SentNA @ ATE_ABSITA: Sentiment Analysis of Customer Reviews Using Boosted Trees with Lexical and Lexicon-based Features

Francesco Mele
Institute of Applied Sciences and Intelligent Systems
National Research Council
f.mele@isasi.cnr.it

Antonio Sorgente
Institute of Applied Sciences and Intelligent Systems
National Research Council
a.sorgente@isasi.cnr.it

Giuseppe Vettigli
Centrica plc,
Institute of Applied Sciences and Intelligent Systems (CNR)
giuseppe.vettigli@centrica.com

Abstract

English. This paper describes our submission to the tasks on Sentiment Analysis of ATE_ABSITA (Aspect Term Extraction and Aspect-Based Sentiment Analysis). In particular, we focused on Task 3 using an approach based on combining frequency of words with lexicon-based polarities and uses Boosted Trees to predict the sentiment score. This approach achieved a competitive error and, thanks to the interpretability of the building blocks, allows us to show the what elements are considered when making the prediction. We also joined Task 1 proposing a hybrid model that joins rule-based and machine learning methodologies in order to combine the advantages of both. The model proposed for Task 1 is only preliminary.

Italiano. *Questo articolo descrive la nostra sottomissione ai tasks sulla Sentiment Analysis ATE_ABSITA (Aspect Term Extraction and Aspect-Based Sentiment Analysis). I nostri sforzi si sono concentrati sul Task 3 per il quale abbiamo adottato gli alberi di predizione (Boosted Trees) utilizzando come features di ingresso una combinazione basata sulla frequenza delle parole con la polarità derivate da un lessico. L'approccio raggiunge un errore competitivo e, grazie all'interpretabilità dei moduli intermedi, ci consente di analizzare in dettaglio gli elementi che caratterizzano maggiormente la fase di predizione. Una proposta è stata realizzata anche per il Task 1, dove abbiamo sviluppato un modello ibrido che combina un approcio basato su regole con tecniche Machine Learning. Il modello sviluppato per il Task 1 è solo in fase preliminare.*

1 Introduction

User feedback has become essential for companies to improve their services and products. Nowadays, we can find user feedback in textual form as online reviews, posts on social media and so on. These resources can express overall opinions but also opinions about some specific details (aspects) of the subject. In this scenario, the tools provided by Sentiment Analysis are crucial to process user feedbacks, the ongoing research in this field is focused on creating models that are more and more accurate and that can also extract fine grained information for the data. As part of this research, the ATE_ABSITA tasks (de Mattei et al., 2020)[1], part of the EVALITA campaign (Basile et al., 2020), challenge the participants in extracting the aspects (Task 1), predict the sentiment towards each expect (Task 2) and also predict the overall sentiment expressed (Task 3) for a dataset containing reviews of items from an online shop.

It's important to notice that the dataset released for the task is one of the few resources for the Italian language that has annotated aspects and sentiment at the same time. Others Italian resources that take into account sentiment with respect to aspects are (Sorgente et al., 2014) and (Croce et al., 2013). The first contains reviews of movies with 8 domain specific aspects and 5 different polarity values while the second contains opinions about wines considering 5 aspects and 3 possible polarity values.

This paper describes our approaches in solving task 1 and task 3. The approach for task 1 is still preliminary.

[1] http://www.di.uniba.it/ swap/ate_absita/index.html

In the last decade top performing approaches to Sentiment Analysis have shifted from using classifiers on hand-crafted features, often based on lexicons (Zhu et al., 2014), to complex models based on deep Neural Networks and advanced word embeddings (Liu et al., 2020). While the latest models require special hardware and significant work to be trained, older approaches are built on top of well understood classification techniques that can be trained on commodity hardware which makes them easy to adapt for new applications. The approach proposed for Task 3 revisits the old fashioned style of doing Sentiment Analysis to see how it performs against more modern methodologies that are used in the competition.

Regarding Task 1 we follow the latest trend of exploiting linguistic patterns (Poria et al., 2016; Liu et al., 2015; Poria et al., 2014; Rana and Cheah, 2019). What distinguishes our approach from others is that we use automatically generated patterns based on POS-Tags (Part of Speech-Tags) following the assumption that they are more robust to bad grammar compared to linguistic dependencies.

In Section 2 we will describe our approach for Task 3 and in Section 2.4 we will discuss the results. In Section 3 we will briefly discuss the preliminary model we build for Task 1 and its results.

