Puccini by mail

Andrea Ierardi
ID: 960188

University of Milan - Data Science and Economics course

Abstract. The aim of the project is the retrieval of Giacomo Puccini letters from official website, developing and application of pre-trained models for letters sentiment polarity classification over the years.

1 Introduction

Giacomo Puccini (22 December 1858 – 29 November 1924) was an Italian composer known primarily for his operas. Regarded as the greatest and most successful proponent of Italian opera after Verdi, he was descended from a long line of composers, stemming to the late-Baroque era. Though his early work was firmly rooted in traditional late-19th-century Romantic Italian opera he later developed his work in the realistic verismo style, of which he became one of the leading exponents. The great interest of Puccini is not only linked to his universal notoriety nor to the constant and still current success of his works, but also to his language, rich in Tuscanisms and inventions. The letters of Puccini can be studied from the perspective of sentiment analysis in order to link the text to several aspects of the temperament of Puccini, the moments of depression and discouragement from which he suffered periodically, his insecurity about his own abilities, certain difficulties in the relationship with his librettists. For this purpose a extensive developing and use of pre-existing classification models for sentiment prediction in Puccini's Letters present in the Ricordi Archive.
Two main models are used:

- SentITA models [1]: Pre-trained Sentiment polarity classification in Italian
- Simple Neural Networks trained on Italian Tweets.

2 Research question and methodology

The Ricordi Archive keeps 350 letters written by Puccini to various recipients of Casa Ricordi. To these are added another 100 letters present in the database but not kept in the archive. In total, therefore, it is about 450 letters to be analyzed.

2.1 Data

Datasets
Two dataset are considered in the study: *The Ricordi Archive* [3] and *Sentipolc-evalita16* [4]. The first is composed by a series of letters from the Ricordi Archive received and sent by Giacomo Puccini. The second is composed by a collection of Italian Tweets with both political and generic topics. The former has been extracted exploiting specific keywords and hashtags marking political topics (topic = 1 in the dataset), while the latter is composed of random tweets on any topic (topic = 0). SentITA models are trained on few datasets (Sentipolc2016 and ABSITA2018). Moreover to obtain a large number of sentiment neutral sentences the model is trained also using extract of around 90,000 Wikipedia sentences and automatically labelled all of them as neutral. Summarizing, the dataset available to train and test the model comprises about 102k sentences of which 7k positives, 7k negatives and 88k neutral.

Data Retrieval
Data retrieval and scraping techniques are applied to automatically extract information from the Ricordi Archive website. The different letters are stored in different pages, an automatic algorithm is constructed in a such way to extract letters, change page, re-extract and so on until the last page is reached. For Sentipolc-evalita16 datasets are in a csv format.

Ricordi Archive Extraction Algorithm
The algorithm uses two main URLs: one for the letter IDs extraction and one for the letter information extraction.
The algorithm starts from pages of the Ricordi Archive filtered by Giacomo Puccini. From the HTML code of this page, IDs are extracted and redirect to the next page containing new IDs. The next pages are obtained changing the final number of the URL. The website allows the exceeding of the number of pages even though it will results in a empty page. For this reason, the algorithm stops whenever the resulting page contains an empty table. Once letter IDs are extracted, the algorithm iterate each ID and passing it as parameter of the URL for letter visualisation. Then, for each letter it extracts all information present in the page referring to specific HTML fields: id, text , source, receiver, date, place, volume, volume signature, year, page, number of pages, named people, named works, named places, named theaters, typology, sub-typology, writing, language.

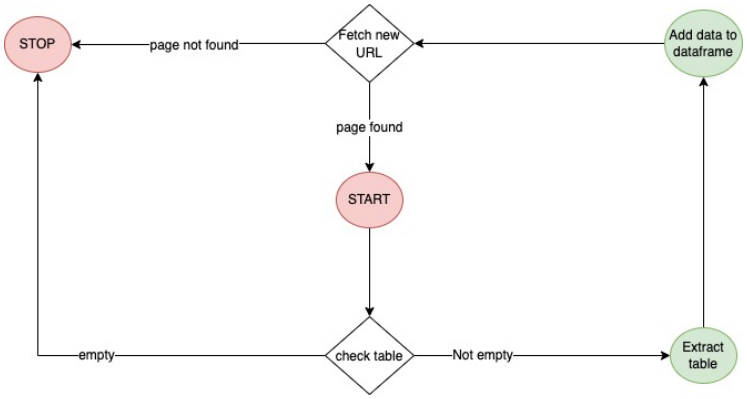

Fig. 1. Extraction Algorithm Flow Chart

Preprocessing
The Sentita used datasets are divided in train and test sets. The library default pre-processing is used: stop words removal, tokenization, padding sequences.
For the custom model the pre-processing approach is different: stemming, stop words removal and tokenization. Then vectorization (bigrams) of the document to apply features selection and standard scaler.

