AriEmozione 2.0: Identifying Emotions in Opera Verses and Arias

Shibingfeng Zhang*
Saarland University

Francesco Fernicola**
Università di Bologna

Federico Garcea†
Università di Bologna

Paolo Bonora‡
Università di Bologna

Alberto Barrón-Cedeño§
Università di Bologna

We present the task of identifying the emotions conveyed by the lyrics of Italian opera arias. We shape the task as a multi-class supervised problem, considering the six emotions from Parrot's tree: love, joy, admiration, anger, sadness, and fear. We manually annotated an opera corpus with $2.5k$ *instances at the verse level and experimented with different classification models and representations to identify the expressed emotions. Our best-performing models consider character* 3*-gram representations and reach relatively low levels of macro-averaged* F_1*. Such performance reflects the difficulty of the task at hand, partially caused by the size and nature of the corpus: relatively short verses written in 18th-century Italian. Building on what we learned from the verse-level setting, we adopt a higher granularity and increase the size of the corpus. First, we switch from verses to arias in order to have longer and more expressive texts. Second, we construct a new corpus with* $40k$ *arias (*$\sim 90k$ *verses). This new dataset contains silver data, annotated by self-learning on the basis of an ensemble of binary classifiers.*

We then experiment with more sophisticated representations, by learning an embedding space and using it to train new models for the identification of emotions at the aria level, obtaining a significant performance boost.

1. Introduction

Arias are used by authors to express the emotional state of the singing character within an opera play. In 17th- and 18th-century Italian operas, characters brought on stage passions ("*affetti*") induced in their souls by the succession of events in the drama. Musicological studies use these affects as one of the interpretative keys of the work

* Dept. of Language Science and Technology - Universität des Saarlandes Campus, 66123 Saarbrücken, Germany. E-mail: shzh00003@stud.uni-saarland.de
** Dept. of Interpreting and Translation - Corso della Repubblica 136, 47121 Forlì, Italy. E-mail: francesco.fernicola2@unibo.it
† Dept. of Interpreting and Translation - Corso della Repubblica 136, 47121 Forlì, Italy. E-mail: federico.garcea2@unibo.it
‡ Dept. of Classical Philology and Italian Studies - Via Zamboni 32, 40126 Bologna, Italy. E-mail: paolo.bonora@unibo.it
§ Dept. of Interpreting and Translation - Corso della Repubblica 136, 47121 Forlì, Italy. E-mail: a.barron@unibo.it – Corresponding author

as a whole (Zoppelli 2001; McClary 2012). In AriEmozione we aim at creating models for the automatic identification of emotions in opera arias. Such models represent a valuable tool for the systematic study and organisation of the vast *repertoire* of arias and characters of this period for musicologists and the lay public alike.

Since an aria may express more than one emotion, we depart at a lower-granularity level: the verse. We first engineer models to identify the emotion of a single verse and then we point higher to identify the emotion(s) expressed by full arias. In the verse-level experiments, the small amount of data available makes it difficult to rely on dense representations or sophisticated models. A 2-layer feed-forward neural network fed with TF-IDF-weighted character 3-grams achieved the best F_1-macro of 0.47. This relatively low performance reflects the difficulty of the task at hand, partially caused by the small amount of supervised data available. In order to overcome these limitations, we produce a significantly larger annotated dataset by means of self-learning. Even if the new data is noisy, the larger amount of supervised instances allows for the application of dense representations and a convolutional neural network, resulting in a performance boost of 0.20 points absolute, passing from an F_1-macro of 0.47 to 0.67.

Our contributions can be summarised as follows.

1. We produced AriEmozione 1.0—a manually-annotated corpus with emotion labels at the verse level including $2.5k$ instances.
2. We produced AriEmozione 2.0—a self-learning-annotated corpus with emotion labels both at the aria and at the verse level including $40k$ arias ($90k$ verses).
3. We produced a FastText embedding space of 17th- and 18th-century Italian operas.
4. We explored supervised models for the identification of emotions in opera at the verse level.
5. We explored supervised models for the identification of emotions in opera at the aria level.

We release both corpora and the embedding space to the research community as well as the implementation of the different models both at verse and aria level.

The rest of the paper is articulated as follows. Section 2 offers some background about both opera and emotions. Section 3 reviews related work on both sentiment analysis and emotion identification. Section 4 presents the work intended to identify emotions at the verse level, including the construction of the dataset and multiple experiments. Section 5 describes the approach after switching to the aria level, including the automatic production of the corpus and the application of deep-learning models. Section 6 closes our contribution by drawing conclusions and identifying interesting research avenues for the future work.

2. Background

In music, *aria* refers to a piece of lyrics within the context of a full opera. An aria usually consists of more than one verse that composes the singer's participation in the dialogue. In general, opera lyrics are highly structured (Burden 1998); usually split in recitative parts where the action occurs, and arias where characters, normally singing a solo, express their feelings and motivations. Arias have a strophic structure, during

the 17th century often dyadic with repetition of the first part (*da capo*). Some times, the first part gives a metaphoric representation of the *affetto* with the second explicating its consequences on the singing character. Each aria is conceived as a whole and as a closed piece, hence being potentially interchangeable between different plays as its function is to convey one or more distinctive *affetti* to the public.

Our research builds on top of CORAGO, the Repertoire and archive of Italian opera librettos.[1] CORAGO constitutes the first implementation of the RADAMES prototype (*Repertoriazione e Archiviazione di Documenti Attinenti al Melodramma E allo Spettacolo*; Repertorisation and Archiving of Documents Related to Melodrama And Entertainment) (Pompilio et al. 2005). All texts are written in 18th-century Italian and articulated in verses and *stanzas* —groups of verses and the way the metric and rhyme structure of a lyric is articulated. The most represented authors in our corpus are two of the most successful librettists of the 18th century: Apostolo Zeno and Pietro Metastasio whose 26 librettos were put in music in more than one thousand operas during the 19th century. Whereas Zeno composed mostly operas on historical and mythological themes, Metastasio is considered the most important writer of *opera seria*.

Most arias in the collection contain between two and three verses. We derive the emotion classification scheme from various previous works. We consider René Descartes' "Les passions de l'âme" (1649) as the reference for the coeval literature for emotions representations and their social expressions and meanings (Garavaglia 2018). We then selected a contemporary model that could be aligned in order to represent the taxonomy of Descartes while being based on the lexical representation of emotion in lyrics. We also consider Shaver et al. (1987)'s prototype approach based on the analysis of the lexicon of emotions. Through this review, carried out together with expert musicologists with extensive experience in the analysis of operas during the studied period, we settled on Parrott (2001)'s hierarchical classification and end up with six primary emotions, which turn into our six classes:

`Amore (love):` a focused sense of belonging, care and attraction toward someone; incl. affection, lust, and longing.

