

Detailing Sentiment Analysis to Consider Entity Aspects: an Approach for Portuguese Short Texts

¹José Saias, ²Mário Mourão, ³Eduardo Oliveira

¹Universidade de Évora, Portugal;

²Decsis, Portugal;

³Digitmarket, Portugal;

jsaias@uevora.pt; mario.c.mourao@decsis.pt; eduardo.oliveira@bizdirect.pt

ABSTRACT

Sentiment analysis is useful for identifying trends, or for discovering user preferences, which can later be applied to campaign targeting or recommendations. In this paper, we describe an approach to classify the sentiment polarity regarding aspects, and how this technique was used in a previous system, for short texts in Portuguese, giving it greater sensitivity to detail.

Aspect extraction is done by locating candidates for aspect as expressions having a relationship with the entity and possibly some polarized term, through rules based on POS tags. For each aspect, the sentiment polarity is determined by a Maximum Entropy classifier, whose features depend on the entity mention, on the aspect and its support text, including negation detection, bigrams, POS tags, and sentiment lexicon-based polarity clues. For aspect sentiment, our classifier evaluation indicated a precision of 68% for the positive class and 73% for the negative class, with the dataset used in our research.

Keywords: Sentiment Analysis, Machine Learning, NLP, Text classification.

1 Introduction

Assessing the level of satisfaction, about a product or entity, is highly valued in marketing. Such information is important for those who decide the direction of a campaign, or to assist in brand or celebrity online image/reputation management, for example. The marketing sector is naturally aware of the phenomena of online social networks and studies the efficiency of advertising in its target users, and fine-tunes criteria for campaigns according to analysis on their data [1].

Sentiment Analysis (SA), sometimes referred to as Opinion Mining, deals with the search of opinion in text, automatically detecting evidences of sentiment with a certain polarity (positive, negative or neutral) [2]. These sentiments can be abstract, or not directed, evidencing something only about the state of mind of the author (E.g.: "*This work makes me sick!*"). They can also be concrete, and focused on a target entity, or some specific aspect of that target (e.g.: "*Restaurant X has fantastic food but prohibitive prices!*").

SA can help at various levels. On one hand, facilitating the identification of trends, pointing out what everyone is liking (or not liking), or who/what is extremely popular (or unpopular), at a given day. In addition, SA can help to create user-specific models by identifying things he likes, his preferences, or aspects he does not like, important elements for tasks as client clustering or recommendation systems.

This paper describes an approach for sensing the overall and the target oriented sentiment polarity over short texts from social media, in Portuguese. Our work is related to the SmartSeg project, which aims to create an intelligent targeting engine that can help marketers define target segments for campaigns. In addition to global polarity and polarity directed to an entity, a greater level of detail is now required, with the identification, whenever possible, of the target entity aspects on which an opinion is expressed. Our focus is an extension to an earlier SA system [3], adding aspect polarity classification, based on previous work by the first author on SA over English text [4, 5].

The following section mentions some recent related work. Section 3 describes how our system works, with more detail on how aspects are handled. Some preliminary results are given on Section 4, and the paper ends with remarks and conclusions in section 5.

2 Related Work

We can easily find a study on SA if we do a web search. However, finding an SA system able to detect different aspects of the same entity and classify their corresponding sentiment, is not easy, and even more if the goal is to work with Portuguese written content.

A doctoral thesis by Nádia Silva [6], in 2016, reports a study on the use of classifier ensembles to SA, feature engineering, and approaches to semi-supervised learning to overall polarity classification. On the Kaggle platform, we can find systems for SA, such as the one Leandro Silva developed for Brazilian Portuguese text [7], described with excerpts of code and pedagogical explanations on how to use NLTK, or a Bag-of-Words model for supervised learning, with Naive Bayes Multinomial and Random Forest classifiers on SkLearn. This is a demonstration purpose system, thought only for overall sentiment, not considering the different target entities. An empirical study of techniques for aspect extraction on SA is presented in Pedro Balage Filho's doctoral thesis [8], exploring three different approaches over English and Portuguese corpus.

Repustate [9] is a paid online service for text analytics, including SA for Portuguese texts. Its site mentions over 100 billion documents analyzed every month. Portuguese-specific tools are employed, including a part of speech tagger, a lemmatizer, and Portuguese-specific sentiment models. As far as we know, and as seen in the free demo, this system performs overall polarity classification only.