2.2 Models

Neural Networks
A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural Networks are trained with GridSearch CV hyperparameters tuning. A simple architecture is used as in Figure 2: 1 Dense layer, 1 Dropout Layer, another 1 Dense Layer and the last Dense layer as output. Two target type of Neural Networks are trained:

- 2 – *sentiments* model: positive and negative.
- 4 – *sentiments* model: positive, negative, neutral and both negative and positive.

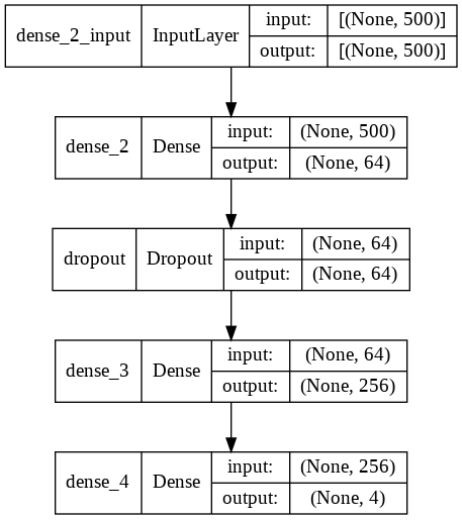

Fig. 2. Simple Neural Network Architecture

Sentita Sentiment and Emotion Classifier
SentITA [2] Python library is used for Sentiment polarity classification in Italian. The deep learning model applied is a Bidirectional LSTM-CNN that operates at word level. The model receives in input a word embedding representation of the single words and outputs two signals ranging between 0 and 1, one for positive sentiment detection and one for negative sentiment detection. The two signals can be triggered both by the same input sentence if this contain both positive and negative sentiment (e.g. "The food is very good, but the location isn't nice"). The model has a reduced number of trainable parameters, around 51k, and dropout to reduce overfitting. The model is written in Python 3.6 and is implemented in Keras 2.2.4 with Tensorflow 1.11 backend. In Figure 3 is possible to see the Sentita Neural Network Architecture.