`Gioia (joy):` a sense of fulfillment and positiveness; incl. cheerfulness, zest, contentment, pride, optimism, enthrallment, and relief.

`Ammirazione (admiration):` admiration or adoration of someone's talent, skill, or other physical or mental qualities; incl. esteem, respect, and approval.

`Rabbia (anger):` a state of repulsion and frustration due to something or someone interfering with one's aims; incl. irritability, exasperation, rage, disgust, envy, and torment.

`Tristezza (sadness):` a state following an unwanted outcome, a loss or a delusion; incl. suffering, disappointment, shame, and neglect.

`Paura (fear):` a state induced by the interpretation of oncoming events as potentially dangerous or threatening; incl. horror and nervousness.

1 `http://corago.unibo.it`.

During the annotation process, We included an extra class: **nessuna (none)**, which applies mostly to verses containing only non-actionable words; the few instances of this class have been neglected in all experiments (cf. Section 4).

3. Related Work

Sentiment analysis, also known as opinion mining, aims at determining the polarity of a text by investigating text features (Liu and Özsu 2009). The decision is often binary —positive vs negative— or ternary, with the addition of an intermediate neutral class. Research on sentiment analysis is vast and we refer the interested reader to Birjali et al. (2021) for a thorough overview. Starting from numerous advances on sentiment analysis, researchers attempted to move towards a finer-degree problem in the more complicated task of multi-class emotion identification. Research has been conducted on various types of text, ranging from social media contents with tweets (Roberts et al. 2012) or Facebook posts (Pool and Nissim 2016) to lyrics (Hu, Chen, and Yang 2009), news (Kirange and Deshmukh 2012), and children's fairy tales (Alm, Roth, and Sproat 2005).

Datasets exist for the analysis and identification of emotions in Italian; most of them focused on social media content. MultiEmotion-It is a corpus with comments from YouTube and Facebook posts responding to music videos and advertisement (Sprugnoli 2020). These comments are annotated according to four aspects: *relatedness*, *polarity*, *emotions* and *sarcasm*. In the specific case of emotions, Plutchik (1980)'s model is used, resulting in classes *joy*, *sadness*, *fear*, *anger*, *trust*, *disgust*, *surprise*, and *anticipation*. MultiEmotions-It (Sprugnoli 2020) and AriEmozione 1.0 (Fernicola et al. 2020) were both released in 2020, representing two of the first manually-annotated corpora for the identification of emotions in Italian. FEEL-IT is a corpus with $2k$ Italian tweets annotated with one label out of *anger*, *fear*, *joy*, and *sadness* (Bianchi, Nozza, and Hovy 2021). The justification in the selection of these labels relies on their "frequent occurrence in text" (Bianchi, Nozza, and Hovy 2021, p. 76)

Our contribution in terms of corpora release go beyond both MultiEmotion-It and FEEL-IT. Similarly to the former, we base our label selection on a formal classification of emotions supported on a psychological and philosophical theory. Similarly to the latter, instead, we narrow such selection by an expert analysis of the emotions that are more present in the analysed genre. Both MultiEmotion-It and FEEL-IT contain annotations at the document level (be it a tweet or a comment). As Strapparava et al. (2012), who released a corpus of popular music in English, we go at the sub-document level and annotate single verses and arias.

Regarding models, some approaches involve rule-based systems. For example, Asghar et al. (2017) proposed a rule-based framework for sentence-level emotion identification of user reviews using an emotion lexicon. Researchers created a mixed-mode classifier that takes into account not only emotion words, but also emoticons and slang and compared the performance of a mixed-mode classifier with another classifier that is created using only emotion words as resources. Both models are tested on a corpus of news texts and the mixed-mode classifier was the one which performed better.

Some researchers opted for a hybrid approach, making use of both supervised and unsupervised methods to achieve a higher accuracy. This is the case of Gievska et al. (2014), who designed an emotion detection approach to deal with the ISEAR

dataset.[2] This study considered seven emotions derived from Ekman's six emotional categories (Ekman 1994): *anger*, *fear*, *sadness*, *disgust*, *joy*, *surprise*, and an additional *neutral* category in order to reduce the effect of misclassified data. They experimented with a lexical-based method alone, the machine learning method alone, and the blended method using both of the previous models. The lexical-based method is developed using a variety of language resources such as WordNetAffect (Strapparava and Valitutti 2004), AFINN (Nielsen 2011), H4Lvd,[3] and the NRC word-emotion association lexicon (Mohammad and Turney 2013). An SVM obtained a significant precision advantage and the hybrid method performed the best.

The existing supervised text emotion analysis research can also be categorised into three general classes: single-label learning (SLL), multi-label learning (MLL), and label distribution learning (LDL) (Zhao and Ma 2019). In SLL, the emotion of a text is represented by a single emotion such as *joyful* or *sad*. In MLL a text is assumed to transmit more than one emotion and can be therefore assigned more than one label. For example, Ye et al. (2012) constructed a reader emotion corpus collecting news articles from the Sina news media. Each news article corresponds to one to three emotion labels. Various feature selection strategies such as document frequency and chi-square statistic are tested with different multi-class classification models. LDL goes a step further and assigns not only a set of emotion labels but also the corresponding emotion intensity. Zhou et al. (2016) proposed a distribution learning approach capable of identifying emotions with their respective intensities of the given sentence. Eight emotion labels are established based on Plutchik's wheel of emotions (Plutchik 1980). Each sentence may express one or more emotions and the sum of emotion intensities for each sentence is normalised to one. This study also captures the relations among the eight emotions of Plutchik's wheel (Plutchik 1980) and incorporate these relations into the learning algorithm in order to enhance the accuracy.

4. Emotion Identification at the Verse level

This section presents our efforts to identify emotion in opera lyrics at the verse level. We cover the creation of the AriEmozione 1.0 corpus as well as the exploration of diverse models and representations.

