be performed by applying it.

We used a lexicon-based sentiment analysis method and evaluate the use of the heuristics and of different models to the scope of negation in our experiments. The sentiment lexicons used were the Sentilex lexicon [22] of adjectives composed of around 6000 annotated human predicatives and a expanded version of the OpLexicon [23], a domain-independent sentiment lexicon for the Portuguese language.

The negation modeling, as discussed by [3,15,25], is of major importance when dealing with sentential and sub-sentential sentiment analysis. In our method, we model negation by a set of negation-indicators which are words and expressions - such as "não" (no, not), "nunca" (never), "ninguém" (no one) - indicating intrasentential negation and a defined scope in the sentence.

The negation-indicators act as polarity shifters of expressions within their scope. An important fact to remember is that the Portuguese language allows three kinds of verbal negation - pre-verbal, post-verbal and a double negation both pre- and post-verbal [19].

Lastly, the polarity of a tweet is identified by the algebraic sum of the polarities, after applying the negation treatment. The internal structure of the sentence may provide more useful information for sentiment analysis, based on the discursive relations it encompass such as discussed by [1]. We do intend to analyze those features in the future in a heuristic-based manner, similar to [6].

4 Experiment and Results

In this section we present our experiments on twitter sentiment analysis. First we describe corpus and lexicon, then we evaluate two negation models, our preprocessing technique and differences based on the choice of lexicon.

4.1 The Corpus

The corpus used as test set for our experiments is composed of 1700 tweets balanced between two classes - positive and negative - and was automatically built using the Twitter API. To collect solely tweets written in Portuguese we used the language parameter choosing the 'pt' value - set language to Portuguese.

As query and to classify the tweets according to their polarity, we used the Twitter hashtags #win and #fail - for positive and negative polarities, respectively. Our approach relies on two special hashtags - the #win and #fail - which denote a positive and a negative sentiment, respectively.

Tweets containing both #win and #fail hashtags were discarded. From the total pool collected on a 3-day process, we randomly selected 1700 tweets - divided by polarity - 540 of which presents negation of terms.

4.2 The Lexicon

We use in this work the OpLexicon - a sentiment lexicon for the Portuguese language built using multiple sources of information [23]. The lexicon is constituted of around 15,000 polarized words classified by their morphological category annotated with polarities positive, negative and neutral. It has been enlarged by re-applying the corpus-based method using a bigger corpus and extracting polar verbs using the thesaurus-based method. Also, the adjectives present in the SentiLex which were not annotated in the OpLexicon were included in the later.

4.3 Results

We implemented two models of negation scope: one based on a pre-fixed window - we used a 5-word window (NegWind), and one with the negator's scope over the entire sentence(NegSent). One method that disregards the effects of the negation (NoNeg) was also implemented aiming to evaluate the impact of the negation on this type of texts.

The Table 1 summarizes the results achieved by our method presenting the accuracy and F-measures ($\beta = 1$, F1) values for the positive and negative classes.

Note that, the Sentential model of negation seems the most adequate, which indicates that the 5 word window is constraint too rigid. Taking negation - in the sentential scope - into consideration has increased the precision which indicates a better modeling of opinion identification. The drop in the recall, on the other hand, may indicate that - since the scope of a negator is too broad - negation may have been applied on polarized terms not syntactically related to the negator.

Table 1. Results using different models for negation

	pos			neg		
	Prec	Rec	F1	Prec	Rec	F1
NoNeg	0.62	0.50	0.55	0.67	0.30	0.42
NegSent	0.66	0.46	0.54	0.74	0.33	0.45
NegWind	0.61	0.5	0.55	0.67	0.29	0.40

We also decided to evaluate the impact of our pre-processing technique in the results by applying the same method without lexical normalization and also we evaluated the performance of our proposed polarity lexicon, OpLexicon, against other available lexicon for Portuguese - the Sentilex lexicon. The results may be seen in Table 2.

Table 2. Results without pre-processing and using the Sentilex lexicon

	pos			neg		
	Prec	Rec	F1	Prec	Rec	F1
Without PP	0.66	0.46	0.54	0.74	0.33	0.45
Sentilex	0.52	0.22	0.31	0.58	0.19	0.29

We can see that the pre-processing technique had minor or no effect on the performance. The OpLexicon, however, yielded higher rates than SentiLex. This is not surprising at all since the SentiLex is built for specific domain analysis and was used to enrich the OpLexicon.

5 Discussion and Conclusions

The better performance obtained by the Oplexicon (Table 1) is due to the fact that, unlike the SentiLex, it is composed by different types of words and not just adjectives as that one. Besides, the SentiLex was constructed for a specific domain and, for that reason, it may yield a low performance when used in other domains. This is not surprising since [23] reports similar results about Sentilex and Oplexicon and since Sentilex was used for the enlargement of our version of the Oplexicon.

