SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
VOLUME 3, ISSUE 3
ISSN 2054-7390

Combining Overall and Target Oriented Sentiment Analysis over Portuguese Text from Social Media

[1]José Saias, [2]Ruben Silva, [3]Eduardo Oliveira and [3]Ruben Ruiz
[1]Universidade de Évora, Portugal;
[2]Cortex Intelligence, Portugal;
[3]BizDirect, Portugal
jsaias@uevora.pt; ruben.silva@cortex-intelligence.com; eduardo.oliveira@bizdirect.pt; ruben.ruiz@bizdirect.pt

ABSTRACT

This document describes an approach to perform sentiment analysis on social media Portuguese content. In a single system, we perform polarity classification for both the overall sentiment, and target oriented sentiment. In both modes we train a Maximum Entropy classifier. The overall model is based on BoW type features, and also features derived from POS tagging and from sentiment lexicons. Target oriented analysis begins with named entity recognition, followed by the classification of sentiment polarity on these entities. This classifier model uses features dedicated to the entity mention textual zone, including negation detection, and the syntactic function of the target occurrence segment. Our experiments have achieved an accuracy of 75% for target oriented polarity classification, and 97% in overall polarity.

Keywords: Sentiment Analysis; Opinion Mining; Text classification; Machine Learning; Natural Language Processing.

1 Introduction

Microblogging and social networks are used by people of all ages. These platforms offer a new form of Web based socialization, simplifying communication to restricted groups or to crowds. They aggregate user-generated content, such as opinions that people write and publish online, and are now valued for market research and trend analysis.

In this paper we describe the use of Natural Language Processing (NLP) techniques for Sentiment Analysis (SA) of social media texts, in Portuguese, sensing the overall and the target oriented sentiment polarity. To automatically extract such information from text, it is necessary to deal with the challenges of natural language, plus the present-day web writing style, full of symbols, tags, abbreviations and misspellings. Given a post or tweet text, the system must find the overall polarity and also the polarity of any specific entity mention. Figure 1 has three examples, of increasing complexity, with the original text in Portuguese, and the respective translation. In the first case we must detect negative polarity in the overall sentiment. For 1(b), beyond the overall sentiment it is necessary to detect the reference to benfica, and that it is positive. The third example is more complicated, because the second and third

DOI: 10.14738/tmlai.33.1297
Publication Date: 26th June, 2015
URL: http://dx.doi.org/10.14738/tmlai.33.1297

sentences both mention two entities, and with opposite polarities. In each sentence, Belenenses has a negative polarity, while Benfica is referred with positive polarity.

The system being described here is a continuation of previous work on overall polarity classification [1] and aspect based sentiment analysis [2] with English texts. We use a supervised machine learning classifier for both overall and target oriented sentiment polarity, with different tuning in feature extraction. Opinion target entities are automatically detected with named entity recognition, complemented with an entity catalog. The analysis pipeline is explained in detail in section 3.

Fiquei aborrecido com o exame nacional.
(in English: "*I was disgusted with the national examination.*")
(a)

Finalizada a época , o benfica é vencedor
(in English: "*Finished the season, the winner is benfica*")
(b)

Adorei o jogo. Benfica 3-1 Belenenses... O Benfica arrasou o Belenenses.
(in English: "*I loved the game. Benfica 3-1 Belenenses... Benfica smashed Belenenses.*")
(c)

Figure 1: Different examples of text to be analyzed: without entities (a); with a single entity (b); and having more than one entity (c).

2 Related Work

There are many works in the SA field, from industry and from academia, which differ in the specific objectives, in their subdomain of expertise, on the language in which they operate, or in the approach taken for polarity classification.

Twitómetro [3] is a system to gauge sentiment towards five political leaders, via Twitter, during the campaign for the 2011 portuguese elections. His politics domain predecessor was OPTIMISM [4], an opinion mining system using an ontology of political entities to assist in entity recognition, and whose polarity classifier combines rules on lexical-syntactic patterns with machine learning. POPSTAR is a research initiative on public opinion and sentiment tracking, producing indicators on entity mentions frequency, and their polarity, over time. Several sub-projects emerged from there, particularly on reputation as described in [5]. Besides Twitter, news feeds are also used for opinion mining, as in [6], through an UIMA-based pipeline for SA, in Portuguese.

