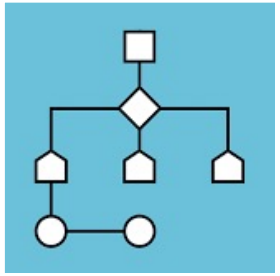
Puccini by mail
Sentiment polarity prediction

Andrea Ierardi

Student’s ID: 960188

Project aim



Build Algorithm for data retrieval from Ricordi Archive

Build prediction models and use pre-trained model to detect sentiment polarity in Puccini’s letters

Comparison between models

Datasets

Two datasets:

- The Ricordi Archive:
  - Extracted with a retrieval algorithm
  - 500 letters received and sent by Giacomo Puccini
- Sentipolc- evalita16:
  - Collection of Italian Tweets
  - Available in csv format in the website
  - 7000 examples for training, 4000 example for testing

Ricordi Archive Extraction Algorithm

Two main URLs:

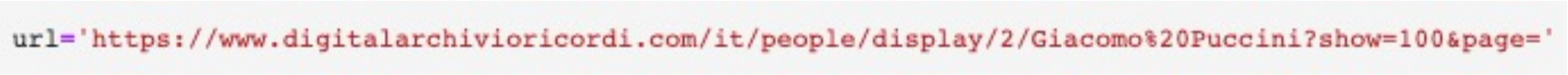
- IDs URL: used for extraction of the letter IDs
- Letters visualization URL: used for the extraction of information for the selected letter's ID.

Extraction of the information using HTML fields:

- **<div> The Content Division element** HTML
- Separating different information searching for specific Div IDs
- For the additional info newline separator have been considered

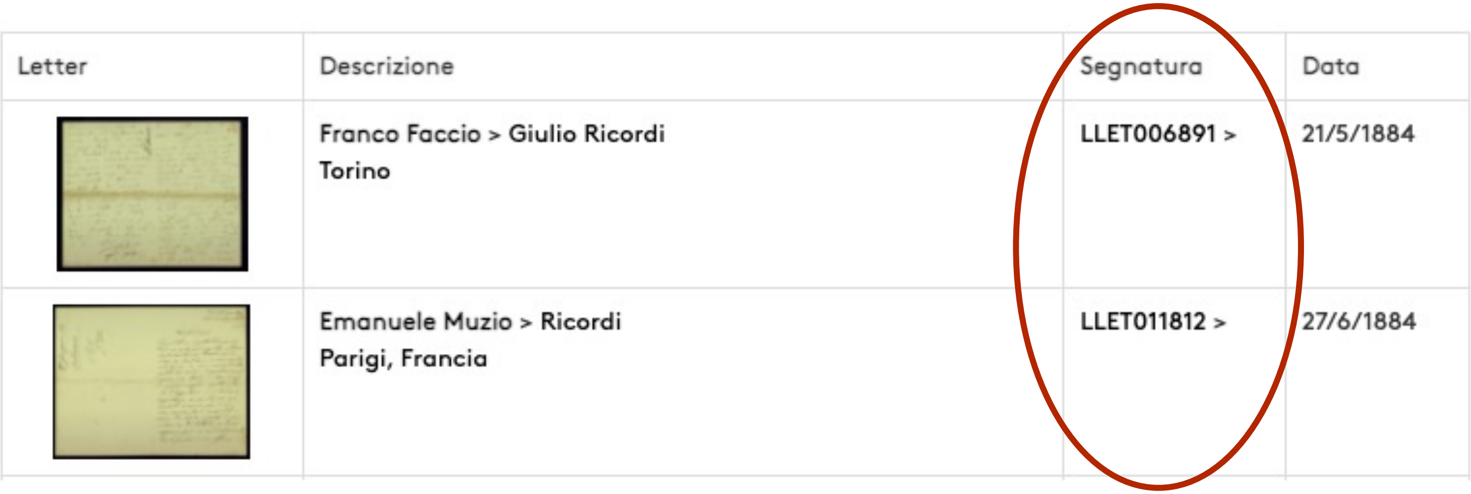
ID Extraction

1) For the iteration of the different pages we pass an incremental number as parameter



```
url='https://www.digitalarchivioricordi.com/it/people/display/2/Giacomo%20Puccini?show=100&page='
```

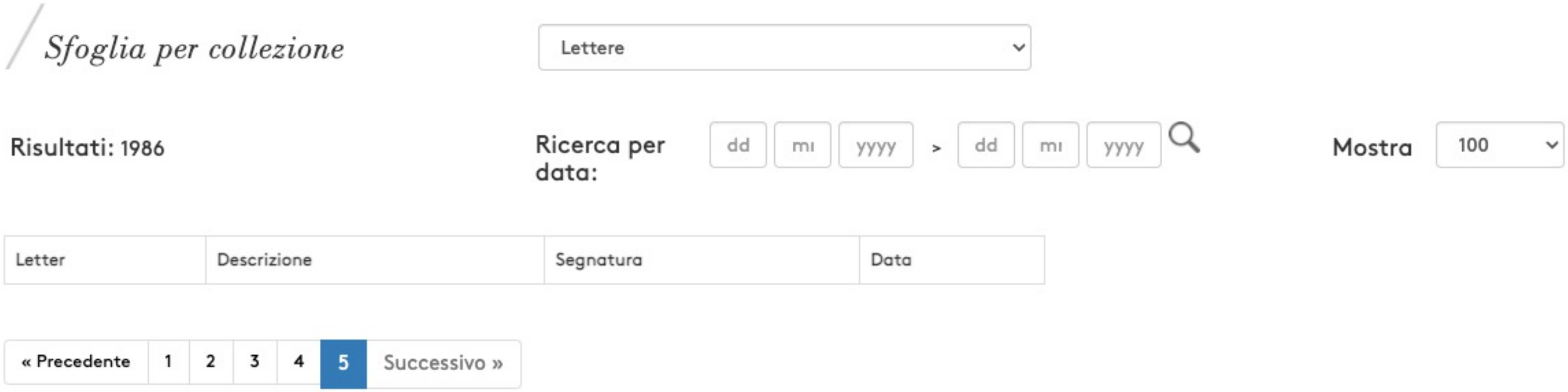
2) For each letter we extract the "segnatura" field which contains the letter's ID



Letter	Descrizione	Segnatura	Data
	Franco Faccio > Giulio Ricordi Torino	LLET006891 >	21/5/1884
	Emanuele Muzio > Ricordi Parigi, Francia	LLET011812 >	27/6/1884

ID Extraction Problems

1) Setting page number high (for example) 200 a empty table is returned



Sfoglia per collezione

Lettere

Risultati: 1986

Ricerca per data: dd mm yyyy > dd mm yyyy

Mostra 100

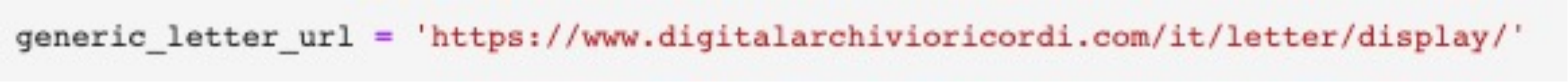
Letter	Descrizione	Segnatura	Data

« Precedente 1 2 3 4 5 Successivo »

