Ensemble of LSTMs for EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA) (Short Paper)

Mauro Bennici
You Are My GUide
mauro@youaremyguide.com

Xileny Seijas Portocarrero
You Are My GUide
xileny@youaremyguide.com

Abstract

English. In identifying the different emotions present in a review, it is necessary to distinguish the single entities present and the specific semantic relations. The number of reviews needed to have a complete dataset for every single possible option is not predictable.

The approach described starts from the possibility to study the aspect and later the polarity and to create an ensemble of the two models to provide a better understanding of the dataset.

Italiano. Nell'identificazione delle diverse emozioni presenti in una recensione è necessario distinguere le singole entità presenti e le singole relazioni semantiche. Il numero di recensioni necessarie per avere un dataset completo per ogni singola opzione possibile non è predicibile.

L'approccio descritto parte dalla possibilità di creare due modelli diversi, uno per la parte di categorizzazione, e l'altro per la parte di polarità. E di unire i due modelli per ottenere una maggiore comprensione del dataset.

1 Introduction

With the increase in interactions between users and businesses across different channels and different languages, it becomes increasingly difficult for businesses to respond promptly and effectively in an effective manner. Not all activities can have a team dedicated to public relations and often rely on external agencies that do not know the internal operations of the company.

Automating the correct recognition of the various problems can lead to the timely addressing of the same to the persons appointed to solve them.

The research was carried out with the dataset provided within the task called ABSITA, Aspect-based Sentiment Analysis at EVALITA 2018[1] (Basile et al., 2018). The task was a combination of two tasks, Aspect Category Detection (ACD) and Aspect Category Polarity (ACP).

The dataset is a selection of hotel reviews taken in Italian from the portal Booking.com.

2 Description of the system

Each review has been cleaned up by special characters, lemmatized and brought to lowercase with the SpaCy[2] framework.
Generic Italian texts have been used, instead of reviews in the accommodation context to be sure that the model will be suitable for more business models, to generate vectors in fastText[3]. The best one has a dimension of 200, with character n-grams of length 5, a window of size 5 and 10 negatives.

The system is the ensemble of two different models to improve the ability to discover hidden properties (Akhtar et al., 2018).

The first model is a bi-directional Long Short-Term Memory (BI-LSTM).
This model is used for the discernment of the ASPECT.

[1] http://sag.art.uniroma2.it/absita/
[2] https://spacy.io
[3] https://fasttext.cc/

Layer (type)	Output Shape	Param #
e (Embedding)	(None, 100, 200)	1420400
b (Bidirection)	(None, 512)	935936
d (Dense)	(None, 7)	3591

A second BI-LSTM model is used for the discernment of POLARITY.

Layer (type)	Output Shape	Param #
e (Embedding)	(None, 100, 200)	1420400
b (Bidirection)	(None, 512)	935936
d (Dense)	(None, 14)	7182

A dropout and a recurrent_dropout of 0.1.
The optimizer for both is the RMSProp.
The loaded embedding is trainable.
Both the systems use Keras[4] to create the RNN models.

The models were trained and tested with a 5-fold cross-validation with a ratio of 80% training and 20% testing. The best model was automatically saved at each iteration.

A threshold of 0.5 was used on the first model to activate the result of the last layer. In the second model, the threshold was of 0.43.

Aspect Category Detection (ACD)		
micro precision	micro recall	micro F1 score
0.8397	0.8050	0.8204

Table 1: micro precision, micro recall and micro F1 score with the gold dataset.

Aspect Category Polarity (ACP)		
micro precision	micro recall	micro F1 score
0.8138	0.6593	0.7172

Table 2: micro precision, micro recall and micro F1 score with the gold dataset.

The results show that the models are useful to understand the category of a review better than its polarity.

After that we ensemble the two models (Choi et al., 2018) to obtain a system able to overcome the results of every single model in the ACP task reducing the result on the ACD task (table 3).

The ensemble has been created in cascade making sure that a system acts as Attention to the underlying system.
The threshold of activation was a range between 0.45 and 0.55.

A third model, a LightGBM[5] (Bennici and Portocarrero, 2018) was also tested, where the following properties are extracted from the reviews text:

- length of the review
- percentage of special characters
- the number of exclamation points
- the number of question marks
- the number of words
- the number of characters
- the number of spaces
- the number of stop words
- the ratio between words and stop words
- the ratio between words and spaces

and they are joined to the vector created by the bigram and trigram of the text itself at word and character level.
The number of leaves is 250, the learner set as 'Feature', and a the learning rate at 0.04.

The result of the union between the three models could not be submitted to the final evaluation, due to the limit of 2 possible submissions, but reported results higher than 83% in the tests carried out after the release of the complete dataset for ASPECT and 75% for POLARITY.
Also, the inference is faster than the RNN models.

[4] https://keras.io

[5] https://github.com/Microsoft/LightGBM

3 Results

Aspect Category Detection (ACD)			
Runs	micro precision	micro recall	micro F1
Run 1	0.8713	0.7504	0.8063
Run 2	0.8697	0.7481	0.8043

Table 3: micro precision, micro recall and micro F1 score for the submitted ACD subtasks.

Aspect Category Polarity (ACP)			
Runs	micro precision	micro recall	micro F1
Run 1	0.7387	0.7206	0.7295
Run 2	0.7472	0.7186	0.7326

Table 4: micro precision, micro recall and micro F1 score for the submitted ACP subtask.

In the evaluation phase, we can see how the results have given reason to the ensemble of the two results.

It is clear that the ACP task (table 4) is the beneficiary of this process, instead of the ACD one (table 3) that lost more than one point.

The study of the dataset is influenced by the little extension of the training dataset and by the specificity of some terms that could refer to different categories such as the comfort of the room and the quality/price ratio.

Various types of data preparation have also been used, including the preservation of special characters, the shape of words (to better identify cities or places written in capital letters), and some SMOTE functions to increase the number of entries but with poor results and noticeable overfitting.

4 Conclusion

Creating an ensemble of models to bring out various properties of a review gave better results than using a single model in the polarity identification.

The terms used in the review are sometimes misleading and can be used both positively or negatively, and to identify different categories of the hotel.

In the near future, we are ready to create a system to split the text of the review to categorize only a single sentence, or less a single subject or object. In this way, we will be ready to evaluate also the polarity of the single object or subject, and only the terms single related to it to improve the result of the ACP task.

The performance of the system will also be evaluated by replacing all the possible entities with variables known as:

- City
- Museum
- Panoramic Point
- Railway station
- Street

and with a pre-category knew a priori as Breakfast for words like Coffee, Cornetto, and Jam.

The expected result is to reduce the variance of the dataset, to improve the ACD result, and to be able to use the system in production.

Finally, we will evaluate the speed and effectiveness of a CNN model in which the tasks, ASPECT, and POLARITY, can be studied separately and then merged.

Reference

Basile, P., Basile, V., Croce, D., & Polignano, M. (2018). Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA). Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)

Akhtar, M., Ghosal, D., Ekbal, A., Bhattacharyya, P., & Kurohashi, S. (2018, October 15). A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction. Retrieved from https://arxiv.org/abs/1808.01216

Choi, J. Y. and Bumshik, L. (2018).“Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting,” Mathematical Problems in Engineering, vol. 2018, Article ID 2470171, 8 pages. doi: https://doi.org/10.1155/2018/2470171.

Bennici, M. and Seijas Portocarrero, X. (2018). The validity of dictionaries over the time in Emoji prediction. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR.org.