Modern systems and SA's state of the art can be seen in NLP Workshops and competitive challenges, such as RepLab [7], SemEval Twitter SA [8], and aspect based SA tasks [9,10]. These events attract participants around the world, but most focused in the English language. One of the top participating systems in SemEval aspect based SA task, in 2014, was the NRC-Canada system [11], which extended the base training corpus with additional customer reviews corpora, and used Stanford CoreNLP to perform tokenization, POS tagging, and dependency analysis. Polarity classification was dealt with a linear SVM classifier, having features for the target and its surrounding words, POS based features, dependency tree based features, unigrams and bigrams, and lexicon based features.

3 Method

In this paper, we assume the existence of a posts and tweets collector module. Our system has a REST API, where the content of those publications can be sent for analysis.

3.1 Underlying Platform and Preprocessing

This work reuses the basic framework of a recent real-time SA system [2] for English texts. Our system is developed in Java, using the tool MALLET [12], a package for statistical natural language processing and machine learning applications to text. Jersey[1] RESTful Web Services framework was used in the system frontend, for making the core functionality available as a service.

The received input is preprocessed through noise removal, tokenization, POS tagging and lemmatization. Data representing social media content is rich in metadata tags and hyperlinks. These noisy parts may hinder the automatic understanding of sentences. Thus, in preprocessing we remove certain elements such as URL addresses and retweet prefixes. However, other elements cannot be removed at this point, because they are potential indicators for polarity, as in the case of hashtags. Instead of using MALLET's default tokenization pipe, we implemented a new tokenization and POS tagging module, based on Apache OpenNLP[2] library, with Portuguese trained models. The lemmatizer is a proprietary software that depends on the POS tag and the textual context of words.

3.2 Overall SA

For this phase, the system must search for sentiment clues on the message transmitted globally by the text. To determine the general sentiment polarity we used a machine learning solution, with a supervised approach and Maximum Entropy classifier, through MALLET classification libraries. To build the model, the classifier was trained from a set of 59000 labeled instances with texts on popular expressions, and online comments on music, festivals, television, sports and politics domains. Two sentiment lexicons are used to assist in feature extraction. The first is SentiLex-PT [13], a sentiment lexicon for Portuguese, made up of 7014 lemmas, and 82347 inflected forms of verbs, nouns, adjectives and idiomatic expressions. The other lexicon is a complementary polarity table, contained in a linguistic knowledge base from a previous work [14]. It has evolved with the gradual introduction of new Portuguese expressions, including idioms, but also popular English expressions, Internet jargon, and common symbols and abbreviations. Our classifier model is then trained with the following features:

- Bag-of-Words (BoW) on lemmas: instead of counting the occurrences of the original text on each token, we consider the frequency of their respective lemmas. To illustrate, let's consider the sentence from the example in Figure 1(a). Some of its features would be: `ficar=1, aborrecer=1, com=1, o=1, exame=1, nacional=1`.
- A pair (POS tag, simple polarity), for each token, and the counting of these values. The second token in the example would have the feature `(v-fin.negative)=1`.
- Bigrams of pairs (POS tag, simple polarity). Like the previous feature, but on each consecutive token pair. The first value in the given example would be `(v-fin.neutral,v-fin.negative)=1`.

[1] https://jersey.java.net
[2] http://opennlp.apache.org

- Trigrams of pairs (POS tag, simple polarity), as before but over three consecutive tokens.
- Bigram before/after polarized terms (positive or negative), according to each sentiment lexicon. If a sentiment lexicon identifies a token T as positive or negative, we generate two features: one with the previous word and T, and another with T and the next token's text, all in lemma form. In the example, it would be `ficar. aborrecer` and `aborrecer. com`.
- Subject/object polarity, if a sentiment lexicon determines the polarity that some expression originates, on the subject and on the object, inside that sentence. As example, the verb *defeated* is positive for the subject, but negative for the entity in the object.
- Presence of terms with positive/negative polarity within the last five tokens. Because sometimes the last words summarize the main idea or polarity.
- Balance of polarity, according to each sentiment lexicon, calculated by the total of positive expressions minus the total of negative expressions, considering also denial detection.
- Bigrams after verbs, and after negation terms, using the lemmatized forms.

Figure 2 shows the result of processing the text in Figure 1(a), with the negative polarity shown in the `overallPolarity` field. If this field is zero, the polarity is neutral; if the value is less than zero, we have negative polarity; and a value greater than zero corresponds to a positive polarity. The remaining two fields denote the absence of entities in the text, and will be explained in the next section. The service output can be in JSON or XML format.

```
{
  "overallPolarity" : -0.9822897,
  "targetCount" : 0,
  "targetPolarityList" : [ ]
}
```

Figure 2: Overall result for sentence in Fig. 1(a).