SOLUTION: if the returned table is empty all the possible IDs have been examined and the first phase of the algorithm stops

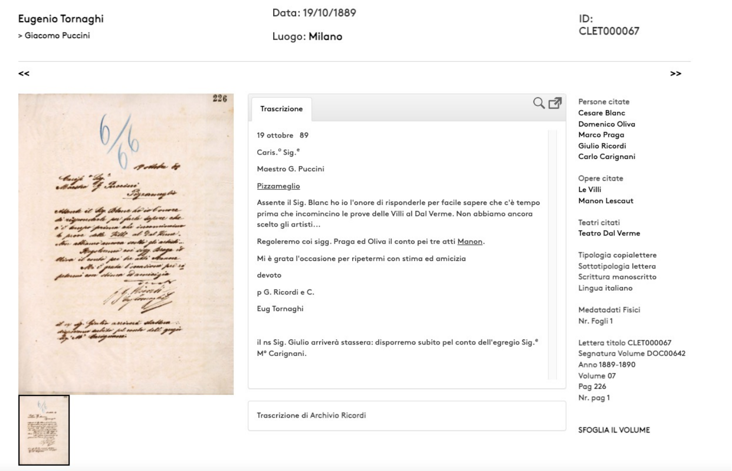
Information Extraction

1) Retrieved IDs in the first phase are appended in the second URL



```
generic_letter_url = 'https://www.digitalarchivioricordi.com/it/letter/display/'
```

2) Appending the letter's ID at the end of the we obtain the letter information



Eugenio Tornaghi
> Giacomo Puccini

Data: 19/10/1889

Luogo: Milano

ID:
CLET000067

19 ottobre 89

Caris.° Sig.^{e}

Maestro G. Puccini

Pizzameglio

Assente il Sig. Blanc ho io l'onore di risponderle per facile sapere che c'è tempo prima che incomincino le prove delle Villi al Dal Verme. Non abbiamo ancora scelto gli artisti...

Regoleremo coi sigg. Praga ed Oliva il conto pei tre atti Manon.

Mi è grata l'occasione per ripetermi con stima ed amicizia

devoto

p G. Ricordi e C.

Eug Tornaghi

il ns Sig. Giulio arriverà stassera: disporremo subito pel conto dell'egregio Sig.^{e} M° Carignani.

Trascrizione di Archivio Ricordi

Persone citate
Cesare Blanc
Domenico Oliva
Marco Praga
Giulio Ricordi
Carlo Carignani

Opere citate
Le Villi
Manon Lescaut

Teatri citati
Teatro Dal Verme

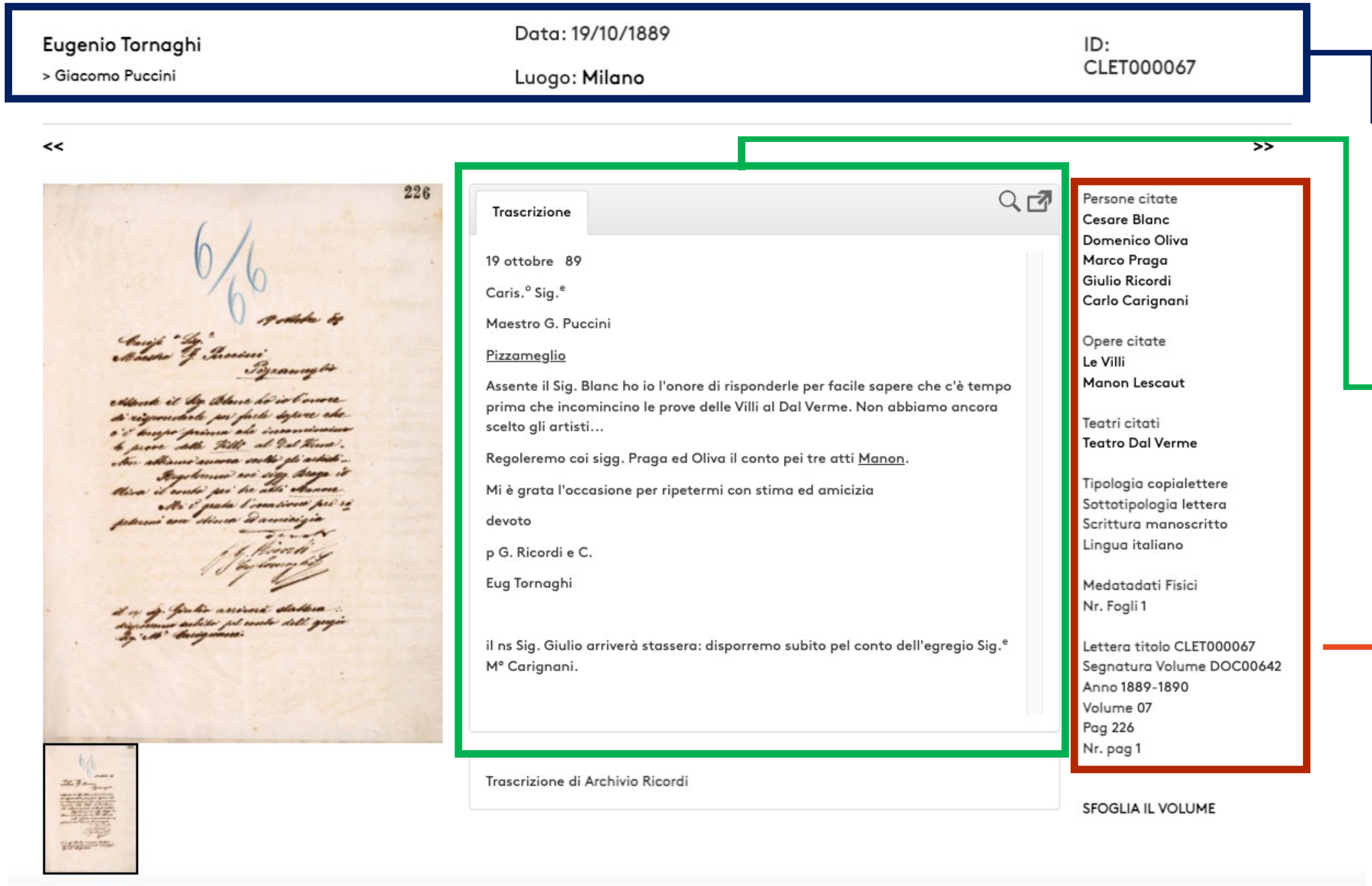
Tipologia copialettere
Sottotipologia lettera
Scrittura manoscritto
Lingua italiano

Medatadati Fisici
Nr. Fogli 1

Lettera titolo CLET000067
Segnatura Volume DOC00642
Anno 1889-1890
Volume 07
Pag 226
Nr. pag 1

SFOGLIA IL VOLUME

Information Extraction



Eugenio Tornaghi
> Giacomo Puccini

Data: 19/10/1889

Luogo: Milano

ID:
CLET000067

19 ottobre 89

Caris.° Sig.e

Maestro G. Puccini

Pizzameglio

Assente il Sig. Blanc ho io l'onore di risponderle per facile sapere che c'è tempo prima che incomincino le prove delle Villi al Dal Verme. Non abbiamo ancora scelto gli artisti…

Regoleremo coi sigg. Praga ed Oliva il conto pei tre atti Manon.

