AriEmozione: Identifying Emotions in Opera Verses

Francesco Fernicola\textsuperscript{1}, Shibingfeng Zhang\textsuperscript{1}, Federico Garcea\textsuperscript{1}, Paolo Bonora\textsuperscript{2}, and Alberto Barrón-Cedeño\textsuperscript{1}

\textsuperscript{1}Department of Interpreting and Translation
Università di Bologna, Forlì, Italy
\textsuperscript{2}Department of Classical Philology and Italian Studies
Università di Bologna, Bologna, Italy

\{francesco.fernicola, zhang.shibingfeng\}@studio.unibo.it
\{federico.garcea2, paolo.bonora, a.barron\}@unibo.it

Abstract

We present a new task: the identification of the emotions transmitted in Italian opera arias at the verse level. This is a relevant problem for the organization of the vast repertoire of Italian Opera arias available and to enable further analyses by both musicologists and the lay public.

We shape the task as a multi-class supervised problem, considering six emotions: love, joy, admiration, anger, sadness, and fear. In order to address it, we manually-annotated an opera corpus with $2.5k$ verses —which we release to the research community—and experimented with different classification models and representations. Our best-performing models reach macro-averaged $F_1$ measures of $\sim0.45$, always considering character 3-grams representations. Such performance reflects the difficulty of the task at hand, partially caused by the size and nature of the corpus, which consists of relatively short verses written in 18th-century Italian.

1 Introduction

Opera lyrics have the function of expressing the emotional state of the singing character. In 17th- and 18th-century operas, characters brought on stage passions 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 as a whole (Zoppelli, 2001; McClary, 2012). Being able to automatically identify the emotions expressed by the different arias of each work would provide scholars with a useful tool for a systematic study of the repertoire. The technology to identify the emotion(s) expressed by an aria represents an effective tool to study the vast repertoire of arias and characters of this period for musicologists and the lay public alike. As an aria may express more than one emotion, we go one granularity level lower—at the verse level. The task is defined as follows:

Identify the emotion expressed in a verse, in the context of an aria.

In order to do that we created the AriEmozione 1.0 corpus: a collection of 678 operas with $2.5k$ verses, each of which has been manually annotated with respect to emotion. We experimented with different supervised models (e.g., SVMs, neural networks) and text (e.g., character $n$-grams and distributed representations).

Our experiments show that, regardless of the model, character 3-grams outperform all other representations, reaching weighted macro-averaged $F_1$ measures of $\sim0.45$. Under-represented classes (e.g., fear) are the hardest to identify. Others, such as anger and sadness, being both negative, are often confused between each other.

The rest of the contribution is distributed as follows. Section 2 describes the AriEmozione 1.0 corpus. Section 3 describe the explored models and representations. Section 4 discusses the experiments and obtained results. Section 5 overviews some related work. Section 6 closes with conclusions and proposals for future work.
First of all, thank you for helping with this work. We are a group of researchers from the Department of Classical Philology and Italian Studies and the Department of Interpreting and Translation, both at UniBo. Your work will help us to produce artificial intelligence models to analyze 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 analyze.
- **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 in the AriEmozione 1.0 corpus.

## 2 The AriEmozione 1.0 Corpus

The corpus AriEmozione 1.0 is a subset of the materials collected by project CORAGO.\(^1\) AriEmozione 1.0 contains a selection of 678 operas composed between 1655 and 1765. We consider the lyrical text in the arias only. A. Zeno and P. Metastasio are among the most represented librettists in the corpus (∼30% of the operas); they are two of the most representative and prolific librettists of the 18th century. All texts are written in the 18th century Italian and articulated in verses and stanzas.

We labeled the emotions transmitted by every single verse, as we observed that this is the right granularity to obtain full text snippets expressing one single emotion. René Descartes wrote in 1649 “Les passions de l’âme”, a sort of compendium of all possible emotions and their possible causes (Garavaglia, 2018). For the sake of concreteness, we leveraged Parrott’s (2001) tree of emotions classification.

An extra class nessuna (none) applies mostly to verses with non-actionable words only, neglected in the current experiments.

Two native speakers of Italian annotated all 2,473 instances independently considering 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 very doubtful.

