BooViews: Aspect-based Sentiment Analysis on Product Reviews combining SVM and CRF in Portuguese

Guilherme Nobre[1,2], Alan Justino[2], Fernando Tadao[2], Danilo Nunes[2], Daniel Takabayashi[2] and Rayssa Küllian[2]

[1] FATEC São Caetano do Sul, São Paulo, Brazil
[2] Boolabs, Artificial Intelligence Research Department, São Paulo, Brazil

Abstract. Customers use product reviews to gather opinion regarding a product before making a purchase decision. Reviews are available in several e-commerce businesses and wrote by real customers. Companies can use these reviews to harness consumer information by artificial intelligence algorithms and automatically extract product information, such as review polarity and product aspects. In this paper, we cover techniques to classify reviews polarity, extract product aspects and classify them. The resulting SVM classifier got 91.7% of precision in classifying the sentiment of the reviews, 74.2% of F1-score using CRF to extract the product aspects and 79.9% of precision classifying aspect's polarity.

Keywords:
aspect-based sentiment analysis, ABSA, sentiment analysis, text mining, SVM, CRF, reviews, books, smartphones, Amazon, Buscapé, pt-br

1 INTRODUCTION

Reviews, as opinionated texts about a product, influence consumers in their shopping decisions about products and services [7]. With e-commerce and other online sources, sites opened space for customers to describe their experiences with products in the form of reviews, leading to a massive growth in its production. Growth of such magnitude that the availability is too high for manual analysis.

This work discusses Artificial Intelligence approaches to discover relevant product data of reviews extracted from the Internet and presents the BooViews system, aimed to extract the aspects of a given product review and classify its sentiment polarity. Such task is described by the 8th International Workshop on Semantic Evaluation series (SemEval 2014) in the form of task 4 subtasks 1 and 2 [7].

The proposed system is divided into two parts: a review polarity classifier, trained to detect a review text as positive or negative, and an aspect extractor that finds relevant characteristics from the reviewed subject from the review text. The latter uses the former to classify the aspects extracted as positive or negative, the final goal of the system.

2 METHODOLOGY

BooViews was based on Support Vector Machine and Conditional Random Fields.

Support Vector Machine (SVM). A classification method that discovers hyper-planes in the data space, providing functions that separate the data in a linear or non-linear fashion, depending on the kernel [4]. These hyper-planes are defined by points in the frontier between classes (Support Vectors) that serve as pivots for a plane dividing the data.

Conditional Random Fields (CRF). Being a supervised learning sequential labeling algorithm capable of taking the context of "neighboring" samples into account when classifying, e.g. between ASPECT/NOT-ASPECT, is described by Balage & Thiago [1] as a way to extract features over an universe of reviews in English about laptop computers, universe provided by the 8th International Workshop on Semantic Evaluation (SemEval 2014).

2.1 Data Origin

Two original corpora were compiled for the present work: one from the "*livros*" (books) section from Amazon.com.br e-commerce brazilian website and one from the "*celulares*" (cellphones) section from the Buscape.com.br price comparison brazilian website. Such information could be considered public and free enough to be source of research data without ethical or legal concerns [3].

The new "Amazon.com.br Books" corpus contains 79.335 unique reviews from books and the new "Buscape.com.br Cellphones" corpus contains 86.926 unique reviews from cellphones. Each review collected contains its author rate as the number of stars given, title and body, among other information. Amazon sourced samples are classified over a 1 to 5 stars range, but Buscapé are classified by 0 to 5 stars.

Rain (2013) [8] argues that it is better to consider only reviews rated from 0 to 2 as negatives and only rated 5 as positives. Rates 3 and 4 could be positive, negative or neutral within the same universe (ambiguous) and should be discarded for polarity classification. This advice was followed.

