Noname manuscript No.
(will be inserted by the editor)

The Italian Music Superdiversity

Geography, Emotion and Language: one resource to find them, one resource to rule them all.

Laura Pollacci · Riccardo Guidotti · Giulio Rossetti · Fosca Giannotti · Dino Pedreschi

Received: date / Accepted: date

Abstract Globalization can lead to a growing standardization of musical contents. Using a cross-service multi-level dataset we investigate the actual Italian music scene. The investigation highlights the musical Italian superdiversity both individually analyzing the geographical and lexical dimensions and combining them. Using different kinds of features over the geographical dimension leads to two similar, comparable and coherent results, confirming the strong and essential correlation between melodies and lyrics. The profiles identified are markedly distinct one from another with respect to sentiment, lexicon, and melodic features. Through a novel application of a sentiment spreading algorithm and songs' melodic features, we are able to highlight discriminant characteristics that violate the standard regional political boundaries, reconfiguring them following the actual musical communicative practices.

Keywords music data analytics · sentiment pattern discovery · music sentiment analytics · multi-source analytics · music sentiment analysis · superdiversity

1 Introduction

Music's origins, like those of language, are hidden in the most ancient past of the humanity's history. For centuries music has been part of human civilization and each culture has given birth to its own music [31]. However, nowadays, the globalization and the progressive reduction of geographical distances facilitated by ubiquitous of the World Wide Web and the media has brought

L. Pollacci ✉ · D. Pedreschi
Department of Computer Science, University of Pisa, Italy
E-mail: laura.pollacci@di.unipi.it, dino.pedreschi@unipi.it

R. Guidotti · G. Rossetti · F. Giannotti
ISTI-CNR, Pisa, Italy
E-mail: {name.surname}@isti.cnr.it

the collapse of the music barriers. During the last decade, we have witnessed to a constant growth of online streaming services that allow a broad open, "democracy" and omnipresent access to the widest choice of music ever seen. Famous artists just like emerging or niche ones gained a global visibility in unimaginable even wrap a few years ago.

In this quickly evolving panorama, music threatens to lose its geographical-cultural characterization. This multifaceted scenario has drawn the attention of the researchers. Indeed several works inspect music from a wide range of perspectives. In fact, the actual music scene offers multiple research suggestions, i.e., geographical and cultural connotations, data collection from online services, specific music' features inference, users behaviors, and tastes, etc. Recently, there is high attention on music sentiment analysis. This sentiment specific context domain exploited several techniques and involved different kinds of features, that will be addressed in follow in Section 2. In the last decade, particular attention turned on corpus-based methods. Despite creating songs polarity tagged datasets is not an easy task [11], datasets of songs labeled with emotions, and polarity tagged lexicons are essentials prerequisite to computer those classification models. As suggested in [11], a music dataset should observe four main characteristics, such as *(a)* strong polarization; *(b)* easily understandable labels taxonomy; *(c)* high coverage and large size (at least 1,000 lyrics); and *(d)* publicly available.

As an example, All Music Guide (AMG)[1] [14] has invested both from the economic and from the human points of view considerable resources, to annotate high-quality emotional music databases. Consequently, they are unlikely to share their data publicly. Besides, is also quite impossible to manually tag large datasets. Indeed, freely available manually annotated datasets are generally small in size. In [39], Trohidis et al. created and publicly shared a dataset of 593 songs all of which have been annotated employing 6 emotions by 3 experts. In [40] Turnbull et al. collected CAL500, a dataset of songs annotated by at least three annotators. CAL500 is composed of one song for 500 artists. The literature underlines some different main solutions to avoid these problems. A first approach to collect emotion annotations is a survey. Surveys and therefore the crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk)[2], represent a straightforward way to gather this kind of musical contents. A second approach to collect intelligence involves social tagging and online services (such as Spotify, Apple Music, Last.fm). In particular, Last.fm[3], is a music discovery website with a wide heterogeneous community of music listeners. Users can contribute social unstructured text tags[21]. As opposed to AMG, some music platforms like Last.fm, Genius, AllMusic, and Spotify have made its data through a public APIs. While Last.fm, as well others like Spotify, represents a useful resource for researchers, in [12] are underlined several problems with

[1] All Music Guide website: All Music Guide

[2] The Amazon Mechanical Turk (MTurk) is a web server for works that require human intelligence. Developers can exploit the service to build human intelligence directly into their applications. MTurk.