4.1 The AriEmozione 1.0 Corpus

This corpus is a subset of the materials from project CORAGO (cf. Section 2). We selected a set of 678 operas composed between 1655 and 1765, considering only the lyrical text in the arias (and neglecting, for instance, recitatives). For the annotation, we split all opera arias into verses, resulting in 2,473 instances. At this stage, we opted for verses because we observed that the snippets hardly express more than one emotion at this level of granularity. Two native speakers of Italian annotated all verses independently following the instructions displayed in Figure 1. They were asked to include (*i*) the emotion transmitted by the verse, (*ii*) an optional secondary label (in case they perceived a second emotion), and (*iii*) their level of confidence: total confidence, partial confidence, or doubtful. Cohen's kappa inter-annotator agreement (Fleiss, Cohen, and Everitt 1969) on the primary emotion was of 0.323, which is considered as a fair agreement —this

2 https://github.com/sinmaniphel/py_isear_dataset
3 http://www.wjh.harvard.edu/~inquirer/Home.html.

First of all, thank you for helping with this work. We are a group of researchers from the D. of Classical Philology and Italian Studies and the D. of Interpreting and Translation, both at UniBO. Your work will help us to produce artificial intelligence models to analyse the lyrics in music.
At this stage we are focused on opera. You will annotate arie in Italian from diverse periods, looking for the emotions that they express. Your work consists of identifying the emotion expressed in each of the verses composing an aria. You can choose among six emotions (or none of them), which are defined next: [...]

Each row is divided in six columns:
id A unique id, tied to the verse. Do not modify it.
verse A verse, inside of an aria. This is the text that you are going to analyse.
emotion Here you can select the expressed emotion (or none of them)
emotion sec. This is available to choose a secondary emotion, in case it is really difficult to choose just one
confidence Not being 100% sure is ok. If that is the case, please let us know by choosing the right confidence level (default: "I am sure").
comments Feel free to tell us something about this instance, if you feel like.

Figure 1
Instructions given to the annotators of the emotions at the verse level in the AriEmozione 1.0 corpus.

Table 1
AriEmozione 1.0 corpus statistics per partition and class.

	amore	gioia	ammirazione	rabbia	tristezza	paura	total
train	289	274	289	414	503	166	**1,973**
dev	36	31	23	84	61	12	**250**
test	37	39	30	64	54	15	**250**
all	362	344	342	562	618	193	**2,473**

value results from the perfect matching between the two annotators in 44% of the instances. When considering the secondary emotion as well, the two annotators were in agreement on 68% of the instances. These numbers reflect the complexity of the task. The same annotators gathered together to discuss and consolidate all dubious instances and produce a consolidated label.

Table 1 shows statistics on the number of instances per class for each corpus partition. The most represented emotions are *tristezza* (sadness) followed by *rabbia* (anger): 25% and 23% of the instances, respectively. The least represented emotion is *paura*, which negatively impacted its prediction results; cf. Section 4.3). A total of 52 verses did not express any emotion and were neglected from the experiments. The average length of the verses is of 72.5 ± 31.6 characters and the corpus contains $34,608$ tokens and $4,458$ types.[4] Appendix A shows the distribution of these classes across time periods. Table 2 shows examples of verses in the corpus, including one for each of the six emotions.

4.2 Models and Representations at the Verse Level

The nature of the corpus —a small amount of short verses written in 18th-century Italian— led us to select a humble set of models and representation alternatives. The

4 The AriEmozione 1.0 corpus is available for download at `https://zenodo.org/record/4022318`.

Table 2
Instances from the AriEmozione 1.0 corpus, including their `English translation`, class, and unique identifier. We include free (unofficial) translations for clarity.

verse	class (id)
Non ho più lagrime; non ho più voce; non posso piangere; non so parlar `I have no more tears; I have no more voice; I cannot cry; I don't know how to speak`	Tristezza (ZAP1593570_03)
Barbaro! Oh dio mi vedi divisa dal mio ben; barbaro, e non concedi ch'io ne dimandi almen `You Barbarian! Oh Lord, you see me separated from my very precious; barbarian, you won't even allow me a question`	Rabbia (ZAP1596431_00)
Guardami e tutto obblio e a vendicarti io volo; di quello sguardo solo io mi ricorderò `Look at me, all else is forgotten and I haste to avenge you; only I shall remember that gaze`	Amore (ZAP1593766_01)
Su la pendice alpina dura la quercia antica e la stagion nemica per lei fatal non è; `Up on the slope of the mountain the ancient oak tree still lives on, and the adverse season poses no fatal threat`	Ammirazione (ZAP1594229_00)
In questa selva oscura entrai poc'anzi ardito; or nel cammin smarrito timido errando io vo `I entered this dark forest not too long ago, boldly; having now lost the path I wander around, shyly`	Paura (ZAP1596807_00)
Vede alfin l'amate sponde, vede il porto, e conforto prende allor di riposar `Finally, the beloved shores, the harbor, are all in sight and with them come solace and sleep`	Gioia (ZAP1599979_01)

baseline is a k–Nearest Neighbors algorithm (kNN), considered due to its simplicity and success in small classification tasks (Zhang and Zhou 2007). We also experiment with multi-class SVMs, logistic regression, and neural networks. Regarding the latter, we experiment with a number of architectures with two and three hidden layers. Finally, we experiment with a FastText classifier (Joulin et al. 2017). Table 3 summarises the configurations explored.[5]

As for the text representations, we consider TF–IDF vectors of both character 3-grams and word 1-grams (no higher values of n are considered due to the size of the corpus). For pre-processing, we employ the spaCy Italian tokenizer[6] and casefold the texts. We also explore with dense representations, derived from the TF–IDF vectors, by means of both LDA (Hoffman, Bach, and Blei 2010) and LSA (Halko, Martinsson, and Tropp 2011). In both cases, we target reductions to 16, 32, and 64 dimensions. As embeddings, we adopted the pre-trained 300-dimensional Italian vectors of FastText (Joulin et al. 2017), and tried with character 3-grams and words.

5 The code is available at `https://github.com/TinfFoil/AriEmozione`. We used Sklearn for the kNN, SVM, and logistic regression models; Keras for the neural networks, and the Facebook-provided library for FastText (cf. `https://scikit-learn.org`, `https://keras.io` and `https://github.com/facebookresearch/fastText`).

6 `https://spacy.io/models/it`

Table 3
Experimental settings for the emotion identification models at the verse level.