Mi è grata l'occasione per ripetermi con stima ed amicizia

devoto

p G. Ricordi e C.

Eug Tornaghi

il ns Sig. Giulio arriverà stassera: disporremo subito pel conto dell'egregio Sig.e M° Carignani.

Trascrizione di Archivio Ricordi

Persone citate
Cesare Blanc
Domenico Oliva
Marco Praga
Giulio Ricordi
Carlo Carignani

Opere citate
Le Villi
Manon Lescaut

Teatri citati
Teatro Dal Verme

Tipologia copialettere
Sottotipologia lettera
Scrittura manoscritto
Lingua italiano

Medatadati Fisici
Nr. Fogli 1

Lettera titolo CLET000067
Segnatura Volume DOC00642
Anno 1889-1890
Volume 07
Pag 226
Nr. pag 1

SFOGLIA IL VOLUME

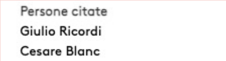
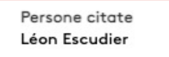
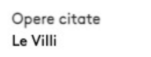
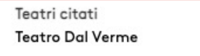
1) Main Information

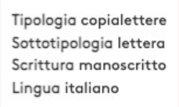
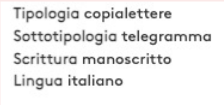
2) Transcribed Text

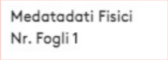
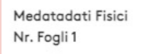
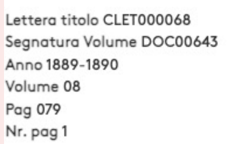
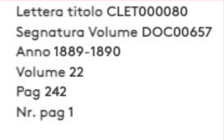
3) Additional Information

Information Extraction Problems

1) Discordance between letters in the additional information field:

Div id	Page 1	Page 2
`letter-show-details-named-people`	 Persone citate Giulio Ricordi Cesare Blanc	 Persone citate Léon Escudier
`letter-show-details-named-works`	 Opere citate Le Villi	NULL
`letter-show-details-named-teatri`	 Teatri citati Teatro Dal Verme	NULL

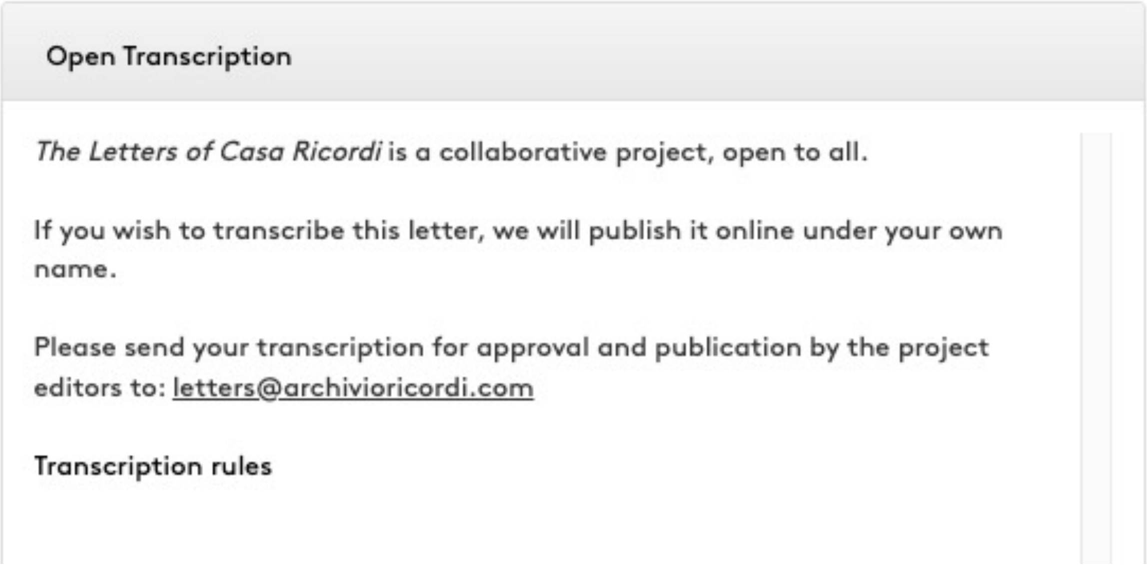
`letter-show-details-tipologia`
`letter-show-details-sottotipologia`
`letter-show-details-scrittura`
`letter-show-details-linuga`
	 Tipologia copialettere
	Sottotipologia lettera
	Scrittura manoscritto
	Lingua italiano
		 Tipologia copialettere
		Sottotipologia telegramma
		Scrittura manoscritto
		Lingua italiano

`letter-show-details-metadati-fisici`	 Medatadati Fisici Nr. Fogli 1	 Medatadati Fisici Nr. Fogli 1
`Outside a div`	 Lettera titolo CLET000068 Segnatura Volume DOC00643 Anno 1889-1890 Volume 08 Pag 079 Nr. pag 1	 Lettera titolo CLET000080 Segnatura Volume DOC00657 Anno 1889-1890 Volume 22 Pag 242 Nr. pag 1

SOLUTION: set dataframe cell with null value if information is missing

Information Extraction Problems

1) Letter's text in some cases is not already transcribed from the original



Open Transcription

The Letters of Casa Ricordi is a collaborative project, open to all.

If you wish to transcribe this letter, we will publish it online under your own name.

Please send your transcription for approval and publication by the project editors to: letters@archivioricordi.com