Conferences related to text processing often have challenges for SA. SemEval-2017 had two of these challenges: Task 4 and Task 5. The purpose in Task 4 was to identify the overall sentiment of a tweet, and the sentiment towards a topic with classification on a two-point and on a five-point polarity scale [10], with English and Arabic languages. For overall sentiment on English language, some teams used Deep Learning and neural network methods such as CNN and LSTM networks. Other supervised learning methods were also used, such as SVM, Maximum Entropy, Logistic Regression, and Random Forest based classifiers. Some systems combined SVM with neural networks. Five of the top-10 scoring systems used ensemble classifiers.

SemEval-2017 Task 5, or Fine-Grained Sentiment Analysis on Financial Microblogs and News, was a challenge that aimed to predict a sentiment score for each company/stock mentioned on text [11]. It was organized into separate subtasks, each with its texts collection: microblogs and news. Most systems used hybrid techniques, with Machine Learning and Lexicon-based approaches (as Jiang et al.'s system [12], which ranked first on subtrack 1), while others used a traditional Machine Learning approach, or a Deep Learning approach possibly combined with lexicon-based techniques.

3 System description

In our previous system [3], we used a supervised Machine Learning classifier for both overall and target oriented sentiment polarity, having a different configuration in each case. This paper describes an extension to that system, which aims to consider the particular entity's aspect about which an opinion is being expressed, and to improve the performance of the sentiment polarity classification.

3.1 Overview

Having as a priority the system sensitivity to aspects, we kept the previous underlying platform and preprocessing stages, coded in Java and MALLET [13], a Machine Learning for Language Toolkit. A REST API is used by the service clients, allowing them to send their requests with short texts, such as tweets or comments on social media. The text received in each request is processed according to the following sequence of steps:

1. Preprocessing: Natural Language Processing techniques, which include the division into sentences and a first level of syntactic analysis
2. Named entity recognition: to identify mentions to persons, brands and other opinion targets
3. Overall sentiment polarity classification: opinion polarity broadly conveyed by the text, even if there is no mention of any entity
4. Target oriented sentiment analysis
 - a. Aspect detection for target entities
 - b. Aspect sentiment polarity classification

Figure 1 shows the system pipeline outcome for the sentence “Évora é bonita.” (In English: “Évora is beautiful.”). It is a JSON structure where we can see the identified aspect “*beleza*” (beauty) and its sentiment polarity. This output can be formatted in JSON or XML.

```
{
  "id" : "2017122709511212",
  "overallPolarity" : 1.0,
  "targetCount" : 1,
  "targetPolarityList" : [ {
    "target" : "Évora",
    "polarity" : 0.7,
    "countPositiveRefs" : 1,
    "countNeutralRefs" : 0,
    "countNegativeRefs" : 0,
    "targetReferencesOverDoc" : [ {
      "aspect" : "beleza",
      "from" : 0,
      "to" : 5,
      "sentenceNumber" : 0,
      "aspectSupportTextTo" : 14,
      "aspectSupportTextFrom" : 6,
      "referencePolarity" : 0.7
    } ]
  } ]
}
```

Figure 1: System output for “Évora é bonita.”

Step 4 of the pipeline, for target oriented SA, was completely reimplemented, and it is described in the following sections. The first three steps (preprocessing, target detection and overall sentiment) are

processed as described in [3]. Preprocessing includes noise removal, tokenization, POS tagging and lemmatization.

The target detection comprises a mixed approach with an OpenNLP classifier for Named Entity Recognition (NER), and an entity catalog. The first has a model trained for Portuguese, and from its outcome, we filter entity categories, by choosing Person, Organization, Brand and Location, while discarding others (currency, time, numeric and abstract). The entity catalog is a lookup table whose entries have the entity canonical name, possible name aliases, and the entity type.

To detect the overall sentiment, we continue to use a supervised Machine Learning approach, in MALLET and a Maximum Entropy classifier. The training set has about 71000 labeled instances, with short texts of posts, tweets or comments on the web about TV celebrities, sports, music festivals and politics.

The overall sentiment classifier model is trained with features for: lemmas, (POS+polarity) pairs for each term, bigrams before/after polarized terms, lemmatized bigrams after verbs and after negation terms, presence of polarized terms in the last 5 tokens. To detect polarized terms or clues, in feature extraction, we use SentiLex-PT [14], a sentiment lexicon for Portuguese, and a complementary polarity table which aggregates a set of Portuguese expressions, including idioms, but also popular English expressions, Internet jargon, and common symbols and abbreviations.