We measured the Cohen’s kappa inter-annotator agreement (Fleiss et al., 1969) at this stage on the primary emotion. The result was 32.30, which is considered as a fair agreement. This 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 coincided in 68% of the instances. These numbers reflect the complexity of the task.

The same annotators gathered together to discuss and consolidate all dubious instances. Table 1 shows the number of instances per class for each corpus partition: training, development, and test set. The verse average length

|     | train | dev | test | overall |
|-----|-------|-----|------|---------|
| amo | 289   | 36  | 37   | 362     |
| gio | 274   | 31  | 39   | 344     |
| amm | 289   | 23  | 30   | 342     |
| rab | 414   | 84  | 64   | 562     |
| str | 503   | 61  | 54   | 618     |
| pau | 166   | 15  | 15   | 193     |
| nessuna | 38 | 3 | 11 | 52 |
| total | 1,973 | 250 | 250 | 2,473 |

Table 1: AriEmozione 1.0 corpus statistics.

The first level of such tree includes six primary emotions: love, joy, surprise, anger, sadness, and fear. Based on the nature of the material under review, we substitute surprise with admiration, ending with the following six classes:

- **Amore (love)** incl. affection, lust, longing.
- **Gioia (joy)** incl. cheerfulness, zest, contentment, pride, optimism, enthrallment, relief.
- **Ammirazione (admiration)** admiration or adoration of someone’s talent, skill, or other physical or mental qualities.
- **Rabbia (anger)** incl. irritability, exasperation, rage, disgust, envy, torment.
- **Tristezza (sadness)** incl. suffering, disappointment, shame, neglect, sympathy.
- **Paura (fear)** incl. horror and nervousness.

\(^{1}\)CORAGO is the Repertoire and archive of Italian opera librettos. It constitutes the first implementation of the RADAMES prototype (Reper- torizzazione e Archiviazione di Documenti Attinenti ai Melo- dramma E allo Spettacolo) (Pompilio et al., 2005); http://corago.unibo.it.
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

Guardami e tutto obblio e a vendicarti io volo; di quello sguardo solo io
Look at me, all else is forgotten and I haste to avenge you;
only I shall remember that gaze

Su la pendice alpina dura la quercia antica e la stagion nemica per lei fatal
Up on the slope of the mountain the ancient oak tree still
lives on, and the adverse season poses no fatal threat

In questa selva oscura entrai poc'anzi ardito; or nel cammin smarrito timido
I entered this dark forest not too long ago, boldly; having now
lost the path I wander around, shyly

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

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

is 72.5 ± 31.6 characters and the corpus contains 34,608 (4, 458) tokens (types).²

Table 2 shows examples of verses in the corpus, including one of each of the six emotions.

3 Models and Representations

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 baseline is a k–Nearest Neighbors algorithm (kNN), considered thanks to its success in 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 summarizes the explored configurations.³

²The corpus is available at https://zenodo.org/record/4022318.
³The 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).

Table 3: Experimental settings overview.

As for the text representations, we consider TF–IDF vectors of both character 3-grams and word 1-grams (no higher n values are considered due to the corpus dimensions). For preprocessing, we employ the spacy Italian tokenizer⁴ and casefold the texts. We also explore with dense representations, derived from the TF–IDF vectors, by means of both LDA (Hoffman et al., 2010) and LSA (Halko et al., 2011).

⁴https://spacy.io/models/it
In both cases, we target reductions to 16, 32, and 64 dimensions. As for embeddings, we adopted the pre-trained 300-dimensional Italian vectors of FastText (Joulin et al., 2017), and tried with character 3-grams and words.

### 4 Experiments

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$_1$ measure 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.\footnote{The full batch of results is available at https://docs.google.com/spreadsheets/d/1Ztjry2mJs6ufCZM1O5CQRy25Yd7nCo0b0KGX1nWO/edit?usp=sharing} Character and word n-grams TF-IDF, 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 appears to be 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.

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 projecting to 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.

It is worth noting that all models tend to confuse rabbia and tristezza. Table 5 shows the confusion matrix for the best model on test. These two emotions get confused between 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.