Model	Settings
k-NN	L2-Norm exploring with $k \in [1, \dots 9]$.
SVM	RBF exploring with $c \in [1, 10, 100, 1000]$ and $\gamma \in [1e-3, 1e-4]$.
Log Reg	Multinomial Logistic Regression with Newton-CG solver.
NN	2 (3) hidden layers with size $\in [32, 64, 96, 128, 256]$ ($\in [8, 16, 32, 64, 96]$); 20% dropout; ReLu for input/hidden layers; softmax for output layer; categorical cross-entropy loss function; Adam optimiser; epochs $\in [1, \dots 15]$
FastText	300d embeddings with or without pre-training; learning rate $\in [0.3, 0.6, 1]$; epochs $\in [1, 3, 5, 10, \dots, 100]$

4.3 Experiments at the Verse Level

We conducted several experiments to find the best combination of parameters and representations. Given the amount of instances available, we merged the training and development partitions and performed 10-fold cross validation. As standard, the test partition was left aside and only one prediction was carried out on it, after identifying the best configurations. We evaluate our models on the basis of accuracy and weighted macro-averaged F_1 to account for the class imbalance. Table 4 shows the results obtained with some interesting configurations and representations both for the cross-validation and on the test set.[7] TF-IDF character and word n-grams, LSA, and LDA were tested with all models except for FastText, on which we test with and without pre-trained embeddings. Notice that we are not interested in combining features, but in observing their performance in isolation.

The most promising representation on cross-validation is the simple character 3-grams, with which we obtained the best results across all models; although it also features the highest variability across folds. Among all 3-gram derived representations, LDA consistently obtained the worst results across all models. Still, it is more stable across folds than the sparse 3-gram representation. LSA performs significantly better than LDA and is always close to the TF-IDF words representation, most notably using the k-NN model. As for FastText, with the same epoch number and learning rate, the character 3-gram vectors always achieved much higher accuracy than the word vectors. Similar patterns are observed when predicting on the unseen test set. The character 3-grams in general hold the best performance, while the 3-gram LDA tends to remain the worst in spite of the model used. This behavior does not hold in all cases. For instance, the logistic regression model achieves F_1 =0.44 on cross-validation, but drops to 0.42 on test. This might be the result of over-fitting.

Table 5 shows the confusion matrix for the best model on test. All models tend to mix `rabbia` and `tristezza`. These two emotions get confused with each other on an average of 18% of the cases. The classifiers tend to confuse `ammirazione` for `gioia` as well, which is understandable given their semantic closeness.

A number of factors contribute to the relatively low performance. First, the verses tend to be very short, causing the identification of emotions difficult. The ancient nature

7 The full batch of results is available at `https://docs.google.com/spreadsheets/d/1Ztjry2mJs6ufCZM1O5CQRyZ8pA5YDnToN0h0NGX1nW0/edit?usp=sharing`

Table 4
F_1 and accuracy for the emotion identification at the verse level on cross-validation and held-out test for some of the model and representation combinations.

model	representation	10-fold CV		test	
		F_1	Acc	F_1	Acc
k–NN	char 3-grams	0.38	38.51	0.35	35.15
	words	0.36	36.08	0.35	34.73
	LSA char	0.36	35.26	0.33	32.64
	LDA char	0.30	29.97	0.31	30.54
SVM–RBF	char 3-grams	0.44	43.70	0.43	43.00
	words	0.42	42.00	0.44	44.00
	LSA char	0.39	39.00	0.40	40.00
	LDA char	0.28	28.00	0.30	30.00
Log reg	char 3-grams	0.44	**45.57**	0.42	43.10
	words	0.41	43.20	0.41	43.10
	LSA char	0.36	36.30	0.34	34.73
	LDA char	0.28	30.63	0.29	30.96
2-layers NN	char 3-grams	0.42	43.61	**0.47**	**46.86**
	words	0.42	42.91	0.43	43.10
	LSA char	0.35	35.63	0.36	37.24
	LDA char	0.27	29.56	0.27	31.80
3-layers NN	char 3-grams	**0.49**	41.86	0.40	41.84
	words	0.47	42.60	0.40	41.84
	LSA char	0.44	41.86	0.41	41.84
	LDA char	0.26	31.41	0.30	31.80
FastText	char 3-grams	0.43	45.00	0.41	42.37
	pre-trained char 3-grams	0.43	47.00	0.41	41.00
	words	0.42	42.56	0.39	44.07
	pre-trained words	0.38	41.00	0.40	42.00

of the language causes pre-trained vectors, such as FastText's, to have a low word coverage. Last, but not least, the number of instances available for training is fairly small. We address these issues in the next section, where we also jump from the verse- to the aria-level emotion identification.

5. Emotion Identification at the Aria Level

We address the issues observed while experimenting at the verse level in different ways. Among them, we expand the size of the supervised data and shift to a higher granularity: the aria. This shift is motivated by the complex structure of the texts, where the lexicon and phrases used to express an *affetto* often span beyond a single verse and even a whole *stanza*. The creation of more annotated instances also opens the door to produce more sophisticated representations; e.g., in-domain embedding spaces. We open the discussion with the creation of the AriEmozione 2.0 corpus and continue exploring with diverse models and representations.

Table 5
Confusion matrix for the 2-layers neural network with TF-IDF character 3-grams on the verse-level prediction task.

	ammirazione	amore	gioia	paura	rabbia	tristezza
ammirazione	**0.37**	0.03	0.18	0.07	0.11	0.06
amore	0.03	**0.43**	0.13	0.00	0.09	0.17
gioia	0.27	0.16	**0.31**	0.20	0.09	0.07
paura	0.10	0.03	0.00	**0.40**	0.02	0.07
rabbia	0.20	0.14	0.03	0.13	**0.64**	0.17
tristezza	0.17	0.14	0.13	0.07	0.19	**0.48**

5.1 The AriEmozione 2.0 Corpus

As observed in Section 4.1, the CORAGO-1700 corpus is composed of Italian operas and lacks any supervision; the annotated AriEmozione 1.0 represents just a tiny subset. We produced corpus AriEmozione 2.0 by performing a self-learning annotation process (Jurkiewicz et al. 2020) on another subset of CORAGO-1700. The first step to label this new corpus would be to automatically identify the class of the new verses with some of our existing models and iteratively add fresh instances to the training set. Nevertheless, even the best-performing multi-class model trained on AriEmozione 1.0 achieves an F_1-measure lower than 0.47 (cf. the 2-layers NN with TF-IDF character 3-grams in Section 4.3). Hence, we adopt a one-versus-all approach (OVA) (Aly 2005). OVA decomposes the k-class classification into k binary classification problems to focus on one emotion class at a time. The instance labels are determined by the class that obtained the maximum classification score. We run parallel processes considering all six classes to iteratively produce the annotations, which end up as the silver data in the AriEmozione 2.0 corpus. Appendix B describes the process in detail.[8]

In order to assess the quality of this pre-selection, we evaluated three different binary models for each class; each model differs with regards to the training material they have access to: (*i*) AE1_{tr} is trained on the manually-annotated instances from AriEmozione 1.0 , (*ii*) $\text{AE1}_{tr} \cup \text{RAW}_{pos}$ considers all training material from AriEmozione 1.0 plus only the instances that have been assigned the class of the corresponding emotion, and (*iii*) $\text{AE1}_{tr} \cup \text{RAW}_{all}$ considers all training material from AriEmozione 1.0 plus all new instances, regardless of the class they were labelled with.

