Explain your opinion.
Aspect-Based Sentiment Analysis with Born Classifier

Gaudenzia Genoni
University of Milan, Italy
gaudenzia.genoni@gmail.com

Abstract

This paper presents an approach to Aspect-Based Sentiment Analysis which leverages the explanatory capabilities of Born Classifier, a self-explainable text classification algorithm inspired by the principles of quantum physics. The research first employs Born Classifier to perform an initial sentiment analysis on a corpus of vectorized documents, obtained from a well-balanced dataset of app reviews (AWARE). The most influential candidate aspects are then extracted from the list of explanation features generated during the classification process, and, finally, all aspects are linked to specific textual segments to predict sentiment polarity more accurately on shorter portions of the texts. Although the evaluation metrics show improvement, further work could be done in the future to refine the methods for aspect detection.

1 Introduction

Sentiment Analysis, sometimes referred to as Opinion Mining, is a field at the intersection of information retrieval, natural language processing and artificial intelligence that deals with the identification and extraction of user’s opinions and emotions [5], [7]. Among the three main models of classification described in [4], aspect-level sentiment analysis is the most fine-grained, as it aims to extract opinions expressed against different aspects/features of the entity [3], allowing for a more detailed analysis of the information provided by the textual reviews. The main challenges of Aspect-Based Sentiment Analysis are the identification of sentiment-target pairs in the text and the classification of the expressed sentiment according to a predefined set of sentiment values, for instance positive and negative [8].

In this paper, Aspect-Based Sentiment Analysis is conducted by exploiting the explanation features given by Born Classifier, which is a self-explainable supervised classification algorithm based on key postulates of quantum mechanics (the Born rule) [2]. In particular, a sentiment analysis classification of a corpus of vectorized documents is performed a first time using Born Classifier; all the explanation features are then extracted from the global Born explanation and the most influential ones are selected as candidate aspects, provided that they appear in the documents as nouns at least once; finally, each aspect is associated to a specific sentence of the texts and the sentiment is predicted a second time using the trained Born classifier, taking insights into the polarity of the specific aspects.

The paper is structured as follows. Section 2 states the goals of the project and provides an overview of the proposed approach, gradually explaining the methodology adopted in the procedure. Section 3 briefly discusses the dataset used for the experiments and presents the results as plots and tables, focusing on the metrics used for the evaluation. The final section is dedicated to concluding remarks, which include a critical discussion on the experimental results and ideas for future work.
All the code is available at `https://github. com/Ggenoni/ABSA_with_Born_Classifier`.

2 Research question and methodology

The research behind this project is meant to understand how much the accuracy of predicted sentiments for a collection of reviews can be influenced by the polarity associated with the main aspects of the products identified by the users; this intrinsically also implies defining a method to detect the aspects, which is one of the most challenging tasks of Aspect-Based Sentiment Analysis. The approach taken to answer this question involved performing a sentiment analysis on a corpus of documents, locating aspects among the explanation features proposed by the classifier, and executing a new sentiment analysis only on the portions of the texts where aspects are explicitly mentioned.

The first step required using Born Classifier to carry out a sentiment analysis on a collection of documents. Since the set of human emotions is very large [6], a binary classification was preferred, distinguishing only between positive and negative polarities. Each document was divided into linguistically meaningful units using spaCy, an open-source library for natural language processing in Python. The part-of-speech (POS) tagging was employed to keep only tokens regarded as 'ADJ', 'ADV', 'NOUN', 'VERB' and 'PNOUN'; furthermore, tokens classified as 'PART' (*particles*) whose lemma is “not” were also taken, in order to include negations as well. During the normalization step, all texts were further processed by transforming each token into its lemma (all lowercase), to limit the number of units passed to the classifier. The documents were then vectorized according to the Bag of Words model, calculating the TF-IDF (*Term Frequency – Inverse Document Frequency)* score for each token and creating a sparse matrix with scikit-learn’s TfidfVectorizer(). After classifying each document, Born Classifier allowed to identify the explanation features that globally influenced its decisions the most.

Features extracted from the global explanation of Born Classifier having part of speech equal to 'NOUN' in at least one document of the original dataset were taken into consideration as candidate aspects, with the assumption that nouns can represent different elements of an object/place/concept better than other lexical categories. Candidate aspects were later divided into positive and negative according to the weight assigned to them by Born Classifier, and for each of these two groups only the most influential aspects were selected, choosing those with weight equal to or higher than the third quartile value of their group. Subsequently, each identified candidate aspect was associated with a portion of the text, more specifically with one or more sentences in which the aspect appears having part of speech equal to 'NOUN': this procedure was repeated twice for both negative and positive aspects. The aim was to map out where each aspect is being discussed in the text, preparing for a more granular sentiment analysis.