3.3 Target Oriented SA

The system starts by identifying entity mentions on the text. These mentions can be opinion targets, and when they are not, the system should give them neutral sentiment classification. Our process to detect target entities comprises a named entity recognition (NER) module, complemented by the use of an entity catalog. For NER, we use an OpenNLP classifier with a model trained for Portuguese. Entities whose categories are currency, time, numeric or abstract are discarded. The most plausible, in categories person, organization, brand, and location, are selected. The entity catalog is an inventory whose records contain the entity canonical name, possible name aliases, and the entity type. This resourse allows the system to realize that, for example, SCP is an alias or alternative designation for Sporting Clube de Portugal, an organization.

In the next step, the system must assign a sentiment polarity (positive, negative or neutral) to every detected entity mention, according to the text content. For such, we prepared a second Maximum Entropy classifier, now based on different features. In this supervised learning, the training labeled instances do not include only the text content, as before. Each instance also has the target entity, and there is an indication of where it is mentioned. Due to the added complexity in corpus annotation, this

training set is smaller than the used for the overall sentiment classifier. Here we have only 13100 instances. We seek the sentiment for each specific entity mention. So for these instances we want features related to that same mention, probably more confined to a short text area. The features for the target oriented classifier are:

- Bag-of-Words for the mention's text area. We define a feature for the original text of each token inside the short sentence that includes the entity mention. The entity name is replaced by TARGET , to facilitate harmonization of cases of sentences with similar structure but in which the opinion focuses on different names. Taking the example shown in Figure 1(b), these first features are [o=1, TARGET=1, é=1, vencedor=1] .
- Lemma bigrams for tokens within the metion's text area. For the same mention, it would be: [m.bigram_o.TARGET=1, m.bigram_TARGET.ser=1, m.bigram_ser.vencedor=1].
- Syntactic function associated with the target. When possible, indicate whether the target appears in the subject or object, according to sentence structure.
- Subject/object polarity. As before, if a sentiment lexicon determines the polarity that some expression originates, on the subject or on the object part, we create a feature for it. Returning to the example, *vencedor* is an adjective with positive polarity to the subject, which in this case is *benfica*.
- Lemma bigrams, after the target, and before the target mention.
- Pairs and bigrams and trigrams of (POS tag, simple polarity) pairs, as before, for the full text.
- Bigrams before/after polarized expressions, as before, across the full text.


```
{
  "overallPolarity" : 0.98907655,
  "targetCount" : 1,
  "targetPolarityList" : [
  { "target" : "benfica",
    "polarity" : 0.8114217,
    "countPositiveRefs" : 1,
    "countNegativeRefs" : 0,
    "countNeutralRefs" : 0,
    "targetReferencesOverDoc" : [
       { "referencePolarity" : 0.8114217,
         "from" : 23, "to" : 30,
         "sentenceNumber" : 0 } ]
  } ]
}
```


Figure 3: Analysis result for the example in Fig. 1(b)

In Figure 3 we can see the detailed output returned by the system, result of analyzing the text in Figure 1(b). In targetCount field we have the number of entities mentioned in the text. Then we have a list with the sentiment polarity for each target entity. In this case we have benfica, an entity with positive polarity, and one (positive) reference, which takes place in the first sentence (by sentenceNumber field), more precisely in the text between the position 23 and the position 30.

Having all this detail in the answer, we can provide a friendly visual output. With the polarity and the precise location of the target entity, we can assign colors to facilitate the interpretation of the results, as shown in Figure 4.

Figure 4: Visual output of the analysis for Fig. 1(b)