2 Our approach for Task 3

The idea behind our approach is to achieve competitive results using well known tools that can be used on commodity hardware. We build the features representing the text using n-grams and adding a set of characteristic annotated in SenticNet (Cambria et al., 2010). Given the large amount of features, we decided to use Boosted Trees as regression model given its ability to sub-sample the features dynamically. For textual preprocessing the libraries Spacy (Honnibal and Montani, 2017) and Scikit-Learn (Pedregosa et al., 2011) were used. We chose XGboost (Chen and Guestrin, 2016) as implementation of Boosted Trees for regression.

2.1 Lexical features

Before extracting the lexical features we remove stop words (apart from words that can be used as negative adverbs) and lemmatized each word. Finally, we extract a set of n-grams from each review. We consider uni-grams, bi-grams and tri-grams at the same time.

2.2 Lexicon-based features

To build the polarity features of our model, we have adopted SenticNet, a resource used for concept-level sentiment analysis. It contains a collection of concepts, including common-sense concepts, provided with values for polarity, attention, pleasantness and sensitivity. These are numerical features that are available for a subset of the words in each review. We take in account the average, the minimum and the maximum of all the values available in each review. We also consider the mood tags provided by SenticNet. They are sets of tags as #tristezza, #rabbia, #felicità[2] attached to each word, we consider them as binary features.

2.3 Regressor

Our final regressor is composed of 800 Decision Trees with a maximum depth of 4 layers. The model was trained using Gradient Boosting with a learning rate of 0.3. The final prediction is computed averaging the output of each tree. The rationale behind our choice is that we have a high number of features that are easy to use with tree based methods for specific cases, hence ensembling allows us to learn a set of shallow trees and each of them can work well for specific cases.

2.4 Results and discussion

To build our model we initially focused on the training set using cross-validation to optimize the parameters achieving a root mean square error of 0.852 (the prediction target is on a scale from 1 to 5), we then tested the optimized model on the development set reaching an error of 0.805. We finally achieved an error of 0.795 on the final test set. The difference in the error across the different stages of validation suggests that the model is well trained as the error doesn't increase when new data is presented. However, it also suggest that the estimation of the error has a wide confidence interval, the standard deviation estimated during cross validation is 0.049.

In Figure 1 we compare the scores predicted and the annotated score on the development set. The chart shows that the model has a tendency to over estimate the error, especially in cases annotated with a low score.

[2] In English: #sadness, #anger, #happiness

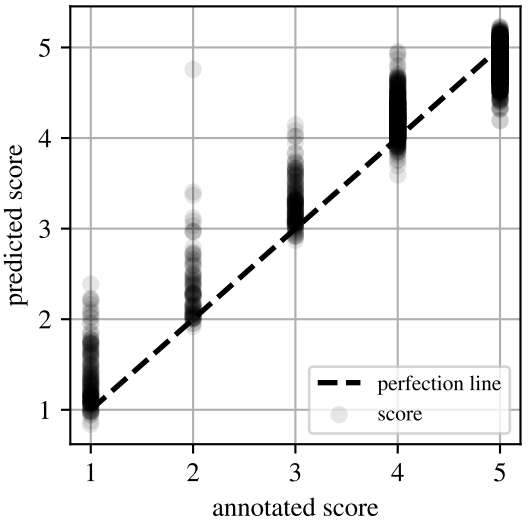

Figure 1: Scatter plot that shows the annotated score against the predicted score on the development set.

term	importance	coverage %
pessimo	0.057123	5.712323
purtroppo	0.038088	9.521134
rimborsare	0.037871	13.308205
non consigliare	0.033299	16.638059
purtroppo essere	0.027965	19.434580
cattivo	0.025690	22.003609
dispiacere	0.024986	24.502171
pensare	0.018631	26.365243
sconsigliare	0.016331	27.998360
dopo	0.016239	29.622279
non funzionare	0.015425	31.164802
delusione	0.015227	32.687547
non riconoscere	0.014809	34.168431
restituire	0.014615	35.629894
bruciare	0.014250	37.054852

Table 1: Important terms highlighted by the model. The column importance reports the importance score of the term while coverage is the cumulative sum of the importance scores.

We will now examine two reviews for which our regressor has the highest error. This is the text of the first review:

"si autospenge proprio quando si necessita di usarla contelecomando"[3].

This review was annotated with a score of 2, but the score assigned by our system is 4.75. This highlights a tendency of the system to give higher scores in uncertain cases. In this specific case we have no adjectives and two typing mistakes that result in no information from the lexicon and most of the words being disregarded as rare by our preprocessing pipeline. This suggests that a special treatment is needed for these specific cases where the classifier has fewer elements to take a decision.

The text of the second review is:

"Per questo prezzo c'è di meglio.. restituita.Gli accessori sono ottimi."[4].