Representative subsets size. Representative subsets had to be produced, having a reduced sample number considered feasible for manual annotation by the research team. The subsets contain 596 samples each, randomly selected from their original corpus. Half from the samples considered positive and half from the samples considered negative. The 596 figure was prescribed via the sample size formula [5], rounded up: $n = (z^2 * p * (1 - p))/c^2$

Let be z the confidence level, be it 96%, c the confidence interval as 4, and p the choice percentile. As the samples polarity is unknown on the original universe, assume p as 0.5 (neutral). Having the universe size available, the formula

could be reduced by this form suitable for finite population, as described by Muniz & Abreu (1999) [5]:

$$sample = \frac{n}{1 + \frac{n-1}{population}}$$

2.2 Production of representative subsets.

For training supervised models and for parameter search purposes, each corpus had produced two representative subsets with 596 samples each: one subset automatically annotated using the given "stars"; one subset manually annotated by the research team. This and other produced subsets are summarized on Table 1.

As it is shown in the next sections, the automatic annotation demonstrates some level of internal ambiguity between the multiple review authors. To circumvent such ambiguity, manually annotated subsets were produced to allow better model training and parameter search.

Polarity annotation from review body. Two manually annotated subsets with 596 balanced samples each were randomly selected for sentiment polarity (positive or negative), one from "Buscape.com.br Cellphones" and one from "Amazon.com.br Books". The manually annotated subsets were considered more accurate than the automatic annotation for model training and validating.

Aspects annotation from review body. The CRF needs an annotated corpus to be applied on. One was produced from the 596 samples subset of the "Buscape.com.br Cellphones", as seen on Table 1. The corpus was manually annotated with the perceived aspects of every review in the subset.

2.3 Model selection

The final classifier was desirable to be equally good classifying positive or negative samples. Was decided to let the subsets be composed of even parts of both classes.

Table 1. Produced representative subsets

Original corpus	Annotation purpose	Annotation origin	Samples
Amazon.com.br Books	Largest balanced universe	Automatic	10910
Buscape.com.br Cellphones	Largest balanced universe	Automatic	9506
Amazon.com.br Books	Sentiment polarity	Automatic	596
Amazon.com.br Books	Sentiment polarity	Manual	596
Buscape.com.br Cellphones	Sentiment polarity	Automatic	596
Buscape.com.br Cellphones	Sentiment polarity	Manual	596
Buscape.com.br Cellphones	Aspect extraction	Manual	596

Table 2. Distribution of classes of the samples on the "Amazon.com.br Books" corpus

Corpus (automatic rating)	Positve	Negative	Neutral	Total	Pos/Neg Ratio
Books Full Corpus	50894	5955	22486	79335	8.55
Books Largest Balanced Universe	5955	5955	0	11910	1.00

Table 3. Sentiment Polarity of whole review results

Train (Annotation)	Test	F1
596 "Amazon Books" balanced samples (manual)	cross-validation	93.60
596 "Amazon Books" balanced samples (automatic)	11910 samples (largest balanced subset)	86.82
	50894 samples (unbalanced universe)	91.72
596 "Buscape Celphones" balanced samples (manual)	cross-validation	96.30
596 "Buscape Celphones" balanced samples (automatic)	9506 samples (largest balanced subset)	84.54
	48330 samples (unbalanced universe)	88.40

The largest possible balanced subset of the universe with automatic rating annotation were produced for model testing purposes. On the "Amazon.com.br Books" corpus, for example, the negatives are relatively scarce. All the 5955 negative samples were combined with 5955 randomly chosen positive samples, as seen on Table 2.

2.4 Sentiment polarity detection over detected aspects

After extraction of aspects via CRF, the aspects sentiment polarity were decided using the SVM sentiment detector described before. The whole sentence containing the detected aspect was evaluated for sentiment polarity and the result was assumed as the aspect polarity. The sentence was detected using the *punkt* module of the NLTK library, configured for Portuguese sentences.

It is aware that two aspects with opposite polarities can feature the same sentence, but the handle of this case was considered matter for next works.