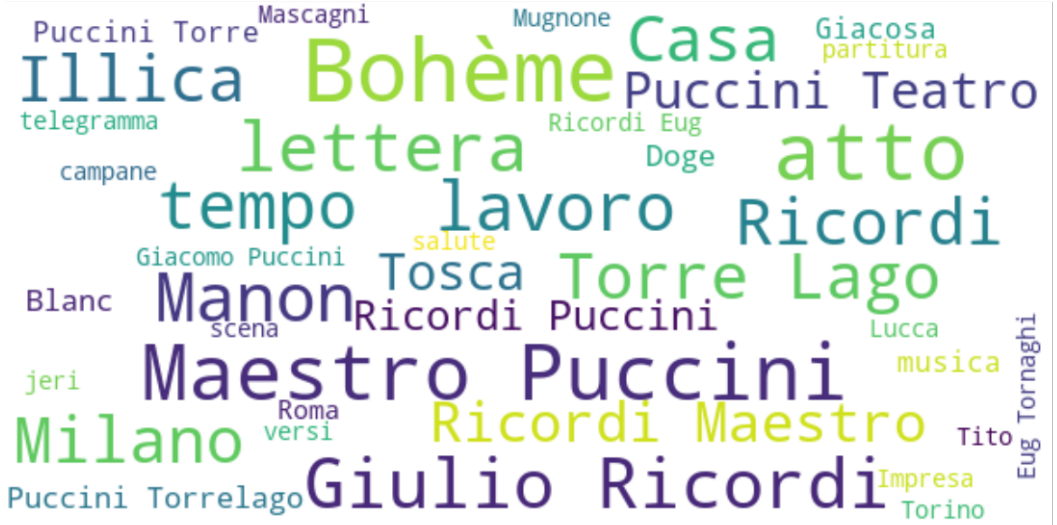
Transcription rules

SOLUTION: Skip this letter in the algorithm

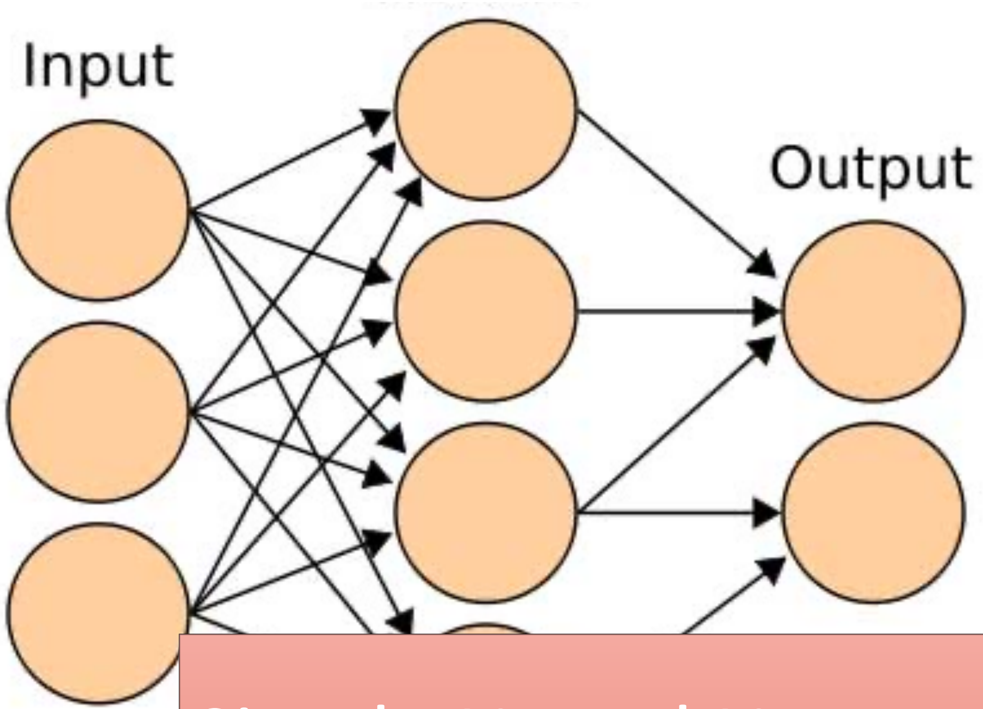
Letters Word Cloud

Most **common words** are related to music, Puccini works, friends and collaborations:

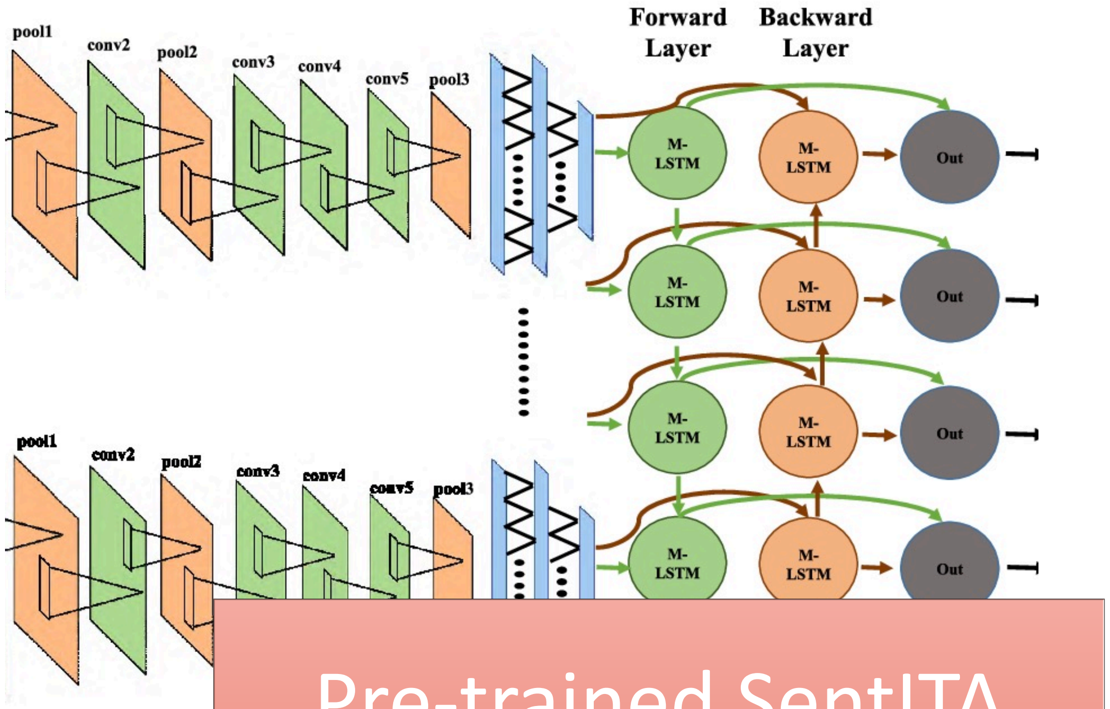
- **Works**:
  - Bohème, Manon Lescaut.
- **Friends** and **colleagues**:
  - Tito is the name of Tito II Ricordi.
  - Luigi Illica and Giuseppe Giacosa are famous librettists whom Puccini worked.
  - Puccini conservatory room mate Pietro Mascagni.
  - Leopoldo Mugnone
- **Italian cities**:
  - Milan, Rome, Turin and Lucca
  - Torre lago: a small community nearby Lucca where from 1891 onwards Puccini, Puccini spent most of his time
- **Opera terms**:
  - Music, score, verse, tempo, scene



Models



Simple Neural Networks

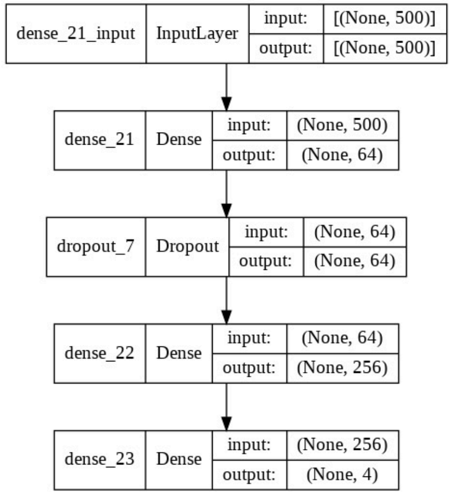


Pre-trained SentITA

Neural Networks Architecture

- **Two types** of Neural Networks have been constructed based on the number of sentiments in the target variable:
  - **2-sentiments Neural Networks**: Positive and Negative
  - **4-sentiments Neural Networks**: Positive, Negative, Neutral, Both Positive and Negative

- Simple Neural Network architecture



dense_21_input | InputLayer | input: [(None, 500)] | output: [(None, 500)]

dense_21 | Dense | input: (None, 500) | output: (None, 64)

dropout_7 | Dropout | input: (None, 64) | output: (None, 64)