[3] Last.fm website: Last.fm

social tags, such their sparsity, the fake tagging, popularity bias, lexical tags variations, etc. Despite this, Last.fm has been broadly used [19,22] in music sentiment analysis tasks, even though obtained datasets are not made public. Indeed, as already underlined in [11], as far as may be difficult to believe, still today no lyrics sentiment dataset fulfills all the four conditions mentioned before. In this context, the Italian music scenario is even more dramatic.

In this paper, we present a data-driven investigation starting from the building of cross-service datasets. For building the dataset, we aggregate information collected from different online resources, such as Spotify, Sound-Cloud, Echonest, and Wikipedia. From the data, we inspect a particular context: the Italian music scene. In particular, we analyzed the peculiarity of regional music leveraging the geographical and emotional dimensions, through melodic and lexical features, to evaluate the Italian music *"superdiversity"*. The concept of "superdiversity" was theorized by Vertovec [41,42] regarding the changes and contexts that have affected migratory flows around the world. These demographic changes, which Vertovec defined superdiversity, are the result of globalization and outline a change in the overall level of migration patterns. In a broader context, the concept of superdiversity aims to acquire the increasingly complex and less predictable set of relationships between ethnicity, citizenship, residence, origin, and language. The superdiversity is based on the perception of sociocultural communities, for example, linguistic minorities, their characteristics, and dynamics. This concept aims at identifying the phenomena that violate the boundaries of the political, historical, social, cultural and linguistic monocentrism, restricted to a closed spatial framework.

The main factors able to identify homogeneous groups of individuals are spatial and linguistic. These dimensions vary in a diachronic sense and their changes can be observed over time. Regarding the language, observations of superdiversification have led to consider the notion of language as the set of trans-idiomatic practices, in order to describe the communicative practices in particular places and situations. In particular, we focused on some questions. Are we observing a growing standardization of Italian music contents? It exists lexical and melodic-specific features able to discriminate the music from a geographical, lexical and emotional point of view? As is well-known, emotions can also be induced by music. If emotions may be caused by music, what is the weight of melodies and lyrics in this process?

To answer these questions, we develop two different kinds of musical profiles through a melodic and a lexical approach. In particular, regarding this latter point, we exploit a sentiment spreading epidemic model[32] originally applied only to English tweets. By choosing to experience this model on Italian music lyrics, we aim to evaluate Italian music superdiversity since the resulting dictionary is strongly population-dependent, "*thus can provide important insight into how language is used by various populations [...]*"[32]. Furthermore, we aim to *(a)* evaluate the model performance on large texts instead of smalls, and *(b)* apply the model to a language different from the English.

The rest of the paper is organized as follows. In Section 2 is discussed the literature involving analysis related to music data gathered by online services and the literature about a sentiment analysis specific domain, the music sentiment analysis. Section 3 introduces the building our cross-service composite dataset, the lexical resources involved, the preprocessing steps performed to make processable the data gathered from different online sources, and the problems encountered during the data mining and the preprocessing phases. In Section 4 we show the Italian regional profiles identified applying an unsupervised learning strategy and a method for enhancing lexicon-based sentiment analysis by extending a base lexicon of terms. In Section 5 we evaluate the obtained profiles by applying the two different procedures, and we discuss the relationships and peculiarities between the lexical and melodic profiles. In addition to this, we show a sentiment analysis experiment. Finally, in Section 6 are discusses the conclusions and some future research directions.

2 Related Work

From the beginning of the 21st century, the music scene is facing ever-increasing growth of attention from the scientific community, empowered by the permeation of World Wide Web and the music-dedicated platforms into daily life. As mentioned above, the research on the Italian music domain is sparse to nonexistent. To the best of our knowledge, the unique contribution focused on this specific domain is proposed in [28]. Authors present the Rapscape corpus, a POS-tagged and lemmatized lexical resource containing about 16,000 Italian rap songs grouped by artist. As declared by authors, the building of this resource aims to overcome the lack of resources about Italian rap music. The initial dataset is gathered by exploiting both Discogs[4] and Spotify APIs. Unfortunately, the article, the relating results, and in particular, the resource is not attested in the music analysis literature and is not publicly available or accessible. Indeed, we are aware of the work only thanks to the external collaboration in the contribution (restricted only to the data preprocessing phase) of one of our authors[5].