Table 6 shows the results on the binary settings over the development set of AriEmozione 1.0 . In terms of precision of the positive (emotion) class, $\text{AE1}_{tr} \cup \text{RAW}_{pos}$ performs consistently the best. Except for emotions *ammirazione* and *tristezza*, $\text{AE1}_{tr} \cup \text{RAW}_{all}$ achieves both the highest accuracy and F_1 scores. However, as precision of the emotion class is the most important metric, we adopt the strategy where only new instances predicted as belonging to the corresponding emotion are integrated.

8 The AriEmozione 2.0 corpus is available at `https://zenodo.org/record/7097913`.

Table 6
Per-class binary evaluation of the models considering different partitions of the self-training pre-labelled instances from the CORAGO corpus. We include the precision of the positive (emotion) class. AE1$_{tr}$=binarised AriEmozione 1.0 training set; RAW$_{all}$=all new instances, regardless of their assigned label; RAW$_{pos}$=new instances labeled as (emotion) positive class.

Emotion	Training material	Acc	F_1	Precision
ammirazione	AE1$_{tr}$	0.851	0.787	0.019
	AE1$_{tr}$∪RAW$_{all}$	0.881	0.872	0.455
	AE1$_{tr}$∪RAW$_{pos}$	**0.882**	**0.878**	**0.530**
amore	AE1$_{tr}$	0.861	0.800	0.024
	AE1$_{tr}$∪RAW$_{all}$	**0.886**	**0.889**	0.671
	AE1$_{tr}$∪RAW$_{pos}$	0.857	0.867	**0.696**
gioia	AE1$_{tr}$	0.853	0.808	0.111
	AE1$_{tr}$∪RAW$_{all}$	**0.866**	**0.856**	0.407
	AE1$_{tr}$∪RAW$_{pos}$	0.847	0.848	**0.504**
paura	AE1$_{tr}$	0.921	0.899	0.111
	AE1$_{tr}$∪RAW$_{all}$	**0.968**	**0.970**	0.917
	AE1$_{tr}$∪RAW$_{pos}$	0.952	0.956	**0.924**
rabbia	AE1$_{tr}$	0.789	0.749	0.241
	AE1$_{tr}$∪RAW$_{all}$	**0.812**	**0.810**	0.586
	AE1$_{tr}$∪RAW$_{pos}$	0.802	0.802	**0.589**
tristezza	AE1$_{tr}$	0.746	0.724	0.296
	AE1$_{tr}$∪RAW$_{all}$	**0.782**	**0.779**	0.516
	AE1$_{tr}$∪RAW$_{pos}$	0.747	0.751	**0.611**

Table 7
Class statistics at the verse level for AriEmozione 2.0 .

	amore	gioia	ammirazione	rabbia	tristezza	paura	total
freq.	13,363	13,226	13,915	17,587	25,499	6,359	**89,949**

To produce the actual annotations that will turn into the AriEmozione 2.0 corpus, we train six new binary neural networks with softmax output layers, each responsible for one one emotion. Each network is trained on the training plus the development sets from AriEmozione 1.0 plus the pre-selected instances belonging to the corresponding emotion class from the previous process. The consolidated —and final— label for each of the new raw instances is the one with the highest score among the six models. This approach to consolidate the labels is inspired by multi-class settings such as multi-class SVMs, where the decision is based on a winner-takes-all strategy (Duan and Keerthi 2005; Crammer and Singer 2001). Table 7 shows the class distribution of the AriEmozione 2.0 corpus. It contains $90k$ verses, a significantly larger amount than its predecessor. Appendix C shows the impact of these new materials on the verse-level identification task.

One of the drawbacks for the models is that the verses tend to be too short. In the rest of the paper, we shift the granularity of our instances from the verse to the aria level. Since an aria is in general composed by more than one verse, and such verses could have

Table 8
Class distribution at the aria level for AriEmozione **1.0 (gold)** —manually annotated— and AriEmozione **2.0 (silver)** —automatically annotated.

ammirazione	amore	gioia	paura	rabbia	tristezza	**1.0 (gold)**	**2.0 (silver)**
One-class instances							
■						109	2,172
	■					122	2,084
		■				107	2,099
			■			57	757
				■		194	3,224
					■	185	5,399
Two-class instances							
■	■					9	1,317
■		■				30	1,769
■			■			9	546
■				■		12	1,878
■					■	14	2,245
	■	■				13	1,407
	■		■			5	404
	■			■		7	1,456
	■				■	24	2,689
		■	■			5	642
		■		■		8	1,381
		■			■	11	2,189
			■	■		5	579
			■		■	22	1,024
				■	■	47	3,119
Overall							
■	■	■	■	■	■	995	38,380

been identified as expressing different emotions, we establish that an aria can belong to up to two emotions. The emotion of an aria is determined by the most frequent emotion label among its verses. In case of draw, the aria keeps the top-two classes.[9] In order to avoid confusion, in the following we refer to the arias derived from AriEmozione 1.0 as "gold instances", whereas those from AriEmozione 2.0 are "silver instances". Table 8 shows the statistics of the resulting dataset.

9 If the draw involves more than two emotions, the instance is considered too noisy and it is discarded. As a result, six arias from AriEmozione 1.0 and 1,623 arias from AriEmozione 2.0 get discarded.

Table 9
Results of the emotion identification task at the aria level for a CNN with different learning rates and number of epochs. The text representation is 300-dimensional pre-trained embeddings on character 3-grams.

learning rate	epochs	accuracy	F_1
0.0001	10	0.491	0.736
0.0001	15	0.628	0.785
0.0001	20	0.614	0.789
0.001	10	0.635	0.795
0.001	15	**0.706**	**0.829**
0.001	20	0.652	0.812

5.2 Models and Representations at the Aria Level

One of the obstacles when dealing with this kind of material is its language: 18th-century Italian. This makes ineffective representing the instances with out-of-the-box pre-trained embeddings, which are built on modern text. To address this issue, we build 300-dimensional embeddings using FastText (Bojanowski et al. 2017) using both AriEmozione 1.0 and AriEmozione 2.0 as unsupervised training material. We produced character 3-gram embeddings by training during 5 epochs with a learning rate of 0.05.