3.2 Aspect based SA

At this point, the target entities have already been tagged throughout the text. In the previous version, we would extract features about the entity mention surrounding text, and, based on those features, deduce the sentiment polarity. This is where we change the procedure to be more meticulous, and to consider entity aspects.

3.2.1 Aspect Detection

As reported by Bing Liu [15], a document where the author gives a positive opinion about an entity or object does not mean that the author always has a positive opinion on all the characteristics of that entity. An aspect is a feature of the target entity. It is something particular about that entity, about which there may be an opinion in the text. For the same target entity, we can have different aspects, each with its sentiment polarity. As an example, in a text relative to a restaurant, there may be a positive opinion about the food, and a negative opinion about the service. In this case we would have only one target entity, the restaurant, and two aspects: the food and the service. Treating polarity over the whole document or sentence may not capture this degree of detail.

We address the aspect extraction as an Information Extraction task, using a relation based method. In the entity's surrounding text, the system looks for expressions having a relationship with the entity, and possibly some polarized term. The relationship between the aspect support expression and the target entity can be identified with the help of a dependency parser and with rules based on the POS tags of neighbouring tokens.

The main extraction rules in our system have the following patterns:

- i. TARGET VTER POL Noun
- ii. Noun FROM TARGET VSER POL
- iii. Noun POL FROM TARGET

- iv. VSER POL ART Noun FROM TARGET
- v. TARGET VSER AQ

In each rule, TARGET is the mention to the entity, and Noun is the candidate aspect. POL is a polarized term, as an adjective with negative or positive sentiment. VTER is a verb semantically compatible with “*ter*” or “*possuir*” (to have). VSER is a verb semantically compatible with “*ser*” or “*estar*” (to be). FROM represents a preposition, such as “*de*”, “*da*”, and “*do*”. ART is a definite article. AQ stands for aspect qualifier, and it is a term or expression that defines the aspect. To detect these, we use a fixed table, whose entries have triples "VERB AQ ASPECT". The example “*ser,barato,preço*” (to be,cheap,price) is one of these triples, which means that if you use the term "cheap" you are referring to "price" aspect.

All the five rules have variants: optional presence of articles, switching the position of POL and Noun, presence of intensity expressions (e.g. adverbs). Another variant is the extraction of aspects in sequence, for texts that describe several attributes of an entity in a single sentence.

Let us consider a concrete example:

“Este Nokia tem um display fantástico mas não tem uma boa bateria.” (In English: “*This Nokia has a fantastic display but does not have a good battery.*”)

At this stage, the system will extract the aspects “*display*” and “*bateria*” (battery) related to entity mention “*Nokia*”. With each aspect, the system also maintains its support text. The intent is to keep the surrounding text including a possible polarized expression.

If near the entity mention we do not find any aspect, then we assume the "general" default aspect, meaning that the opinion is towards the target entity as a whole.

3.2.2 Aspect Sentiment Polarity

In this step, the system will determine the sentiment polarity (positive, negative or neutral) for each aspect found for all detected target entities. In cases where the aspect is global, that is, a more specific aspect has not been extracted, polarity classification is done towards the target entity mention, considering its location in the text, as done in the previous version of the system. In more specific cases, with a concrete aspect, the polarity is focused on that aspect rather than on the entity, and so the system considers the aspect support text first, and then the remaining text.

Sentiment polarity is determined by a Maximum Entropy classifier, whose supervised learning model was now trained with 13600 instances. In each instance, we now needed tags with the target entity, the location of the mention to the target entity, the aspect and the aspect support text, and the sentiment polarity regarding the aspect. As we did not have all this detail in the previous model's training instances, it was necessary to reach a compromise solution, that would allow to take advantage of what already existed, and to add new instances, with the additional annotation tags.

For the first part, backward compatibility, some corrections were made on the previous instances, and some were eliminated because they did not fit properly. In these instances, without aspect annotation, the system assumes the "general" aspect, and therefore trains these cases with the polarity directed to the entity as a whole.