The pre-processing techniques seem to have no impact on the results. We believe that the set of heuristics is yet quite small and simple, which is probably the main reason for such a result. Other methods of lexical normalization, e.g. [11], may be applied to improve the performance. Our work is, however, the first work on Sentiment Analysis for Twitter, in our knowledge, that evaluates the impact of the preprocessing techniques applied.

It is important to note that the language detection of the Twitter API is not perfect. From the texts which the system could not classify either as positive nor negative - due to lack of opinion-bearing words from the lexicon - for the

OpLexicon using the NegSent strategy, around 15% where texts written in the Spanish language. Besides, about 5% from those which could not be classified, were spam using the #fail hashtag in its text.

The low impact of the negation may be explained firstly by the high coverage of the lexicon and by the specificities of the data, as discussed. We intend to explore further different models of negation, the inclusion of other negators and the best scope of negation for this source of data. Our model of negation, nevertheless has increased the precision of the classification - indication a better modeling of opinion.

Finally, note that the accuracy of our system is yet quite below others, as the Twittómetro. This is because the polarity identification method relies solely on the algebraic sum of the polarities and a simple model of negation. We intend to further explore methods for tweet classification including pseudo-syntactic and discursive information. We also plan to propose a sub-sentential entity-centered method for sentiment analysis for Twitter messages.

References

1. Asher, N., Benamara, F., Mathieu, Y.: Appraisal of opinion expressions in discourse. Lingvisticæ Investigationes 31.2, 279–292 (2009)
2. Calais Guerra, P.H., Veloso, A., Meira Jr., W., Almeida, V.: From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, pp. 150–158. ACM, New York (2011), http://doi.acm.org/10.1145/2020408.2020438
3. Choi, Y., Cardie, C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: EMNLP 2008, pp. 793–801. ACL, Stroudsburg (2008)
4. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: COLING 2010, pp. 241–249. ACL, Stroudsburg (2010)
5. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: CHI 2010, pp. 1195–1198. ACM, New York (2010)
6. Ding, X., Liu, B., Zhang, L.: Entity discovery and assignment for opinion mining applications. In: KDD 2009, pp. 1125–1134. ACM, New York (2009)
7. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: ACL 2011 (Short Papers), pp. 42–47. ACL (2011)
8. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. Entropy, 17 (2009)
9. Golder, S.A., Macy, M.W.: Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science 333(6051), 1878–1881 (2011)
10. Grefenstette, G., Qu, Y., Shanahan, J.G., Evans, D.A.: Coupling niche browsers and affect analysis for an opinion mining application. In: RIAO 2004, pp. 186–194. CID (2004)
11. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a #twitter. In: ACL-HLT 2011 (2011)
12. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11), 2169–2188 (2009)

13. Kamps, J., Marx, M., Mokken, R.J., de Rijke, M.: Using WordNet to measure semantic orientation of adjectives. In: LREC 2004 (2004)
14. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Artificial Intelligence, pp. 538–541 (2011)
15. Moilanen, K., Pulman, S.: Sentiment composition. In: RANLP 2007, Borovets, Bulgaria, pp. 378–382 (2007)
16. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Computer, 1320–1326 (2010)
17. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL, pp. 271–278 (2004)
18. Riloff, E., Wiebe, J., Phillips, W.: Exploiting subjectivity classification to improve information extraction. In: AAAI 2005 (2005)
19. Schwenter, S.A.: The pragmatics of negation in Brazilian Portuguese. Lingua 115(10), 1427–1456 (2005)
20. Silva, I.S., Gomide, J., Veloso, A., Meira Jr., W., Ferreira, R.: Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information, SIGIR 2011, pp. 475–484. ACM, New York (2011), http://doi.acm.org/10.1145/2009916.2009981
21. Silva, M.J., Team, R.: Notas sobre a realizao e qualidade do twitómetro. Tech. rep., University of Lisbon, Faculty of Sciences, LASIGE (May 2011)
22. Silva, M.J., Carvalho, P., Costa, C., Sarmento, L.: Automatic expansion of a social judgment lexicon for sentiment analysis. Technical Report TR 1008 University of Lisbon Faculty of Sciences LASIGE (2010)
23. Souza, M., Vieira, R., Busetti, D., Chishman, R., Alves, I.M.: Construction of a portuguese opinion lexicon from multiple resources. In: STIL 2011, Cuiabá, Brazil (2011)
24. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ACL 2002, pp. 417–424. Association for Computational Linguistics, Morristown (2002)
25. Wiegand, M., Balahur, A., Roth, B., Klakow, D.: A survey on the role of negation in sentiment analysis. Imagine, 60–68 (July 2010)
26. Wilson, T., Wiebe, J., Hwa, R.: Recognizing strong and weak opinion clauses. Computational Intelligence 22, 73–99 (2006)
27. Wu, Y., Zhang, Q., Huang, X., Wu, L.: Phrase dependency parsing for opinion mining. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1533–1541. Association for Computational Linguistics, Singapore (2009), http://www.aclweb.org/anthology/D/D09/D09-1159