As for the classification models, we opt for a multi-label setting to predict up to two classes per instance. We use a CNN with one convolutional layer (ReLU activation functions and a stride of 3), two hidden layers and the output layer. Both hidden layers have 2,500 neurons, dropout of 0.1 and sigmoid activation functions. The output layer has a six-units sigmoid function. We use binary cross-entropy and the Adam optimizer. The classification threshold is set at 0.5.[10]

5.3 Experiments at the Aria Level

The CNNs are trained on all arias in AriEmozione 2.0 (silver instances) and tested on all arias in AriEmozione 1.0 (gold instances). Table 9 shows the results after training during different epochs and with two learning rates. The best performance is obtained when training for 15 epochs with a learning rate of 0.001: $F_1 = 0.829$. Even if this score is not directly comparable to the numbers in Table 4 (different data partitions, different granularity), the allocation of more training data and the aria granularity clearly allow for much better figures.

Table 10 shows the confusion matrices of such model. Having a multi-label setting, we opt for dissecting into six matrices: one emotion against the rest. Instances of class *tristezza* are identified the best, with a precision of 0.921, whereas instances of *paura* are the most difficult, with a precision of 0.825. These outcomes can be attributed to the imbalanced distribution of instances with double labels in gold instances and silver instances. Table 8 shows that about 55% of the silver instances have two labels, while

10 The implementation code is available at `https://github.com/TinfFoil/AriEmozione-2.0`. We used Sklearn for label encoding, Keras for the neural networks, and the Facebook-provided library for FastText (cf. `https://scikit-learn.org`, `https://keras.io` and `https://github.com/facebookresearch/fastText`).

Table 10
Normalised confusion matrices for the emotion identification task at the aria level zoomed into each of the six classes against the rest. Absolute values shown in parenthesis.

	rest	**ammirazione**
rest	0.889 (722)	0.111 (90)
ammirazione	0.104 (19)	0.896 (164)

	rest	**amore**
rest	0.948 (773)	0.052 (42)
amore	0.100 (18)	0.900 (162)

	rest	**gioia**
rest	0.965 (792)	0.035 (29)
gioia	0.161 (28)	0.839 (146)

	rest	**paura**
rest	0.983 (877)	0.017 (15)
paura	0.175 (18)	0.825 (85)

	rest	**rabbia**
rest	0.939 (678)	0.061 (44)
rabbia	0.128 (35)	0.872 (238)

	rest	**tristezza**
rest	0.893 (618)	0.107 (74)
tristezza	0.079 (24)	0.921 (279)

only 17% of the gold instances do. Many single-label instances in AriEmozione 1.0 are assigned two (or even three) labels. The best-performing model assigned three labels to 14 arias and two labels to 345 in the test set, whereas in reality only 222 arias have two labels associated.

Overall, the performance is good considering the difficulty of the task. However, there is room for improvement: the model shows robustness in the identification of each singleton emotion, but it struggles with multi-label classification.

6. Conclusions

We addressed the novel problem of identifying the emotions expressed by opera aria lyrics. This is an interesting problem because it opens the door to the creation of search engines and to the assisted organisation and curation of repertoires —both based on emotion. It is challenging because there is a lack of supervised (and unsupervised) data in the domain, and its language —17th- and 18th-century Italian— makes the use of modern semantic representations non straightforward.

We address the problem at two granularity levels: the verse and the aria. For the former, we annotated a small collection of verses with six emotions and performed numerous experiments with different models (e.g., support vector machines, logistic regression, and neural networks) and representations (e.g., character and word n-grams and word embeddings). Our results showed that neither the amount of supervised data nor the representations were enough. We then applied a self-learning approach to produce silver data to train on, produced an embedding representation out of a large-collection of non-supervised operas, and shifted to the aria granularity level, within a multi-label setting in which each instance could express up to two emotions. These efforts enabled us to try convolutional neural networks on better representations, which resulted in a large performance boost, bringing the approach closer to be applied in a practical setting.

The work on emotion identification in opera (and other kind of musical arts) can be further refined. For instance, rather than a multi-label setting, the emotion of an aria could be judged on the basis of a distribution, which considers that each item might have non-zero intensities for every single emotion (Zhao and Ma 2019). Another

interesting avenue would be considering multi-modal aspects. That is, not only the written verses but also music sheets. The parallel corpus from Strapparava et al. (2012), which includes annotations on the notes and lyrics of popular music in English, can be leveraged to investigate the cooperation between textual features and musical features for emotion identification (Mihalcea and Strapparava 2012). In the case of operas, even scene representations could be taken into consideration in the decision process. Videos could play that role for popular music.

Acknowledgments

This research was carried out in the framework of CRICC: *Centro di Ricerca per l'interazione con le Industrie Culturali e Creative dell'Università di Bologna*; a POR-FESR 2014-2020 Regione Emilia-Romagna project (`https://site.unibo.it/cricc`).
We thank Ilaria Gozzi and Marco Schillaci, students at Università di Bologna, for their support in the manual annotation of the AriEmozione 1.0 corpus.

References

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Aly, Mohamed. 2005. Survey on multiclass classification methods. *Technical Report*, 19:1–9.

Asghar, Muhammad Zubair, Aurangzeb Khan, Afsana Bibi, Fazal Masud Kundi, and Hussain Ahmad. 2017. Sentence-level emotion detection framework using rule-based classification. *Cognitive Computation*, 9(6):868–894.

Bianchi, Federico, Debora Nozza, and Dirk Hovy. 2021. FEEL-IT: Emotion and sentiment classification for the Italian language. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 76–83, Online, April. Association for Computational Linguistics.

Birjali, Marouane, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107–134.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Burden, Michael. 1998. The new grove dictionary of opera, ed. Stanley Sadie. *Early Music*, 26(4):669–670.

Crammer, Koby and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, December.

Duan, Kai-Bo and S. Sathiya Keerthi. 2005. Which is the best multiclass svm method? an empirical study. In Nikunj C. Oza, Robi Polikar, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, pages 278–285, Seaside, CA, USA, June. Springer Berlin Heidelberg.

Ekman, Paul. 1994. All emotions are basic. *The nature of emotion: Fundamental questions*, pages 15–19. Oxford University Press.