The last step consisted in predicting the sentiment for each sentence collected in the previous phase, this time with a focus specifically on the identified aspects. The predictions were made using the already trained Born Classifier, with the intention of verifying if isolating portions of the texts in which the aspects are mentioned can help improving the outcomes of the classification. The analysis was performed first on the subset of sentences containing negative aspects, and then on the subset of sentences containing positive aspects. The metrics used for the evaluation, together with a short overview of the experimental results of the research, are presented in the following section.

3 Experimental results

The dataset used for this project is AWARE, *Aspect-Based Sentiment Analysis Dataset of Apps Reviews for Requirements Elicitation* [1]. It contains 11321 apps reviews from three different domains (Productivity, Social Networking, and Games), where each review is annotated with labels relative to aspect terms (one word or a couple of words mentioned in the reviews that describe an aspect expressed by the sentiment), domain-specific categories for each aspect, and the corresponding sentiment (either positive or negative). Since each review is annotated with a Boolean value indicating whether it expresses an opinion or not, all reviews that do not contain an opinion were discarded. After eliminating all the rows with nulls, the total number of reviews left amounted to 6824, 3337 of which positive and 3487 negative: the dataset can therefore be regarded as well-balanced. It must be considered that the sentiment analysis discussed below was performed only on the first 3000 reviews of the dataset.

All the results were obtained using Python 3.12.0.

The results of the first sentiment analysis preformed with Born Classifier are summarized below in Table 1:

	precision	recall	f1-score	support
negative	0.62	0.68	0.65	437
positive	0.67	0.61	0.64	463
accuracy			0.64	900
macro avg	0.64	0.64	0.64	900
weighted avg	0.65	0.64	0.64	900

Table 1: Precision is the ratio between the number of relevant retrieved instances and all retrieved instances, while Recall measures the number of relevant retrieved instances over the number of all relevant instances. F1-score is the Harmonic mean of Precision and Recall ($2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$).

In this case, all the evaluation metrics remain below 70% and the overall accuracy stands at 64%. These poor results may suggest that there is too much noise in the feature set.

The evaluation is graphically shown below in the Confusion Matrix of Figure 1.

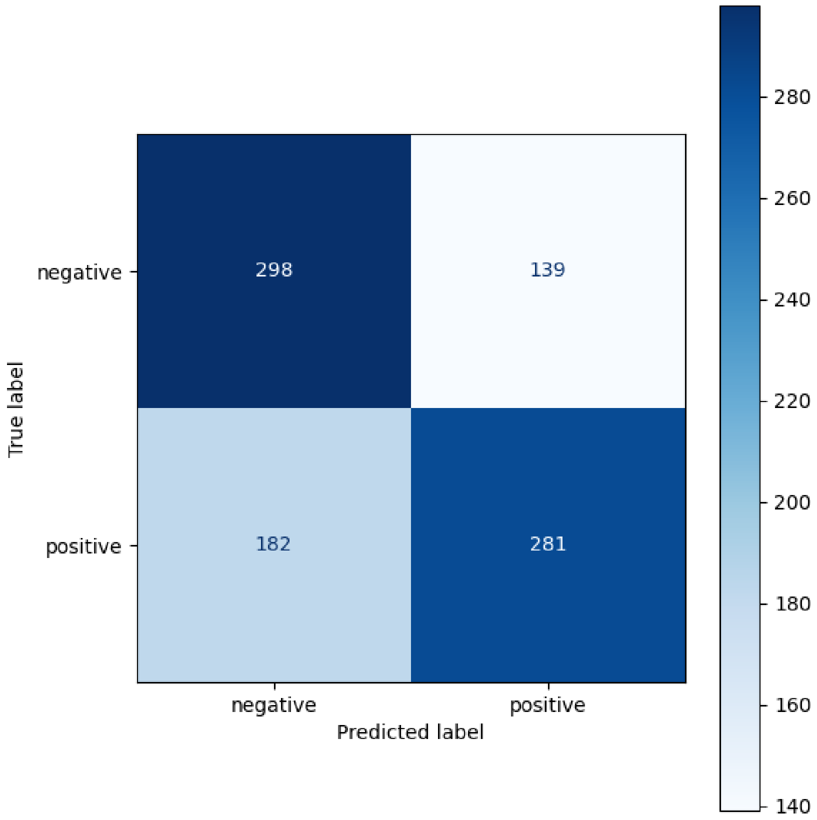

Figure 1: Confusion Matrix for the first sentiment classification.