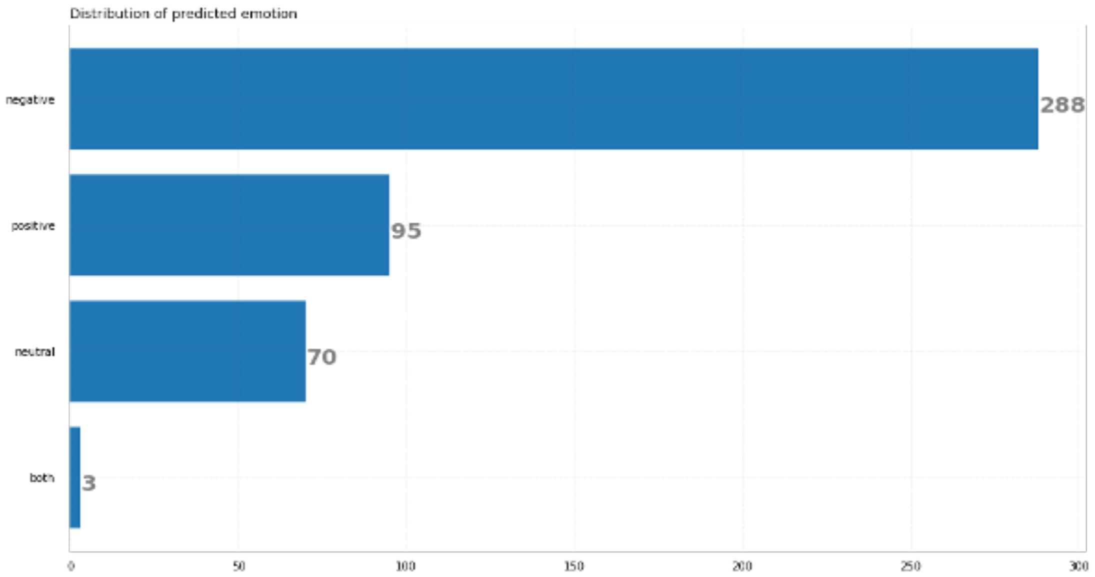
dense_22 | Dense | input: (None, 64) | output: (None, 256)

dense_23 | Dense | input: (None, 256) | output: (None, 4)

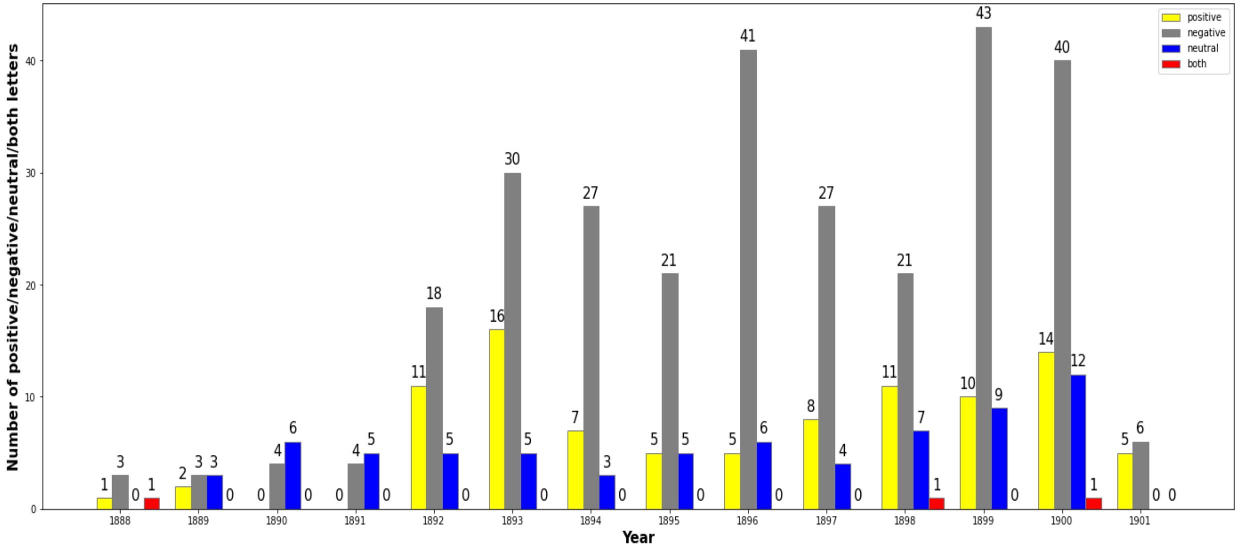
Neural Networks 4-sentiment

- SentiPolc Test set result:
  - F1 score: 0.467 - Accuracy: 46.22 %
- Model result on Puccini’s letters:

Predictions with 4-sentiments model



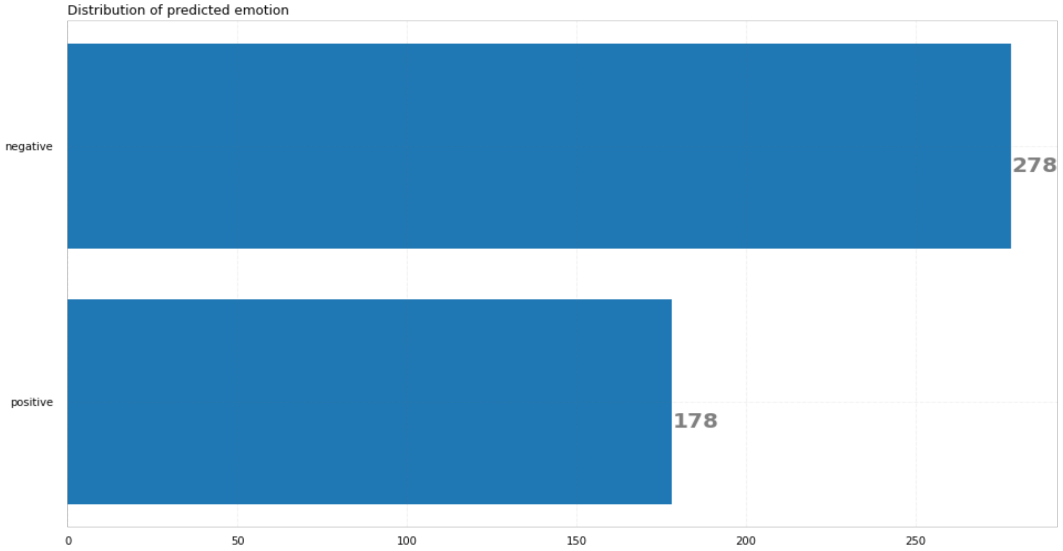
Predictions with 4-sentiments model over the years