3 RESULTS

Sentiment polarity detection on the whole review. On cross-validation of the manually annotated balanced subsets of 596 samples, the SVM sentiment polarity detector model achieved 93.60% of F1 metric for books and 96.39% for cellphones. The experiments were carried out with a 3-fold cross-validation scheme, with 80% of data for training and 20% for testing.

When trained with the 596 automatic annotated balanced samples from each corpus and tested over the entire original unbalanced corpus, the model achieved 91.72% for F1 on the "Amazon.com.br Books" corpus and 88.40% on the "Buscape.com.br cellphones" corpus.

On Table 3 is shown the results achieved by the BooViews classifier, linear-kernel based using TF/IDF as feature extractor. Implementation came from

Scikit-learn [6]. Fang 2015 [2] showed a SVM classifier going from 0.61% to 0.94% as its training data increased from 180 to 1.8 million product reviews.

Aspect extraction. The resulting CRF model achieved 74.28% for F1 when applied on the 596 manually annotated samples reviewing cellphones, after boosting. For comparison, the best team on Semeval 2014 Task 4 Subtask 1 [7] training using an in-domain dataset (laptops) achieved 73.78% for F1.

Sentiment polarity detection on each extracted aspect. Applying the model trained for cellphone reviews over the aspects extracted via CRF from the 596 manually annotated samples reviewing cellphones, it achieved 79.95% for F1 on the aspects sentiment polarity. The best system on the Semeval 2014 Task 4 Subtask 2 [7] achieved 70.48% for F1 over the laptops domain.

4 ANALYSIS AND CONCLUSIONS

In this paper, we elaborate on the uses of e-commerce data to extract information from subjective and opinionated texts. This work can be adapted in websites for better search results, for filters directed to the client that wishes for specific traits on a product and for manufacturers that want to cater and listen to consumer feedback. Classifier results are on par with past works in SemEval, portraying the benchmark for this area.

Representativeness of the subsets over the original universes. The model which achieved 91.72% of F1 over the whole "Amazon.com.br Books" corpus was trained with only 596 samples, less than 1% of this researched universe. It could infer that the books review universe have a relatively low number of meaningful words for sentiment polarity detection, but this hypothesis is to be verified in future works.

Generalization between the researched universes. The models trained from cellphones or books corpora samples were cross tested using the other domain, in other words, trained from books and tested on cellphones and *vice-versa*. As seem on Table 4, each model still have good results on their "unfamiliar" corpus, indicating that this two domains could be partially generalized in future works. Specially, the cellphones model achieved 77.02% for F1 on the books universe balanced samples.

Table 4. Cross tests over the books and cellphone domain models

Metric: F1	Book test set (11190 balanced samples)	Cellphone test set (9506 balanced samples)
Books train set (596 balanced samples)	**86.82%**	60.26%
Cellphones train set (596 balanced samples)	77.02%	**85.54%**

References

1. Balage Filho, P. P., & Pardo, T. A. (2014). *NILC USP: Aspect Extraction using Semantic Labels*. SemEval 2014, 433.
2. Fang, X., & Zhan, J. (2015). *Sentiment analysis using product review data*. Journal of Big Data, 2(1), 1-14.
3. Giles, C. L., Sun, Y., & Councill, I. G. (2010, April). *Measuring the web crawler ethics*. In Proceedings of the 19th international conference on World wide web (pp. 1101-1102). ACM.
4. Meyer, D., & Wien, F. T. (2015). *Support vector machines. The Interface to libsvm in package e1071*.
5. Muniz, J. A., & Abreu, A. R. (1999). *Técnicas de amostragem*. Lavras: UFLA/FAEP (pp. 102)
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). *Scikit-learn: Machine learning in Python*. The Journal of Machine Learning Research, 12, 2825-2830.
7. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014, August). *Semeval-2014 task 4: Aspect based sentiment analysis*. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) (pp. 27-35).
8. Rain, C. (2013). *Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning*. Swarthmore College.