Sometimes we may have different opinions on the same document, or even in the same sentence, on the same entity or not. This is the case of the example shown in Figure 1(c). The JSON code with our system's output for such example is listed on Figure 5. The targetCount field shows us that there are references to two entities. Belenenses is referred to twice, in the second and in the third sentences, and both times having negative polarity, resulting in a target polarity value of -1.47. Benfica is also referred to twice, but with positive polarity. With four entity mentions, this is an example where the visual output is clearly easier to read. The result for this case is shown in Figure 6.


```
{
  "overallPolarity" : 0.28949898,
  "targetCount" : 2,
  "targetPolarityList" : [
  { "target" : "Belenenses",
    "polarity" : -1.4753892,
    "countPositiveRefs" : 0,
    "countNegativeRefs" : 2,
    "countNeutralRefs" : 0,
    "targetReferencesOverDoc" : [
       { "referencePolarity" : -0.5799957,
         "from" : 12, "to" : 22,
         "sentenceNumber" : 1 },
       { "referencePolarity" : -0.89539355,
         "from" : 20, "to" : 30,
         "sentenceNumber" : 2 } ]
  },
  { "target" : "Benfica",
    "polarity" : 0.92901266,
    "countPositiveRefs" : 2,
    "countNegativeRefs" : 0,
    "countNeutralRefs" : 0,
    "targetReferencesOverDoc" : [
       { "referencePolarity" : 0.09627137,
         "from" : 0, "to" : 7,
         "sentenceNumber" : 1 },
       { "referencePolarity" : 0.83274126,
         "from" : 2, "to" : 9,
         "sentenceNumber" : 2 } ]
  } ]
}
```


Figure 5: Analysis result for the example in Fig. 1(c)

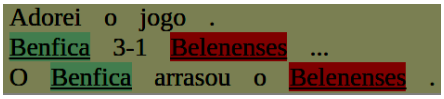

Figure 6: Visual output of the analysis for Fig. 1(c).

3.4 Model Improvement

The system can evolve, by adjustments in the framework that contribute to a faster running, and improvements in the classification model, for better performance in terms of accuracy and precision. The priority is the second case, on system output quality. For the detection of entities, the opinion

targets, the main NER tool may be replaced, if necessary. And minor flaws can be overcomed by the introduction of new entries in the supplementary entities catalog.

For this system's most important component, the sentiment polarity classification, we chose two aspects to monitor: the adequacy of the model features, and the training set. By studying the output of the system, we seek clues to distinguish the characterization of mislabeled instances. Observing their respective text, we can add features on the sentences structure, about new symbols, or context aspects based. On the other hand, and in parallel, the training set will grow, being added new annotated instances, both for overall sentiment and target oriented model tuning. For this purpose, we set up an interface for collecting feedback, which facilitates marking the sentiment polarity. This way, on the next model iteration, these new instances are already used to train the classifier. Figure 7 illustrates the collection of feedback on the third sentence of the last example. The user may indicate the sentiment polarity regarding Belenenses, leading to a new labeled instance. Also, when collecting this feedback, if a marked target entity was not yet known, it will be added to the system catalog.

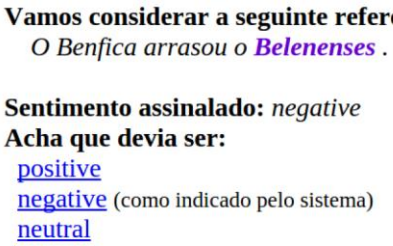

Figure 7: Web interface for gathering feedback and corpus annotation

4 Results

The perception of the performance in the main components of a system is critical, because only then we can realistically improve the service. To evaluate our classification model, on both analysis types, we have used a k-fold cross-validation method. The labeled instances are partitioned into k subsets. Then there are k rounds of evaluation, in which, and in turn, each of the k instance sets is used to test the classifier trained with the other k-1 sets. At the end of the process, all subsets were used only once for testing, and their results are combined. We used the typical 10-fold evaluation, which means that each training round has 90% of the instances. Table 1 has the accuracy values, for both overall and target oriented SA. It is a general measure for success on the predicted polarity. The underlying set of labeled instances is not equal in both cases. In overall SA we have more instances, but fundamentally with short texts, often with a single short sentence, for which the classifier worked fine.

Increasing the evaluation detail, we calculated the precision, recall and F-measure, for each polarity class, and these metrics' results are shown in Table 2. The overall polarity classifier has the top performance, reaching 98% precision and recall for the negative class, and slightly less in the positive class. The target oriented classifier achieved poorer results, particularly on the effectively polarized classes (positive and negative). Its best precision and recall were obtained in neutral class, being noticed, on the other side, a weak coverage for positive class, with only 64%.

Table 3 shows the weight of each polarity class in training, for both the classifiers. Considering also the information from the previous table, we notice, as expected, that in classes with more training instances

the evaluation results are better. An important part in target oriented SA is the detection of the sentiment target entities. When evaluating our entity recognition method with the annotated corpus we used for polarity training, the accuracy is 88%.

Table 1: Accuracy for sentiment polarity.