Fernicola, Francesco, Shibingfeng Zhang, Federico Garcea, Paolo Bonora, and Alberto Barrón-Cedeño. 2020. Ariemozione: Identifying emotions in opera verses. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Bologna, Italy [online], March, 2021.

Fleiss, Joseph L., Jacob Cohen, and B.S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327.

Garavaglia, Andrea. 2018. Funzioni espressive dell'aria a metà seicento secondo il "Giasone" di Cicognini e Cavalli. *Il Saggiatore Musicale*, Anno XXV(1):5–31.

Gievska, Sonja, Kiril Koroveshovski, and Tatjana Chavdarova. 2014. A hybrid approach for emotion detection in support of affective interaction. In *2014 IEEE International Conference on Data Mining Workshop*, pages 352–359, Shenzhen, China, December. IEEE.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Hoffman, Matthew, Francis Bach, and David Blei. 2010. Online learning for Latent Dirichlet Allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Hu, Yajie, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *International Society for Music Information Retrieval*, pages 123–128, Kobe, Japan, October.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Jurkiewicz, Dawid, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online), December. International Committee for Computational Linguistics.

Kirange, Dnyaneshwar and Ratnadeep Deshmukh. 2012. Emotion classification of news headlines using SVM. *Asian Journal of Computer Science and Information Technology*, 5(2):104–106.

Liu, Ling and M. Tamer Özsu. 2009. *Encyclopedia of database systems*, volume 6. Springer New York, NY, USA:.

McClary, Susan. 2012. *Desire and Pleasure in Seventeenth-Century Music*. University of California Press, Berkeley, CA, 1 edition.

Mihalcea, Rada and Carlo Strapparava. 2012. Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea, July. Association for Computational Linguistics.

Mohammad, Saif M. and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Nielsen, Finn Årup. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98, Heraklion, Crete, May.

Parrott, W. Gerrod. 2001. *Emotions in Social Psychology: Essential Readings*. Psychology Press.

Plutchik, Robert. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Pompilio, Angelo, Lorenzo Bianconi, Fabio Regazzi, and Paolo Bonora. 2005. RADAMES: A new management approach to opera: Repertory, archives and related documents. In *Proceedings - First International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, Florence, Italy, November-December. Institute of Electrical and Electronics Engineers.

Pool, Chris and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39.

Roberts, Kirk, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, volume 12, pages 3806–3813, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shaver, Philip, Judith Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061–1086.

Sprugnoli, Rachele. 2020. Multiemotions-it: a new dataset for opinion polarity and emotion analysis for italian. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769 of *CEUR Workshop Proceedings*, Bologna, Italy, March 2021. CEUR-WS.org.

Strapparava, Carlo, Rada Mihalcea, and Alberto Battocchi. 2012. A parallel corpus of music and lyrics annotated with emotions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, volume 12, pages 2343–2346, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Ye, Lu, Rui-Feng Xu, and Jun Xu. 2012. Emotion prediction of news articles from reader's perspective based on multi-label classification. In *International Conference on Machine Learning and Cybernetics*, volume 5, pages 2019–2024, Xian, Shaanxi, China, July. IEEE.
Zhang, Min-Ling and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.
Zhao, Zhenjie and Xiaojuan Ma. 2019. Text emotion distribution learning from small sample: A meta-learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3948–3958, Hong Kong, China, November.
Zhou, Deyu, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Austin, Texas, November. Association for Computational Linguistics.
Zoppelli, Luca. 2001. Il teatro dell'umane passioni: note sull'antropologia dell'aria secentesca. In *I luoghi dell'immaginario barocco*. Liguori, Napoli, Italia.

Appendix A: Label Distribution across Periods

Figure 1 shows the label distribution at the aria level in both AriEmozione 1.0 and AriEmozione 2.0 . We use either the year of creation or the year of the first performance to allocate each opera and produce non-overlapping bins of five years.

Tristezza is the most represented emotion in both corpora, being the most frequent in eight out of twelve periods in AriEmozione 1.0 and in all nineteen periods in AriEmozione 2.0 . *Rabbia* is the second one, being the most frequent emotion in the four periods of AriEmozione 1.0 and the second in all periods but one from AriEmozione 2.0 . *Paura* is the least represented emotion in almost all intervals of both collections. By looking at all the emotions across periods, the emotion distribution is fairly stable in both the original and extended corpus.

Appendix B: One-Versus-All Self Learning Annotation of the AriEmozione 2.0 Corpus

Here we describe in detail the process to produce the silver annotations for the instances in the AriEmozione 2.0 corpus (cf. Section 5.1). We started by merging the training and development partitions of AriEmozione 1.0 and produced six one-versus-all collections, each corresponding to one emotion with the instances belonging to the other five classes simply turned into class all. Each of the six collections is then re-partitioned into training and development partitions on an 8:2 ratio. Since we are interested in spotting the actual emotion of each new instance, we adopt precision as our single evaluation metric. The model we use is the best one from our experiments on corpus AriEmozione 1.0 (cf. Section 4.3): a 2-layer NN with TF-IDF character 3-grams.

Algorithm 1 sketches the iterative self-learning annotation process, which is applied in parallel for each of the six emotions. The input to the process includes the new training and development collections for each binary task and the raw instances, which lack annotation (lines 2–4). The output consists of the instances in the raw dataset, with emotions pre-labeled. In each iteration, baseline binary classifiers are trained on the existing labeled training data and evaluated on a fix development set (lines 7–8). The same model is applied to the set of raw instances, which are then ranked according to the classification confidence score, and the top instances are selected as candidates to join the training material (lines 9–10). Such candidates are added to the original training material at this iteration, a new model is trained from scratch, and its performance on the development set gets measured (lines 12–13). If the resulting precision is higher than