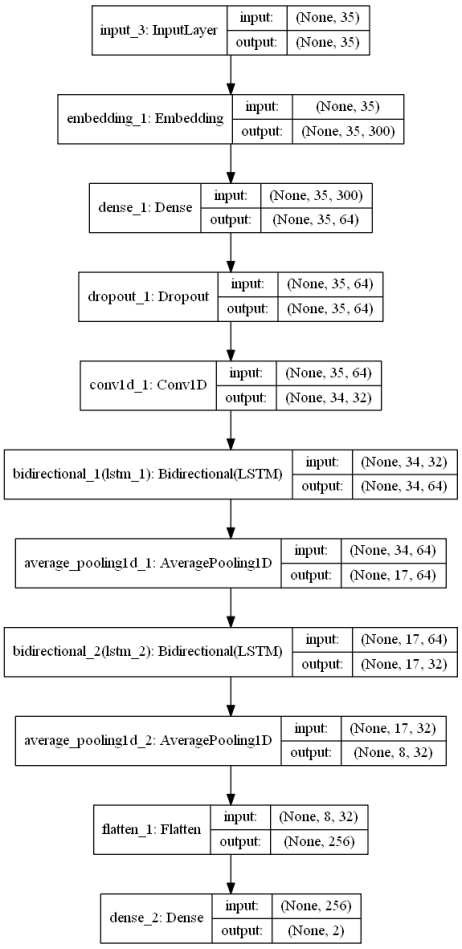

Fig. 3. Sentita Sentiment and Emotion Classifier Neural Network Architecture

3 Experiments

3.1 Word Cloud

Word Cloud visualization allows the identification of the most used words in Puccini's letters. Most common words are related to music, Puccini works, friends and collaborations.
From Figure 4 is possible to find Puccini's work such as Bohème, Manon Lescaut. Doge is the nickname of Giacomo assigned by Giulio Ricordi. Tito is the name of Tito II Ricordi. Luigi Illica and Giuseppe Giacosa are famous librettists whom Puccini worked. Puccini studied at the conservatory for three years, sharing a room with Pietro Mascagni. Leopoldo Mugnone friend of Giacomo Puccini, of whom he edited the "premiere" of Tosca (1900). Common Italian large cities such as Milan, Rome, Turin and Lucca. A place like Torre Lago [5] is present. From 1891 onwards, Puccini spent most of his time, at Torre del Lago, a small community about fifteen miles from Lucca. Torre del Lago was the primary place for Puccini to indulge his love of hunting. "I love hunting, I love cars: and for these things, in the isolation of Torre del Lago, I keep the faith." Moreover, other common words refers to the music such as: music ("musica"), score ("partitura"), verse ("versi"), tempo ("tempo"), scene ("scena").

Fig. 4. Word Cloud

3.2 Neural Networks model results

4-sentiments Neural Networks
Neural Networks trained using a target of 4 sentiments(positive, neutral, negative, both positive and negative) have an accuracy on the test set of 46.22 and $f1$ score of 0.467.
Figure 5 shows the number of predicted letters that are categorized as negative, positive, neutral and both positive and negative. 63% are categorized as negative, 21% positive, 15% neutral and less than 1% as both.

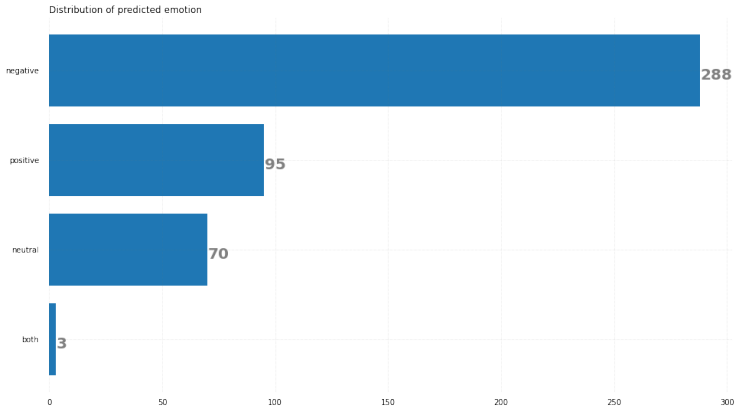

Fig. 5. Predicted sentiment with 4-sentiments model

Figure 6 shows the number of predicted letters that are categorized as negative, positive, neutral and both with a focus on the year. Most of the letters are categorized as negative over the years, while positive and neutral have similar numbers. Both predicted class (positive and negative) is present only during three years: 1888, 1898 and 1900.

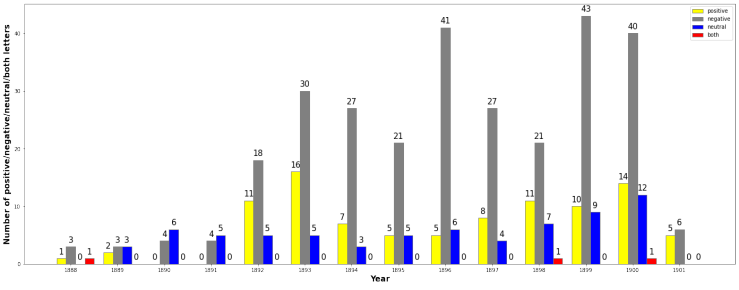

Fig. 6. Predicted sentiments by 4-sentiments Neural Network each year

2-sentiments Neural Networks
Neural Networks trained only using a target of 2 sentiments (positive and negative) have an accuracy on the test set of 61.89 and $f1$ score of 0.627.
Figure 7 shows the number of predicted letters that are categorized as negative or positive by the model. 61% are categorized as negative, 39% positive.

