Text Mining and Sentiment Analysis

Explain Your Opinion

Sofia Introzzi
sofia.introzzi@studenti.unimi.it

October 2023

1 Introduction

Sentiment Analysis is a core subject in the field of Text mining and Sentiment Analysis and it is comprehended in the field of Text Classification. Precisely, it consists in binary classification in which the output classes reveal the polarity of the portion of text that is considered. Although it belongs to this field, this feature distinguishes it from the others tasks. The output gives information concerning the polarity of the textual document.

Concerning this field, a significant subarea is represented by the Aspect Based Sentiment Analysis (ABSA). The peculiarity of this field is the finer level of granularity to which the sentiment is applied, the aspect indeed. It consists of two main related processing: the aspect detection and sentiment analysis on the aspect. This field is also referred as fine-grained target opinion. It finds indeed many real-word applications in costumer analysis.

In order to conduct this study the classification algorithm that has been exploited is the Born Classifier. The textual dataset instead comes from the juridical field, it is indeed a collection of legal labelled documents.

The main purpose of this research project is understanding how Sentiment Analysis differs at different levels of granularity and targets changes and if consistency in the different levels of analysis and results is observable.

In the first section, the Sentiment Analysis for documents classification is presented, afterwards the Aspect Based Sentiment Analysis pipeline is explored, and finally results are presented.

2 Data

The data implemented for this analysis has been retrieved from the "Open Science Framework (OSF)" public platform. The dataset, named SigmaLaw-ABSA, comprises textual data related to judicial opinions, concerning therefore the legal opinion domain, in which ABSA processing is still pioneer. Each document in this dataset is associated to a human-annotated label related to the polarity score, identified either in a positive (1) or negative (-1) value.

	Sentence	Sentiment
0	[2008, feder, offici, receiv, tip, confidenti,...	-1
1	[2008, feder, offici, receiv, tip, confidenti,...	1
2	[lee, sold, inform, ecstasi, marijuana]	-1
3	[obtain, warrant, offici, search, lee, 's, hou...	-1
4	[found, drug, cash, load, rifl]	1
...	...	...
1615	[carri, elebi, linda, jacob, stestifi, heard, ...	-1
1616	[final, govern, play, videotap, petition, yarb...	-1
1617	[none, defend, rebut, prosecut, wit, ', claim,...	-1
1618	[govern, not, contest, petition, ', claim, wit...	-1
1619	[govern, present, sever, wit, corrobor, aspect...	1

1620 rows × 2 columns

Figure 1: Textual dataset

The dataset comprises a total of 1,620 observations, with a slight imbalance as it contains more observations associated with a negative score.

3 Methodology

In this section the main techniques for information retrieval and sentiment analysis will be presented.

3.1 Sentiment Analysis for documents

The initial step in the analysis of textual datasets involves text preprocessing, where fundamental techniques in Natural Language Processing are applied. As previously mentioned, the classification algorithm utilized for Sentiment

Analysis is the Born Classifier: a supervised classification algorithm based on the Born rule of quantum mechanics.

In order to score the polarity of the documents in the textual corpora, the first requirement was to vectorize data. The Term Frequency-Inverse Document Frequency (TF-IDF) consists in measuring the frequency of a term against the overall terms frequency. (in order to obtain the TF-IDF matrix). After partitioning the data into a 75% training set and a 25% test set we are able to fit the Born Classifier. The resulting metrics present quite decent results, moreover it is essential to remember the slightly imbalanced in the number of instances per class. The model indeed tends to perform better on the class with a larger number of observations.

3.2 Aspect Based Sentiment Analysis

The second part of this research project has been addressed in two phases. The first section focuses on the approach for retrieving the relevant features whilst the second one in performing the sentiment analysis at different level. Concerning the first stage, identifying explanatory features is a relevant step, indeed recalling the fact that our first objective at this stage is detecting aspects in the sentences, finding relevant features, such as noun-adjective pairs, would support this task: the explanatory nouns are indeed likely to be aspects. If the noun is informative with respect to the sentence, then it could be an aspect.