The results of the sentiment analysis performed on the subset of sentences containing negative aspects are summarized in Table 2 and graphically shown in the Confusion Matrix of Figure 2:

	precision	recall	f1-score	support
negative	0.80	0.85	0.83	1995
positive	0.45	0.36	0.40	654
accuracy			0.73	2649
macro avg	0.63	0.61	0.61	2649
weighted avg	0.72	0.73	0.72	2649

Table 2: In this case, the evaluation metrics for negative predictions are rather good (80% of Precision and 85% of Recall), but the positive predictions seem to be heavily penalized. The overall accuracy, however, stands above 70%.

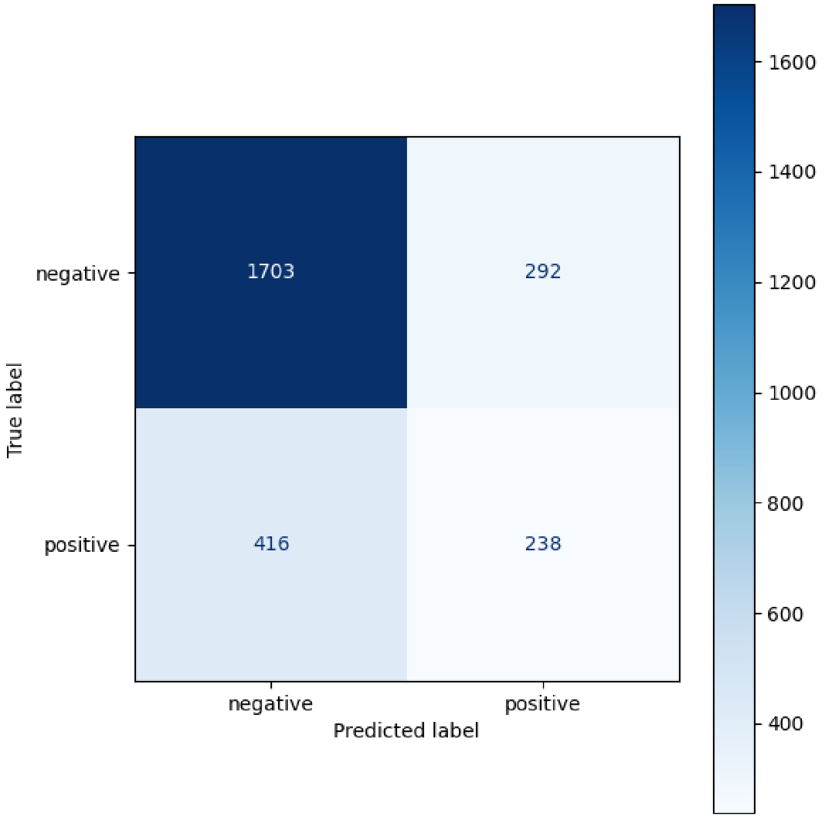

Figure 2: Confusion Matrix for the sentiment classification of sentences containing negative aspects.

The results of the sentiment analysis performed on the subset of sentences containing positive aspects are summarized in Table 3 and graphically shown in the confusion matrix of Figure 3:

	precision	recall	f1-score	support
negative	0.46	0.33	0.38	636
positive	0.84	0.90	0.87	2441
accuracy			0.78	3077
macro avg	0.65	0.61	0.62	3077
weighted avg	0.76	0.78	0.77	3077

Table 3: Contrary to the previous case, here the evaluation metrics for positive predictions are satisfying (84% of Precision and 90% of Recall), while the negative predictions are very poor. Again, the overall accuracy stands well above 70%, almost reaching 80%.