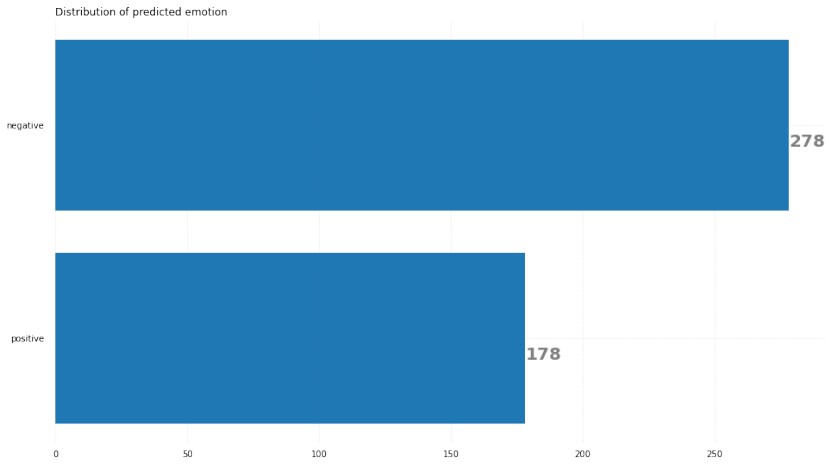

Fig. 7. Predicted sentiment by 2-sentiments Neural Network

Figure 8 shows the number of predicted letters that are categorized as negative and positive over the years. Most of the letters are categorized as negative, while positive letters are in majority only in the years 1889 and 1901.

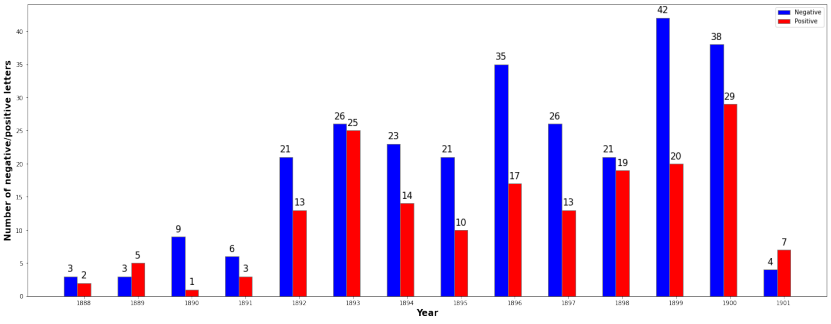

Fig. 8. Predicted sentiments by 2-sentiments Neural Network each year

3.3 SentITA model results

SentITA sentiments analysis
SentITA model has $f1$ score of 0.85 on the same test set of the Neural Networks models. Figure 9 shows the sentiment predicted by Sentita model. It predicted mostly negative sentiment 240 over 456 letters and 216 positive over 456 letters. 53% negative, 47% positive.

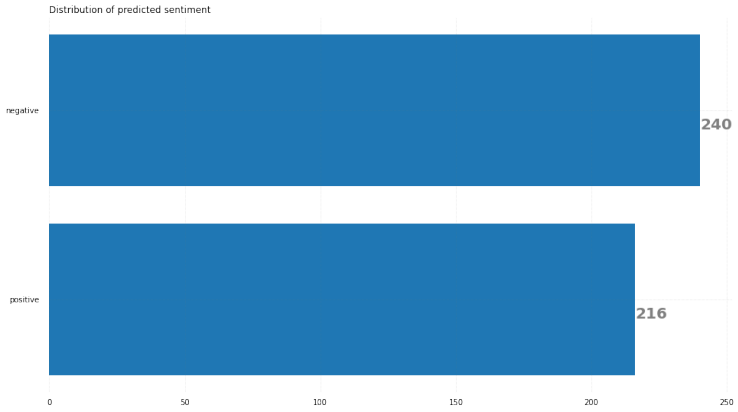

Fig. 9. Sentita predicted sentiment in Puccini's letters

Figure 10 shows the predicted sentiments by Sentita model over the years. Most categorized letters concentrate in the 1893-1900 years range. In this case there is a majority of positive for 6 years, while a majority of negative for 6 and equal number for 2 years.