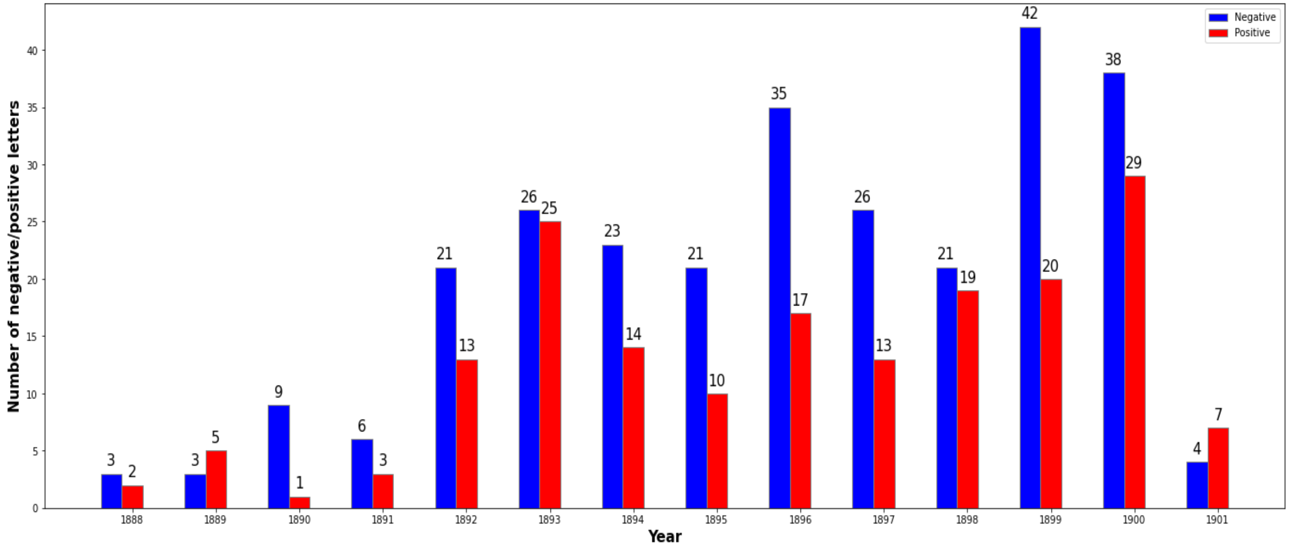
Neural Networks 2-sentiment

- SentiPolc Test set result:
  - F1 score: 0.627 - Accuracy: 61.89 %
- Model result on Puccini’s letters:

Predictions with 2-sentiments model

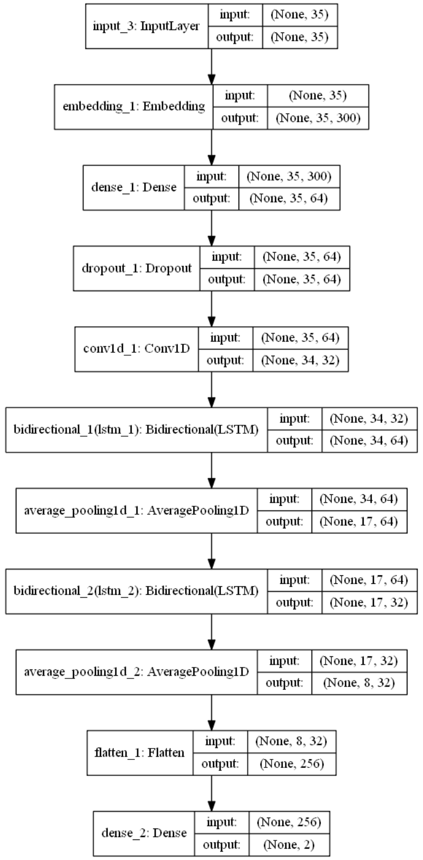


Predictions with 2-sentiments model over the years



SentITA Architecture

- The model receives in input a word embedding representation of the single words
- Trained on few datasets (Sentipolc2016, AB- SITA2018 + Wikipedia).
- Train and test the model comprises about 102k sentences of which 7k positives, 7k negatives and 88k neutral.
- Bidirectional LSTM-CNN Neural Network architecture



input_3: InputLayer | input: (None, 35) | output: (None, 35)

embedding_1: Embedding | input: (None, 35) | output: (None, 35, 300)

dense_1: Dense | input: (None, 35, 300) | output: (None, 35, 64)

dropout_1: Dropout | input: (None, 35, 64) | output: (None, 35, 64)

conv1d_1: Conv1D | input: (None, 35, 64) | output: (None, 34, 32)

bidirectional_1(lstm_1): Bidirectional(LSTM) | input: (None, 34, 32) | output: (None, 34, 64)

average_pooling1d_1: AveragePooling1D | input: (None, 34, 64) | output: (None, 17, 64)

bidirectional_2(lstm_2): Bidirectional(LSTM) | input: (None, 17, 64) | output: (None, 17, 32)

average_pooling1d_2: AveragePooling1D | input: (None, 17, 32) | output: (None, 8, 32)

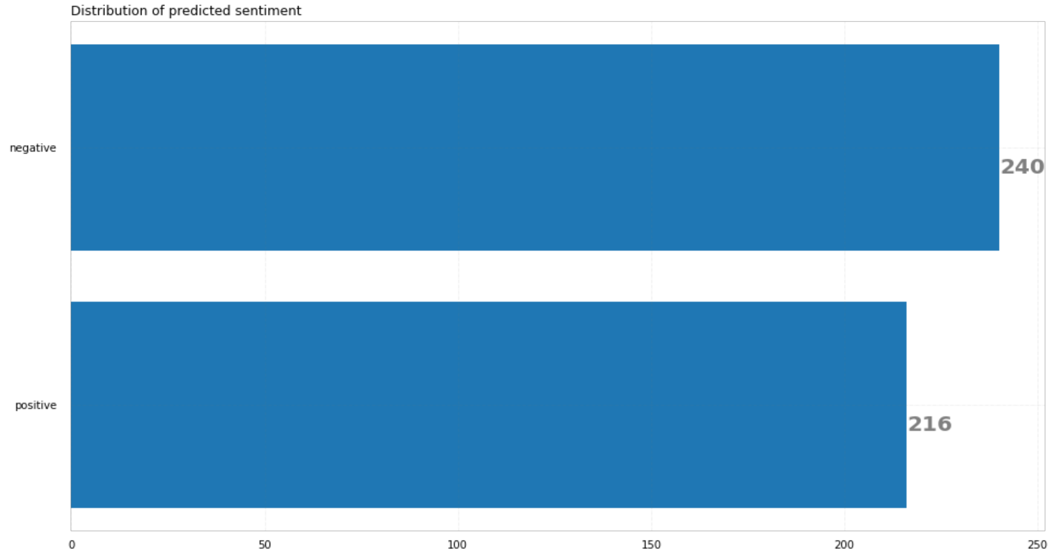
flatten_1: Flatten | input: (None, 8, 32) | output: (None, 256)

dense_2: Dense | input: (None, 256) | output: (None, 2)