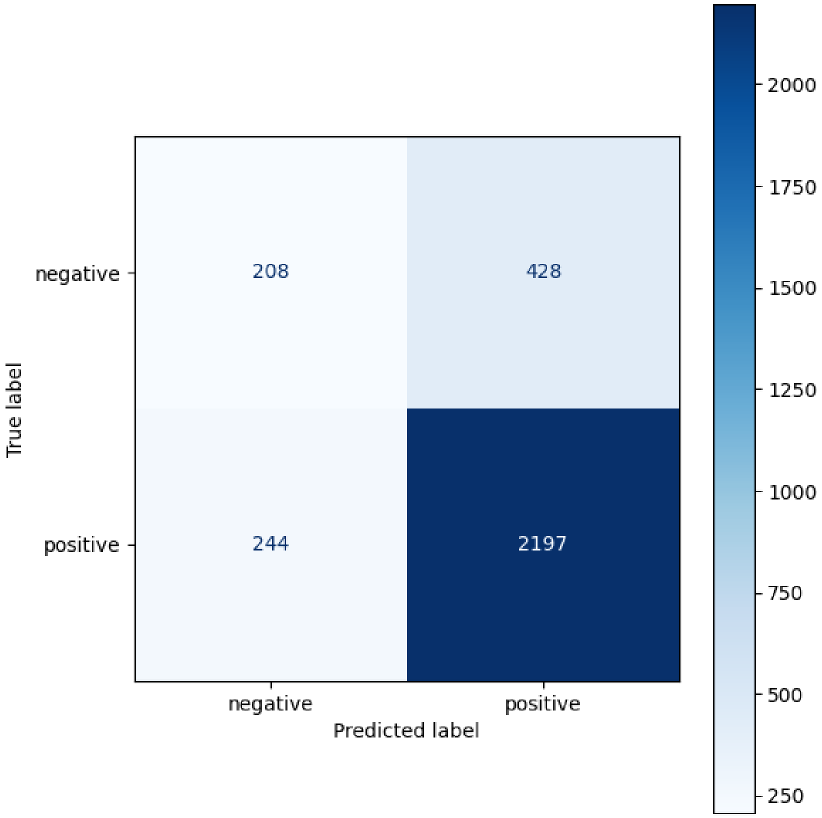

Figure 3: Confusion Matrix for the sentiment classification of sentences containing positive aspects.

In general, it can be said that the results of the second and third classifications, run on a subset of sentences containing – respectively – negative and positive aspects, are superior to the first sentiment analysis performed with Born Classifier on the test data. Limiting the analysis to the portions of the texts where the most influential terms are mentioned can help reduce noise and improve the accuracy of the classification. On the other hand, however, it must be considered that the presence of strongly polarized terms can force the predictions towards the sentiment that is prevalent in the data.

Concluding remarks

As made evident by the experimental results, the classification improves when performed on a smaller set of data composed only of sentences containing relevant aspects, showing a correlation between the sentiment orientations of the aspects and the general opinions expressed by reviewers.

The main issue that must be addressed here is the method chosen to detect the aspects. Since the explanation features are provided by Born Classifier as a list in alphabetical order and all information about syntactical dependency and word order is lost, the method adopted in this project was to pick aspects according to their POS tagging, considering only nouns and using a statistical criterion to select the most influential terms. This is however a naïve approach that could be refined in future work. Furthermore, spaCy POS tagging is prone to errors: for instance, in the sentence “I use macOS and find the syncing between the Mac and iPhone flawless”, from the review indexed 205 (its id in the dataset is fa397047-d276-43f9-895b-a5c25bf9ba21), the term “flawless” is annotated as 'NOUN', while it is clear to every human reader that it is an adjective. Unfortunately, it is impossible to assess the number of such errors.

In addition to finding an alternative solution for aspect detection, an interesting idea for future projects could be to evaluate the accuracy of the predictions when the classification is performed on all the remaining sentences in which aspects are not mentioned, in order to acquire more information on how strongly the aspects can influence the sentiment of the reviews.

References

[1] Nouf Alturaief, Hamoud Aljamaan, and Malak Baslyman. AWARE: Aspect-Based Sentiment Analysis Dataset of Apps Reviews for Requirements Elicitation. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 211-218, 2021, doi: 10.1109/ASEW52652.2021.00049.

[2] Emanuele Guidotti and Alfio Ferrara. Text Classification with Born’s rule. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[3] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 168-177, 2004.

[4] Bing Liu. Sentiment Analysis and Opinion Mining. Series *Synthesis Lectures on Human Language Technologies*, volume 16. San Mateo, CA, USA: Morgan, 2012.

[5] Toqir Ahmad Rana and Yu-N Cheah. Aspect extraction in sentiment analysis: comparative analysis and survey. In *Artificial Intelligence Review*, volume 46, issue 4, pages 459-483, 2016.

[6] Robert Plutchik. Emotion: A Psychoevolutionary Synthesis. New York, NY, USA: Harper & Row, 1980.

[7] Kim Schouten and Flavius Frasincar. Survey on Aspect-Level Sentiment Analysis. In *IEEE Transactions on Knowledge and Data Engineering*, volume 28, issue 3, pages 813-830, 2016, doi: 10.1109/TKDE.2015.2485209.

[8] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. In *Data Mining Knowledge Discovery,* volume 24, issue 3, pages 478-514, 2012.