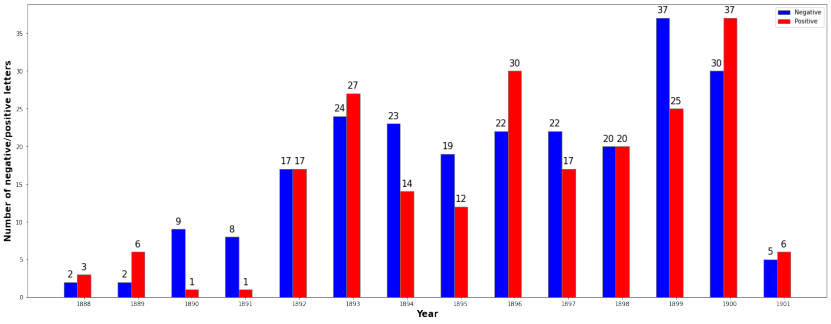

Fig. 10. Sentita predicted sentiments each year

SentITA emotions analysis
Figure 11 shows the predicted emotions by Sentita model. It predicted mostly joy and sadness emotions while a little number of letters are categorized with fear and anger emotions.
55% Joy, 36% Sadness, 5% Anger, 4% Fear.

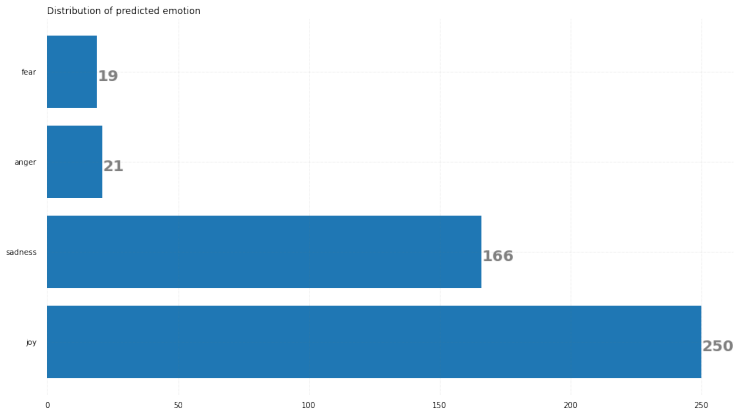

Fig. 11. Sentita predicted emotions in Puccini's letters

Figure 12 shows the model predicted emotions by Sentita model with respect to the year. Figure 11 shows that Joy and Sadness emotions are the most present with respect to the other.

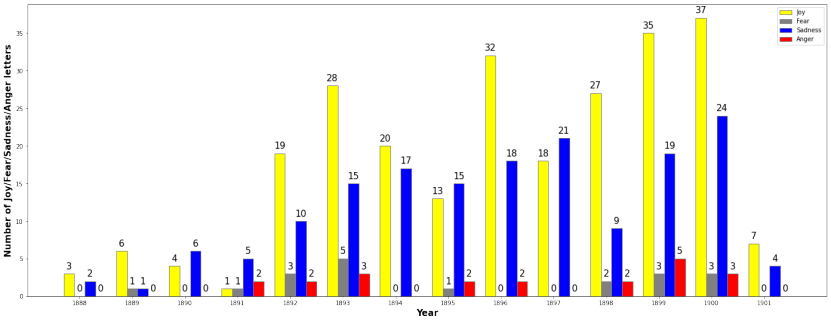

Fig. 12. Sentita predicted emotions each year

SentITA emotions and sentiments analysis
Figure 13 shows the model predicted emotions and sentiments by Sentita model. It predicted mostly joy-positive and sadness-negative combinations.
42% Joy-positive, 31% Sadness-negative, 13% Joy-negative, 5% Sadness-positive, 5% Anger-negative, 4% Fear-negative.

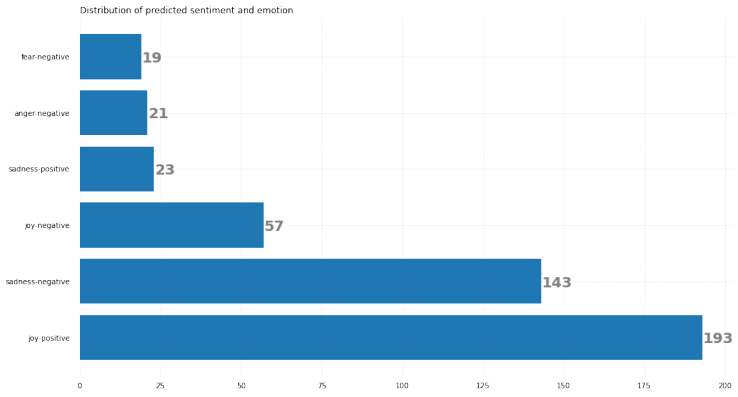

Fig. 13. Sentita predicted emotion and sentiment in Puccini's letters

3.4 Model comparison

Comparison between Sentita model and Neural Networks with 2 sentiments model.
Figure 14 shows the number of letters predicted with the same sentiment by the two model and those that are not. Most of the letters are predicted with the same sentiment, while 92 are predicted as negative by 2-sentiments model and positive by Sentita model. 54 are predicted as positive by 2-sentiment model and negative by Sentita. In conclusion 68% of the letters are classified by the same sentiment by the two models.