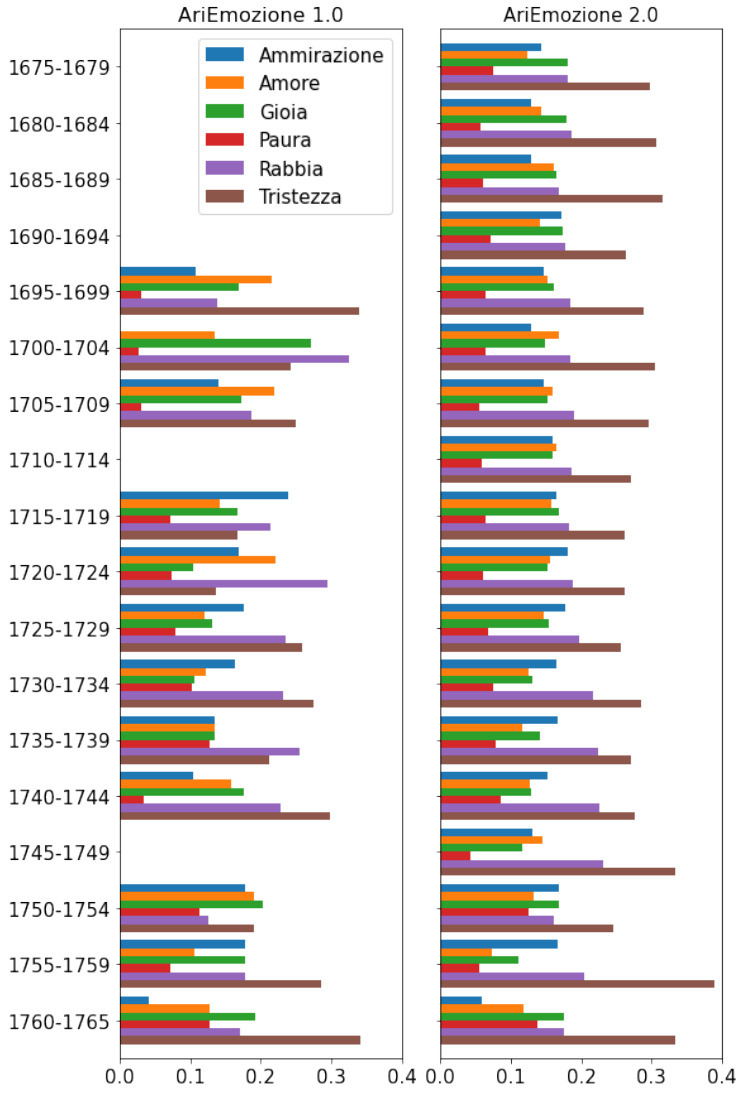

Figure 1
Emotion distribution across 5-year periods in both AriEmozione 1.0 (manual annotation; left) and AriEmozione 2.0 (automatic annotation; right) starting in 1675.

the baseline model, we transfer the new instances from the raw dataset to the training material for the next iteration (lines 15–16). Otherwise, the instances are kept in the raw set and a new iteration begins. The process runs until a minimum evaluation score is reached or the raw material gets all integrated to the training one. In our experiments, the second condition was never met. At last, 952 verses of `ammirazione`, 2,070 verses of `amore`, 1,120 verses of `gioia`, 1,320 verses of `paura`, 1,890 verses of `rabbia`, and 2,820 verses of `tristezza` were pre-selected.

Algorithm 1 Pseudo-code for the self-learning annotation process.

$E \leftarrow$ [amore, gioia, ammirazione, rabbia, tristezza, paura]
$tr[e] \leftarrow$ 80% of AriEmozione 1.0 , binarised as e vs rest $\forall e \in E$
$te[e] \leftarrow$ 20% of AriEmozione 1.0 , binarised as e vs rest $\forall e \in E$
$raw \leftarrow$ full set of fresh, non-annotated, instances
while $precision[e] < thres \forall e \in E$ **or** $raw \neq \emptyset$ **do**
    for e $\in E$ **do**
        $model[e] \leftarrow$ train binary classifier on $tr[e]$
        $precision[e] \leftarrow$ evaluate $model[e]$ on $te[e]$ and record the performance
        $scores[e] \leftarrow$ predict all raw instances record $model[e]$'s confidence scores
        $top[e] \leftarrow$ top-k instances in raw with the highest confidence scores

        $model'[e] \leftarrow$ train binary classifier on $tr[e] \cup top[e]$
        $precision'[e] \leftarrow$ evaluate $model'[e]$ on $te[e]$) and record the new performance
        if $precision'[e] > precision[e]$ **then**
            $tr[e] \leftarrow tr[e] \cup top[e]$
            $raw \leftarrow raw \setminus top[e]$
        else
            continue
        end if
    end for
end while

Table 1
Accuracy and F_1-measure on the test set of the AriEmozione 1.0 corpus using different training partitions: AE1.0$_{tr}$=training set from AriEmozione 1.0 ; AE1.0$_{de}$=development set from AriEmozione 1.0 ; AE2.0=full AriEmozione 2.0 .

train material	**Acc**	**F_1**
AE1.0$_{tr}$∪AE1.0$_{de}$	0.413	0.394
AE2.0	0.417	0.411
AE1.0$_{tr}$∪AE1.0$_{de}$∪AE2.0	0.419	0.413

Appendix C: Impact of AriEmozione 2.0 on the Performance at the Verse Level

Before shifting to the aria granularity level, we performed an experiment to observe the impact of the silver data from AriEmozione 2.0 in the verse-level classification. We trained a 2-layer neural networks with TF–IDF character 3-grams (the best configuration in Table 4) on (*i*) training plus development sets from AriEmozione 1.0 , (*ii*) AriEmozione 2.0 alone, and (*iii*) the union of both. We evaluated the three models on the testing partition of AriEmozione 1.0 . We repeat each experiment three times to enhance the reliability of the results and report the arithmetic mean of the outcomes.

Table 1 shows the results. The presence of the instances from AriEmozione 2.0 , even when used alone enhance the overall performance only slightly. Still, it boosts significantly the prediction performance for some of the classes; in particular *amore* and *paura*. Table 2 shows the diagonal values of the associated confusion matrices. When the model is exposed to instances from AriEmozione 1.0 alone, the precision on both class

Table 2
Diagonal values of the confusion matrices of the predictions on the test set of AriEmozione 1.0 when the models get trained with different data partitions: AE1.0$_{tr}$=training set from AriEmozione 1.0 ; AE1.0$_{de}$=development set from AriEmozione 1.0 ; AE2.0=full AriEmozione 2.0 .

train material	**ammiraz.**	**amore**	**gioia**	**paura**	**rabbia**	**tristezza**
AE1.0$_{tr}$∪AE1.0$_{de}$	0.333	0.006	0.300	0.067	0.532	0.600
AE2.0	0.556	0.234	0.276	0.400	0.441	0.550
AE1.0$_{tr}$∪AE1.0$_{de}$∪AE2.0	0.556	0.243	0.279	0.400	0.439	0.549

amore and *paura* tend to zero. Adding the new material from AriEmozione 2.0 rises the precision on both classes 0.243 and 0.400, at the cost of a lower performance on some of the other emotions.