This sentence was annotated with a score of 2, but the score assigned by our system is 3.36. We have again a case of over estimation of the score. This time the review has two contrasting sentences. A very negative one where the user states of having returned the item and a very positive one regarding the accessories. This ambivalence makes the review a borderline case for our model.

We attribute this tendency to overestimate the target to the fact that the model is optimized to minimize the root-mean-square error, this makes the model predict values closer to the average annotated score. While this is acceptable in an academic competition, it's less than ideal in an industrial setting. One way to solve the overestimation problem, without changing the formulation of the error to minimize, would be to balance the data so to have a similar number of occurrences for each score. Sub-sampling the data is unpractical as it would reduce the sample size too drastically. This leaves open only the option to add more samples.

In Table 1 we see the 15 terms most influential on the model. Here we note that most of the terms have a negative connotation. Interestingly, all the bi-grams in the list contain the word *non* (not). Taking in account that the terms reported in the table add up to 37% of the importance of all the features, this highlights the fact that the regressor puts particular attention in the prediction of reviews with a low score even if they are a minority.

3 Preliminary results on Task 1

Task 1 asks to identify terms and phrases that contain an aspect of the customer review when it co-

[3]In English: It turns off on its own when you need to use it with the remote control. (The original sentence contains a two typos.)

[4]In English: There's a better choice for the same price.. I returned it.The accessories are great.

occurs with opinion words that bring information about the sentiment polarity. [5]

For this task we have designed a hybrid model that joins a rule-based approach with machine learning. The main idea is to identify a set of plausible aspects via some pre-defined rules, then use a classifier to filter out the wrong candidates. The rules are defined on POS-Tagging patterns. For example the review

"Ottimo rasoio dal semplice utilizzo."

with annotated as aspect *"semplice"* matches the rule defined by the following pattern

ADJ NOUN PROPN **ADJ** NOUN.

The bold tag indicates the position of the plausible aspect. We have defined a set of about 3000 rules. The rules have been discovered picking the most common POS-Tagging patterns that match the annotated aspects. In particular we have found the position of the aspects in the sentence and selected the POS of close words (three on each side) taking in account the punctuation.

Each aspect found can match one or more rules. The activation of each rule is used as binary feature for the final classifier. The final classifier is implemented using Logistic Regression (Hastie et al., 2001), its target is to predict if each candidate found by the rules is an actual candidate or a false positive.

This preliminary effort achieves a F1-score of 0.340, which is above the baseline (0.255) but below the average score of the submissions (0.504).

4 Conclusions

The submission confirmed the effectiveness of using a simple approach to predict the sentiment score of customer reviews in Italian (Task 3). The approach consists in combining simple word embedding, specifically tri-grams, and a lexicon as SenticNet to build features for Boosted Trees. Our system achieved a competitive error which is lower than the baseline by 0.209 points and higher than the best model by 0.131 points. The error achieved above the average official score by 0.067 points (the estimates includes baseline models).

The submission also highlights that we were able to beat the baseline for Task 1 with a rudimentary approach. We will build upon this approach in our future work.

[5] Detailed description of the task at http://www.di.uniba.it/ swap/ate_absita/task.html

References

[Basile et al.2020] Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

[Cambria et al.2010] Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10. Citeseer.

[Chen and Guestrin2016] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

[Croce et al.2013] Danilo Croce, Francesco Garzoli, Marco Montesi, Diego De Cao, and Roberto Basili. 2013. Enabling advanced business intelligence in divino. In *DART@AI*IA*, pages 61–72.

[de Mattei et al.2020] Lorenzo de Mattei, Graziella de Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020. ATE_ABSITA@EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

[Hastie et al.2001] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Honnibal and Montani2017] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

[Liu et al.2015] Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Twenty-Fourth international joint conference on artificial intelligence*.

[Liu et al.2020] Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. kk2018 at semeval-2020 task 9: Adversarial training for code-mixing sentiment classification. *arXiv preprint arXiv:2009.03673*.

[Pedregosa et al.2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,

M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Poria et al.2014] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, and Alexander Gelbukh. 2014. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

[Poria et al.2016] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

[Rana and Cheah2019] Toqir A Rana and Yu-N Cheah. 2019. Sequential patterns rule-based approach for opinion target extraction from customer reviews. *Journal of Information Science*, 45(5):643–655.

[Sorgente et al.2014] Antonio Sorgente, Giuseppe Vettigli, and Francesco Mele. 2014. An italian corpus for aspect based sentiment analysis of movie reviews. In *First Italian Conference on Computational Linguistics CLiC-it*.

[Zhu et al.2014] Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. 2014. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 443–447.