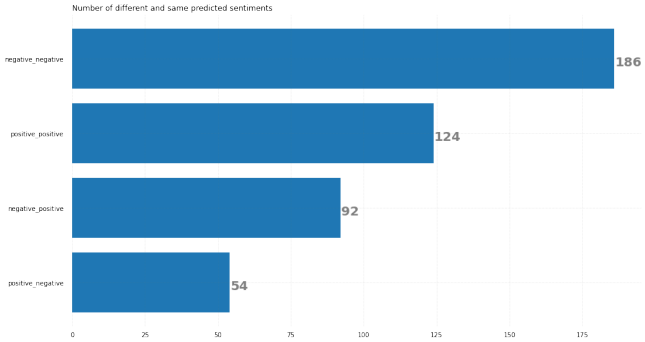

Fig. 14. Comparison between 2-sentiments Neural Networks and Sentita models

Figure 15 shows the number of letters predicted with same and different sentiment between Sentita and 2-sentiment model over the years.
The two model discord in the 1888 and 1902 over 60% of the time, while for the other years the percentage stays under 40% of discordance.

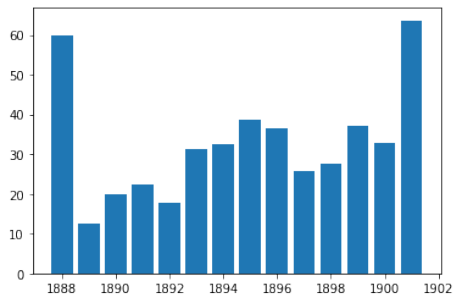

Fig. 15. Percentage of discordance in predicted sentiments

4 Conclusion and Next Steps

The focus of the study is the construction of algorithm for data retrieval, preprocessing and developing of sentiment polarity prediction models for Giacomo Puccini's letters. Sentipolc dataset is used for model training and testing, while data retrieval algorithm techniques are used for extracting letters information and text from Archivio Ricordi website. Two different target models are constructed: 2-sentiments and 4-sentiments Neural Networks. Moreover, SentITA pre-trained model is used for results comparison and for understanding if a simple Neural Networks may perform similarly. In fact, the models performed similar 60% of the time as shows in Figure 14. SentITA performs better, due to the large training datasets combination and deep study but 2-sentiments classifier may be a substitute. Emotion comparison and 4-sentiments comparison cannot be taken into consideration, but is still interesting to see how the models classified the letters.

Future steps may be taking into consideration the developing of a simple Neural Networks for emotions prediction and compare it with Sentita emotion classifier or extension of the dataset for the simple Neural Network. Moreover, another steps would be to add some feature extracted in the retrieval phase (source, receiver, date, place, volume, volume signature, year, page, number of pages, named people, named works, named places, named theaters, typology, sub-typology, writing, language.) to train the model and understand if they can contribute in the sentiment prediction.

References

1. NICOLA, Giancarlo. Bidirectional Attentional LSTM for Aspect Based Sentiment Analysis on Italian In: EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples [online]. Torino: Accademia University Press, 2018 (creato il 16 décembre 2021). http://books.openedition.org/aaccademia/4550. ISBN: 9788831978699. DOI: https://doi.org/10.4000/books.aaccademia.4550.
2. https://github.com/NicGian/SentITA
3. https://www.archivioricordi.com/
4. BARBIERI, Francesco ; et al. Overview of the Evalita 2016 SENTIment POLarity Classification Task In: EVALITA. Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 7 December 2016, Naples [online]. Torino: Accademia University Press, 2016 (creato il 16 décembre 2021). Disponibile su Internet: ¡http://books.openedition.org/aaccademia/1992¿. ISBN: 9788899982553. DOI: https://doi.org/10.4000/books.aaccademia.1992. http://www.di.unito.it/ tutreeb/sentipolc-evalita16/
5. https://en.wikipedia.org/wiki/Giacomo_Puccini