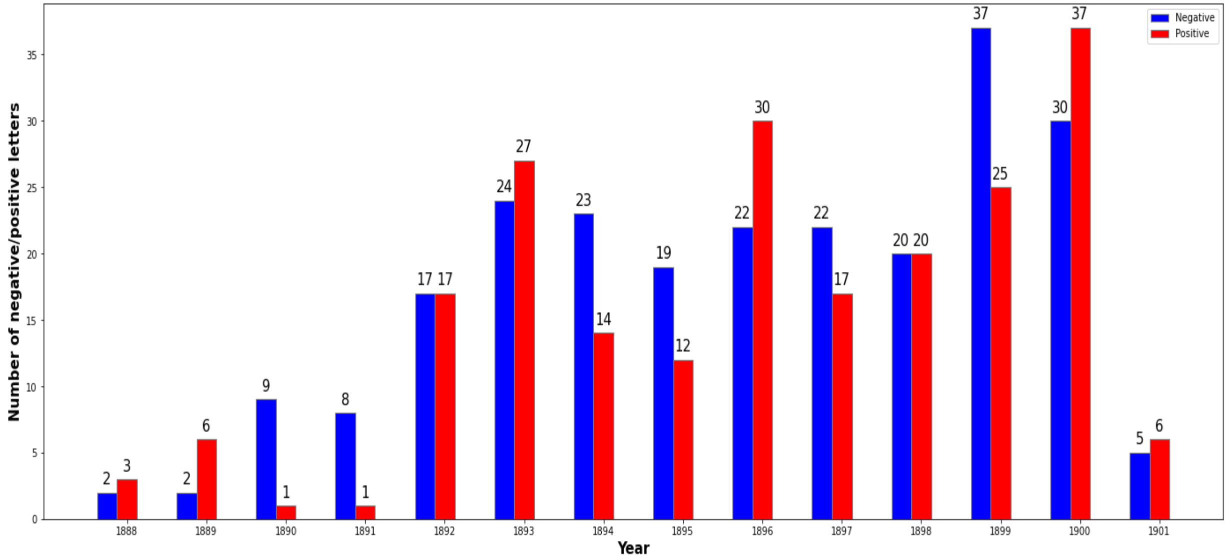
SentITA sentiment polarity

- SentiPolc Test set result:
  - F1 score: 0.85
- Model result on Puccini’s letters:

Predicted sentiment with SentITA model



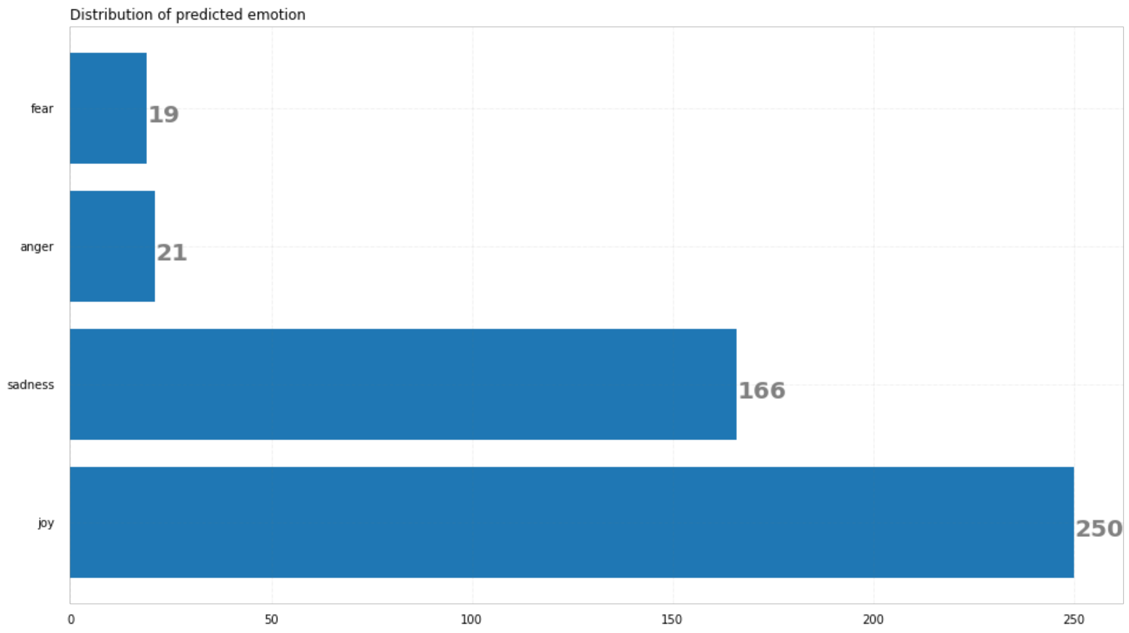
Predicted sentiment with SentITA model over the years



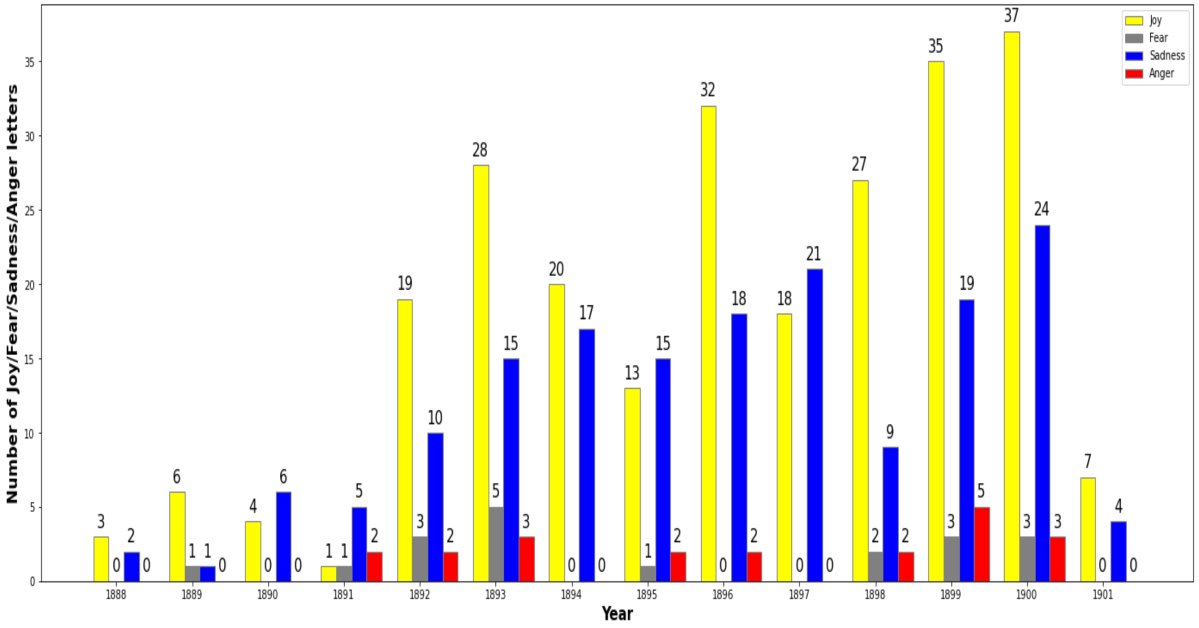
SentITA emotion

- There is not information about emotions in SentiPolc dataset
- Model result on Puccini's letters:

Predicted sentiment with SentITA model



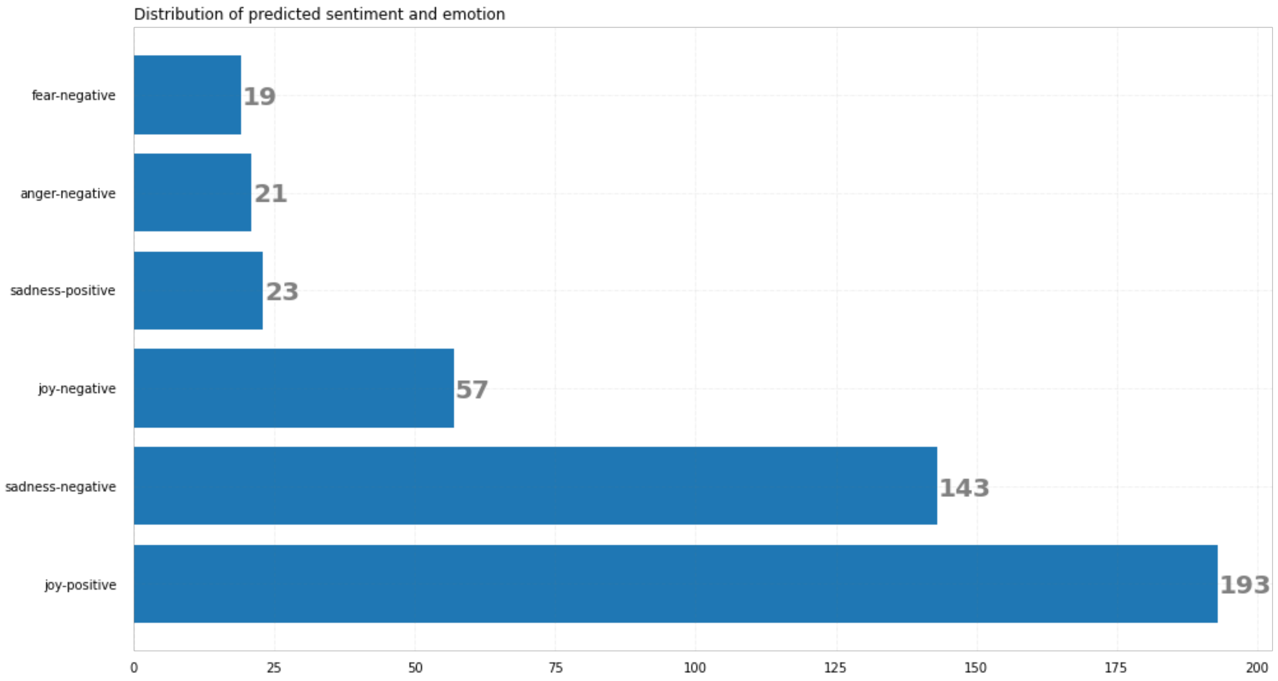
Predicted sentiment with SentITA model over the years



SentITA emotion and sentiment

- There is not information about emotions in SentiPolc dataset
- Model result on Puccini’s letters:

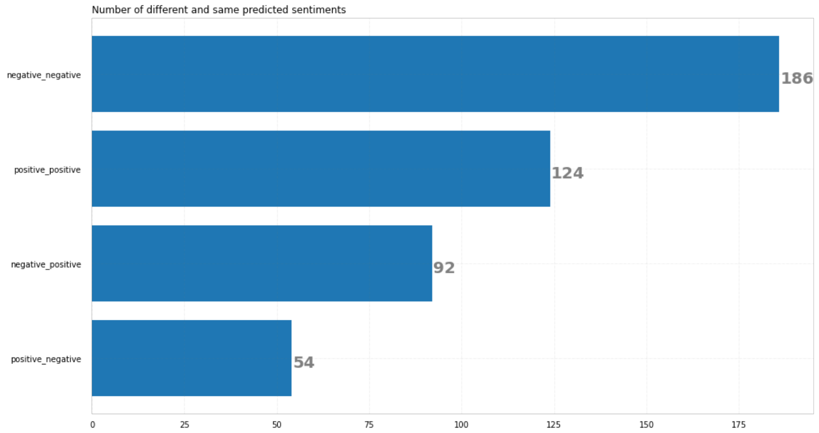
Predicted sentiment with SentITA model



Neural Networks vs. SentITA

- Comparison between Simple Neural Networks model and SentITA model
- Only the sentiment polarity model have been compared
- Model result on Puccini’s letters comparison:

Comparison between 2-sentiments Neural Networks and Sentita models

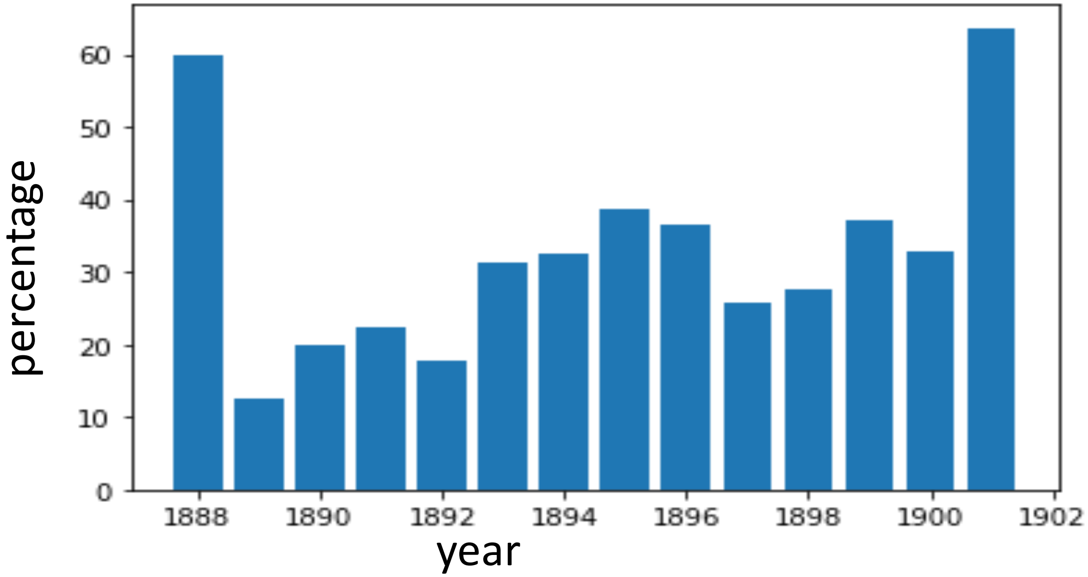


- Most of the letters are predicted with the same sentiments
- 68% of the letters are classified by the same sentiment by the two models

Neural Networks vs. SentITA

- Comparison between Simple Neural Networks model and SentITA model
- Model result on Puccini’s letters comparison:

Percentage of discordance in predicted sentiments - Neural Networks vs. SentITA



- Most letters are distribuited in the years in the middle
- Percentage is higher in 1888 and 1902 since number of letters in that years is lower
- Percentage discordance stays under 40% in the other years

Conclusions & Next Steps

- Conclusions
  - Retrieval algorithm worked well
  - SentITA performs better, due to the large training datasets combination but 2-sentiments is much faster to train and for this reason it may be a simple and reliable baseline
- Next Steps:
  - Develope a model for emotions detection
  - Use also other features retrieved to train the model (source, receiver, date, place, volume, volume signature, year, page, number of pages ecc.)

Questions?