### 5 Related Work

Building on the numerous pre-existing studies focusing on sentiment analysis (Ain et al., 2017; Shi et al., 2019), some researchers have...
been seeking to dig deeper, towards multi-class emotion analysis. Most of the work thus far has focused on social media (e.g. Twitter). Bouazizi and Ohtsuki (2016) built a classifier for seven emotions: happiness, sadness, anger, hate, sarcasm and neutral; i.e. an overlap of five classes with respect to the ones in ariEmozione. In contrast to our experiments, they focused on exploiting the polarity of the words from each instance to be fed to a random forest classifier.

Balabantaray et al. (2012) tried to distinguish among happy, sad, anger, disgust, fear and surprise using WordNet Affect (Valitutti et al., 2004). Given that no WordNet-Affect is currently available for Italian, such an approach is unfeasible.

Promising work has been carried out on news articles (Ye et al., 2012), news headlines (Strapparava and Mihalcea, 2007) and children’s narrative (Alm et al., 2005). While a lexical-based approach is the most frequent to determine the binary positive vs negative classification, Strapparava and Mihalcea (2007) combined a high-dimensional word space produced from word TF-IDF vectors with a set of seed words to predict the valence of a text exploiting the syntagmatic relations between words. A bottom-up semantic approach has also been proposed (Seal et al., 2020).

To the best of our knowledge, no work in the field of either emotion or sentiment analysis has been performed on operas.

6 Conclusions and Future Work

We addressed the novel problem of emotion classification of opera arias at the verse level. The task is interesting because of the lack of automated tools for the analysis of operas and challenging due to both the language used in 17th- and 18th-century lyrics and the complexity to produce the necessary amount of quality supervised data.

We explored with various classification models and representations. A neural network with two hidden layers fed with a simple TF-IDF character 3-gram representation is among the most promising approaches to the problem. Among the six possible emotions, the most difficult to identify are rabbia and tristezza, which tend to be confused with each other, followed by ammirazione, which is often confused by gioia. In order to foster the research on this topic, we release the AriEmozione 1.0 corpus to the community (cf. footnote 2).

As for the future work, we intend to increase the size of the AriEmozione 1.0 corpus by means of active learning (Yang et al., 2009). Once a larger data volume is produced, we plan to explore with models to identify the emotion at the aria rather than at the verse level. Following the theory of emotion proposed by Plutchik (1980), we could identify the emotion of a whole aria by combining the emotions at the verse level, and then conduct experiments to verify which granularity is more adequate as a single emotion unit. In order to address the issue of emotional polysemy and ambiguity of aria verses, we aim at producing explainable models by highlighting the specific fragments expressing the emotion.

Another interesting alternative is the one highlighted by Zhao and Ma (2019), who adopted an efficient meta-learning approach to augment the learning ability of emotion distribution; i.e. the intensity values of a set of emotions within a single sentence, when the training dataset is small, as in the AriEmozione 1.0 corpus.

Acknowledgments

This research is 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

Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6):424.

Cecilia O. Alm, 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 (HLT–EMNLP)*, pages 579–586.

Rakesh C. Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53.

Mondher Bouazizi and Tomoaki Ohtsuki. 2016. Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter. In *IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

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

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

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

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Armand Joulin, 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. ACL, April.

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

W. Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. Psychology Press.

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

Angelo Pompilio, 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*. IEEE.

Dibyendu Seal, Uttam K. Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In Milan Tuba, Shyam Akashe, and Amit Joshi, editors, *Information and Communication Technology for Sustainable Development*, pages 423–430, Singapore. Springer Singapore.

Yong Shi, Luyao Zhu, Wei Li, Kun Guo, and Yuanchun Zheng. 2019. Survey on classic and latest textual sentiment analysis articles and techniques. *International Journal of Information Technology & Decision Making*, 18(04):1243–1287.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1).

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, page 917–925, New York, NY. Association for Computing Machinery.

Lu Ye, Rui-Feng Xu, and Jun Xu. 2012. Emotion prediction of news articles from reader’s perspective based on multi-label classification. In *2012 international conference on machine learning and cybernetics*, volume 5, pages 2019–2024. IEEE.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Zhenjie Zhao 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 (EMNLP-IJCNLP)*, pages 3957–3967, Hong Kong, China, November. Association for Computational Linguistics.

Luca Zoppelli. 2001. Il teatro dell’umane passioni: note sull’antropologia dell’aria secentesca. In *I luoghi dell’immaginario barocco*. Liguori, Napoli, Italia.