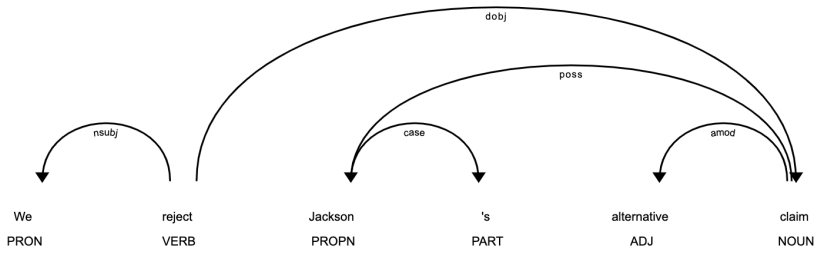

Figure 2: Sample of syntax dependency

An essential step to address this analysis is given by the syntax. The syntactic dependency of tokens are used indeed to detect noun phrases. First of all, Spacy library has been used to sparse documents through the English module. For each span, a token map has been created in order climb the tree structure, i.e., to explore the sentences to collect the syntactic information and finally adjectives and nouns for each document have been returned. The concern is to evaluate whether those pairs are informative, therefore they can be used them as features. In order to evaluate this, the Pointwise Mutual Information method has been applied. The PMI consists in comparing the joint distribution of the pair over the probability of observing the variables independently. The PMI will return a value that is explicative of how much information the pair is providing.
Afterwards, the identified nouns have been used for clustering words, that have been previously retrieved. In other words, a matrix between the aspects and the words has been generated in order to retrieve the pairwise similarity.

3.3 Sentiment classification

As a final step, the Pointwise Sentiment Information has been used to apply the Born classification algorithm for predicting sentiment scores while considering the aspects. It is interesting to observe that the average precision is 0.68% that rises to 0.74% when weighted for the classes. Concerning the sentiment classification of documents the average accuracy was 0.71%, and 0.72% for the weighted score.

	0	1
(absolute, immunity)	0.021987	0.016453
(adequate, warning)	0.036871	0.059605
(first, degree)	0.377964	0.000000
(full, value)	0.004475	0.005115
(grand, jury)	0.000021	0.000021
(guilty, plea)	0.002064	0.001892
(ineffective, assistance)	0.011180	0.009037
(prior, conviction)	0.000000	0.000000
(reasonable, person)	0.000021	0.000021
(summary, judgment)	0.038576	0.059161

Figure 3: Explanation results

It is interesting to observe the explanatory table. It is evident, the pair

"first-degree" exhibits the highest explanatory power in predicting the negative class.

4 Conclusions

In conclusion, it can be observed that the results in sentiment scores remains relatively consistent across different levels of sentiment classification. Rather it can be confirmed the consistency in the results.

However, it's worth noting that the Born Classifier demonstrates slightly improved performance for document classification. Nevertheless, in ABSA results appears to be more affected by the imbalanced dataset. Finally, the Born Classifier has proven to be an efficient performer.

The field of Aspect Based Sentiment Analysis lets plenty of room for further analysis due to its double nature of information mining and sentiment analysis making it a fascinating field. Interesting further investigations could concern comparison analysis of

References

E. Guidotti and A. Ferrara. Text classification with born's rule. In *Advances in Neural Information Processing Systems*, volume 35, pages 30990–31001, 2022.

C. R. Mudalige, D. d. S. Karunarathna, A. S. Perera, and R. Pathirana. SigmaLaw-ABSA: Dataset for Aspect-Based Sentiment Analysis in Legal Opinion Texts. , 2020.

T. A. Rana and Y. N. Cheah. Aspect extraction in sentiment analysis: comparative analysis and survey. , 2016.

K. Schouten and F. Frasincar. Survey on aspect-level sentiment analysis. . pages 813–830, 2015.