To get new instances, with aspects annotation, we used the outputs from *PolarityGame* [16], a system for building an annotated corpus. That system is based on gamification [17], which is the adoption of game techniques and dynamics in processes that do not have this nature. Such may encourage users to try a system, and instill elements of reward or competition, as a stimulus for the user to continue their participation. Tagging to SA is a time consuming process, so it is not easy to find volunteers. In this way, the use of gamification is an effort to capture some contributions to the annotation task. A *PolarityGame* session includes: insert a short text; tagging it; and validate or fix system tags for another existing instance that is chosen by the system. The scoring scheme of the game is based on the estimated contribution to the system. By contribution, we mean the insertion and tagging of a text that effectively adds value to the knowledge base, and that is supported by cross-validation.

To isolate the polarity of each aspect, before extracting the features, the system defines the classification chosen text, which consists of the cancellation of the remaining aspects. If there are multiple aspects for the same entity, the other aspects' support text is replaced by white space. The aspect polarity classifier's model is built with the following features:

- Bag-of-Words for each token inside the chosen text, and replacing the entity name by TARGET, to simplify pattern processing and be transparent to the variability of names.
- Lemma bigrams for tokens within the chosen text.
- Syntactic function associated with the target. When possible, indicate whether the target appears in the subject or object, according to sentence structure.
- Syntactic function associated with the aspect term.
- 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.
- Lemma bigrams, after and before the target mention.
- Lemma bigrams, after and before the aspect term.
- Lemma bigrams, after and before the aspect support text.
- Pairs and bigrams and trigrams of (POS tag, simple polarity) pairs, as before, for the chosen text.
- Bigrams before/after polarized expressions, as before, across the chosen text.
- Presence of negation terms before the target mention, and before the aspect support text.

In cases where a target entity has more than one aspect, the polarity assigned to the entity is the sum of the polarities of its aspects.

In Figure 2 we can see the detailed output returned by the system for the previous example. In this JSON data we can find an entity "*Nokia*", and the two aspects associated with it: "*display*" and "*bateria*". The aspect support text offset is also set in each case. As we can see in this example, the "*display*" aspect has a positive polarity, while for the "*bateria*" aspect the sentiment is negative.

4 Results

The training set has a class distribution as indicated in Table 1. For both the classifiers, most of the annotated instances belong to the negative class. As in the previous version, there is still a large imbalance between the positive and negative classes.

To evaluate the impact of this system's change with the approach for aspect support, we performed an evaluation similar to the previous one. We used a 10-fold evaluation, which means that each training round has 90% of the instances. The labeled instances are partitioned into 10 subsets. Then there are 10

rounds of evaluation, in which, and in turn, each of the 10 instance sets is used to test the classifier trained with the remaining 9 sets.

```
{
  "id" : "sample0031",
  "overallPolarity" : 0.08,
  "targetCount" : 1,
  "targetPolarityList" : [
    {
      "target" : "Nokia",
      "polarity" : 0.07,
      "targetReferencesOverDoc" : [ {
        "aspect" : "display",
        "from" : 5,
        "to" : 10,
        "sentenceNumber" : 0,
        "aspectSupportTextTo" : 36,
        "aspectSupportTextFrom" : 18,
        "referencePolarity" : 0.78
      }, {
        "aspect" : "bateria",
        "from" : 5,
        "to" : 10,
        "sentenceNumber" : 0,
        "aspectSupportTextTo" : 64,
        "aspectSupportTextFrom" : 53,
        "referencePolarity" : -0.71
      } ],
      "countPositiveRefs" : 1,
      "countNegativeRefs" : 1,
      "countNeutralRefs" : 0
    } ]
  } ]
}
```

Figure 2: Sample output with two aspects for the same entity

Table 2 has the results of the evaluation, for both overall and target+aspect oriented SA, with precision, recall and F-measure metrics per class. As expected, the overall polarity classifier has the best performance, with high values for F1, as seen in the system's previous version. The results of the new aspect based polarity classifier are shown in the second half of the table. The positive class has the lowest values in all three metrics. This already happened in the previous system, but there was a 1% increase in precision, recall and F1, perhaps by the presence of new/more instances in the dataset. In the opposite direction, precision fell in both the negative class and the neutral class by approximately 2%, reaching now 73% on the negative class, and 75% on the neutral class. The best F1 measure in target oriented was obtained in the neutral class with 76%.