Analysis	accuracy
overall	0.97
target oriented	0.75

Table 2: Detailed evaluation result.

Analysis	class	precision	recall	f-measure
overall	positive	0.96	0.96	0.96
	negative	0.98	0.98	0.98
	neutral	0.78	0.77	0.78
target oriented	positive	0.67	0.64	0.65
	negative	0.75	0.75	0.75
	neutral	0.77	0.80	0.78

Table 3: Polarity class weight in the training instances.

Analysis	positive	negative	neutral
overall	19%	72%	9%
target oriented	21%	40%	39%

5 Conclusion

We described a sentiment analysis system for social media content, thought to classify both overall sentiment, and sentiment towards specific targets mentioned in the text. Despite the good accuracy, in development time, of our overall sentiment classifier, its use in post-development period reveals more errors. We think this is due to the differences between training instances and these recent input texts, which have greater length and a more complex structure than the former. Portuguese corpus for overall sentiment usually do not have many long texts.

As in other NLP tasks, small errors in the modules that carry the first part of the analysis can compromise the quality of the final classification result. Establishing a comparison with our previous experiences in English language target oriented SA [2], in this work we got 3% less accuracy. But in English we have more tools to work the text, and more labeled data resources, than those available for Portuguese.

As future work, we plan to increase the size of the corpus annotated for training, as a measure to mitigate the difference in results between classes. At the same time, we will continue to experiment with new features, looking to improve precision and recall. Apart from the performance of our current system, there is a feature that could be introduced: aspect classification. The analysis result would be richer, pointing out a particular aspect of the target that is affected by some sentiment polarity. In the case of comments on some restaurant, possible aspects could be the price or the food. And for each of them we could then examine the sentiment polarity, independently.

Although there still is a significant error rate, of approximately 25% for target oriented SA, this tool can greatly facilitate the analyst work. Let us consider a collection of posts or tweets, where only 40% are not neutral. On average, to find 40 positive or negative opinions, a human would have to read 100 documents, whereas using a system like the one presented in this paper, the human analyst would only have to filter the classification results. Even on the worst case, the gain is that the work will be halved.

ACKNOWLEDGMENTS

This project was approved under the "SI ID&T – projeto individual", and according to AAC nº 07/SI/2012, project number 38601.

REFERENCES

[1] J. Saias, Senti.ue: Tweet overall sentiment classification approach for SemEval-2014 task 9. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, August 2014. ISBN 978-1-941643-24-2, p. 546–550.

[2] J. Saias, Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Colorado, USA, 2015. ACL.

[3] M. J. Silva et al., Notas sobre a Realização e Qualidade do Twitómetro. Technical Report. University of Lisbon, LASIGE. 2011.

[4] M.J. Silva et al., The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. In New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence, 2009, p. 565-576.

[5] J. Filgueiras and S. Amir, POPSTAR at RepLab 2013: Polarity for Reputation Classification. In Proceedings of the 4th International Conference of the CLEF initiative, CLEF 2013, Valencia, Spain.

[6] P. Lambert and C. Rodriguez-Penagos, Adapting Freely Available Resources to Build an Opinion Mining Pipeline in Portuguese. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Iceland, 2014.

[7] E. Amigó et al., Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management. In Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science, Volume 8685, 2014, p. 307-322.

[8] S. Rosenthal et al., SemEval-2014 Task 9: Sentiment Analysis in Twitter. In Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval'14). August 23-24, 2014, Dublin, Ireland.

[9] M. Pontiki et al., SemEval-2014 Task 4: Aspect Based Sentiment Analysis. Proceedings of the 8th SemEval, Dublin, Ireland. 2014.

[10] M. Pontiki et al., SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, USA. 2015.

[11] S. Kiritchenko et al., NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland, 2014, p. 437–442.

[12] A. K. McCallum, MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet.cs.umass.edu

[13] M. J. Silva et al., Building a Sentiment Lexicon for Social Judgement Mining. In Lecture Notes in Computer Science (LNCS) / Lecture Notes in Artificial Intelligence (LNAI), International Conference on Computational Processing of Portuguese (PROPOR), Coimbra, 2012.

[14] M. Mourão and J. Saias, BCLaaS: implementação de uma base de conhecimento linguístico as-a-service. In L. Ferreira and V. Pedro, editors, Actas das 3as Jornadas de Informática da Universidade de Évora. ECT, Universidade de Évora, Portugal, 2013.