Table 1: Polarity class weight in the training instances

Analysis	positive	negative	neutral
overall	19%	72%	9%
target+aspect oriented	22%	41%	37%

Table 2: Sentiment polarity classifier evaluation

Sentiment	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+aspect oriented	positive	0.68	0.65	0.66
	negative	0.73	0.75	0.74
	neutral	0.75	0.78	0.76

5 Conclusion

We described a sentiment analysis system for short texts in Portuguese, and how it was improved to deal with entity aspects. Considering weight of each class in the target oriented training set, we can estimate a weighted average of the precision at 72.6%. This is roughly the same precision as the previous homologous classifier, but with the important difference of granularity in detail, which is now richer, having indication of polarity per each aspect of the target entity.

The evaluation presented for target oriented in section 4 refers to the appreciation of the sentiment classifier, on a large number of instances without aspects ("general" aspect) and approximately 1000 instances with a tagged aspect and its polarity. We believe that the number of instances with aspects is still reduced.

For future developments, we have many lines of work. We intend to collect more instances with aspects for the training set. The aspect extraction module needs a specific evaluation, which in this work did not happen. Still regarding aspect extraction, we can restrict more, limiting aspects to a well-known set for each type of entity, or be more permissive, allowing the discovery of unknown/unseen aspects, but adding the risk of increasing the error with false positives.

In future we also plan to experiment with other classification techniques such as deep neural networks. We did not do this already because the first experiences did not give good indicators. As we can see in the ranking of SemEval-Task 5's participating systems, the use of Deep Learning does not automatically mean better results. See the case of the winning system for the microblog messages subtask, which was ahead of other systems that used Deep Learning methods. In our case, with few training instances, we have chosen to stabilize a system based on conventional Machine Learning. Later, we can appreciate the possible benefit of Deep Learning for this dataset, now having a reference for comparison.

ACKNOWLEDGMENTS

This research is partially supported by the SmartSeg project, which is co-funded through Portugal 2020's "R&D Incentive System – Individual Projects" program, grant number "POCI-01-0247-FEDER-011192".

REFERENCES

- [1] Catherine E. Tucker (2014) Social Networks, Personalized Advertising, and Privacy Controls. *Journal of Marketing Research*: October 2014, Vol. 51, No. 5, pp. 546-562.
- [2] Pang, Bo & Lee, Lillian (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Vol. 2(1-2), pp. 1-135.

- [3] J. Saias, R. Silva, E. Oliveira, and R. Ruiz (2015). "Combining overall and target oriented sentiment analysis over portuguese text from social media," Transactions on Machine Learning and Artificial Intelligence, vol. 3, pp. 46–55, June 2015.
- [4] José Saias (2015). Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, USA. June 2015. p. 767-771, ACL
- [5] E. Dovidon and J. Saias (2017). "ej-sa-2017 at semeval-2017 task 4: Experiments for target oriented sentiment analysis in twitter," in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), (Vancouver, Canada), pp. 635–638, Association for Computational Linguistics
- [6] Silva, N. F. F. (2016). Análise de sentimentos em textos curtos provenientes de redes sociais. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil.
- [7] Silva, L. (2018) Análise de Sentimentos em língua portuguesa do Brasil. Last accessed January 2018, at: <https://www.kaggle.com/leandrodoze/sentiment-analysis-in-portuguese>
- [8] Balage Filho, P. P. (2017). Aspect extraction in sentiment analysis for portuguese language. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brasil.
- [9] Repustate. Sentiment analysis, social media sentiment and text analytics. Last accessed February 2018, at: <https://www.repustate.com/portuguese-sentiment-analysis/>
- [10] S. Rosenthal, N. Farra and P. Nakov (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Pages: 502–518. ACL, 2017.
- [11] K. Cortis et al. (2017). SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Pages: 519–535. ACL, 2017.
- [12] Mengxiao Jiang, et al. (2017). Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval2017). Association for Computational Linguistics, Vancouver, Canada.
- [13] McCallum, Andrew Kachites (2002). "MALLET: A Machine Learning for Language Toolkit". <http://mallet.cs.umass.edu>
- [14] 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.
- [15] Bing Liu (2010). Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, 2nd Edition. Taylor and Francis Group, Boca, 2010.

- [16] José Saias (2017). Ludificação: experiências para construção e marcação de um corpus para Análise de Sentimentos. In I Congresso Luso-Extremadurense de Ciências e Tecnologia, Universidade de Évora. Outubro de 2017. ISBN: 978-989-8550-45-3

- [17] Kim, B. (2015). Understanding Gamification. Library Technology Reports, 51(2), 1–35