Regarding the worldwide music analysis panorama, the literature on music analysis is noticeably large. Several works [33, 34, 29, 8, 19, 20, 18] have analyzed data gathered by online services in order to analyze different phenomena related to the online music consumption, such model diffusion of new music genres/artists, behaviors and tastes of users, recommendation models, classification of semantic states inferred by lyrics, music classification, and so on. In particular, the music sentiment analysis, also called *music mood recognition*, exploits and combine several techniques, such as machine learning, and data mining to classify songs into polarity classes. The literature displays that, to accomplish this task, can be involved and also combined several different kinds

[4] Discogs website: Discogs

[5] AIDAinformazioni, Anno 34, numero 1-2, 2016: Il linguaggio del Rap. Possibilità di un'analisi multidisciplinare

of features like as melodic/audio, lyrics or metadata. In [18], AllMusic metadata are used to create a categorical representation of music emotions. Their results show that many individual mood terms are highly synonymous or express different aspects of a more general mood class. This leads in some cases, to better identification of the underlying mood by decreasing mood vocabulary size. Authors also propose a five-class music categorization and a set of not even thirty mood's popular terms, recommending to reduce mood lexicons in a set of classes rather than using immoderate individual mood terms.

One of the two most attested trends in music emotion recognition is to use self-created datasets. In [19] Last.fm tags are used to build an about 5000 songs dataset tagged with 18 mood categories. Authors employ a binary approach for all the mood categories, independently if songs have or not category tags.

As underlined above, a second trend is gathering intelligence about music by human feedback employing surveys, in particular exploiting Amazon MTurk. In [23] is proposed an inquiry of the possibility to apply this service to music mood ground truth data creation. By comparing MTurk data with those of MIREX AMC 2007 task[6], authors report a similar distribution on the MIREX clusters. Despite the fact that authors warn about problems which can diminish annotations qualities, such as spamming, they conclude that MTurk represents a valid approach option. In [26] a lyrics dataset based on Valence-Arousal model of Russell [35] is created employing (AMG) tags. Likewise [13, 31] and the work we present, Affective Norms for English Words (ANEW)[9] is used as a lexical resource. Once classified AMG tags in the four Russell's model quadrants using ANEW, songs are categorized using the obtained tags, and finally, annotations are evaluated by employing human evaluators. As underlined in [11], this tagged dataset is one of the few public lyrics datasets.

Finally, another attested tendency both in music sentiment analysis and in the analysis of the phenomena related to the music is to build multimodal music datasets [27,36] which merge and combine several kinds of features. In [24] a semi-supervised approach is used to study the problem of the music artist genre identification both from lyrics and melodic features, as acoustic ones. The similarity between the 45 analyzed artists is identified by exploring AMG artists' pages. Obtained results report that lyrics and sound performances are comparable. Authors of [36] presents Musiclef, a professionally created multimodal dataset of about 1300 popular songs. Musiclef combines several features such general metadata, Last.fm tags, audio features together with web pages and labels provided by expert annotators. In Musiclef songs are first labeled using a seed set of 188 terms, after reduced to 94. Nevertheless, the professionally of the dataset, this wide range of labels may seem redundant, superfluous and not reliable [11]. The approach proposed in [27] merges either textual and melodic features. The work exploits a 100 items dataset of popular songs annotated for emotions at line level using Amazon MTurk. The dataset is then used to explore the automatic recognition of emotions in songs. The results demonstrate that *(a)* emotion recognition can be performed using either melodic or

[6] 2007 Audio Music Mood Classification: 2007:Audio Music Mood Classification

textual features, and especially that *(b)* the joint usage of these two dimensions leads to improve significantly classifications based on only one dimension at a time. The obtained dataset is available for research upon request to the authors, but due to its small size cannot be used as experimentation set [11]. On the contrary, a rich set of lyrics like the one presented in [10] lacks in the human evaluation, and therefore it too cannot be used as a ground truth set.

3 Data and Preprocessing

In this section, we describe the different types of data sources, the preprocessing steps, the lexical resources we used and the problems encountered.

To obtain a wide overview of the actual Italian music scene, we exploit musical dataset showing two different levels of spatial granularity: national and regional. These two datasets refer to *Italian* musicians (both famous and less famous), and emerging youth bands in *Tuscany*, respectively. These datasets are built by exploiting the Spotify API[5] and are composed of all the song retrieved from the Spotify[4] platform for a selected artist. In particular, the `TUSCANY` dataset refers to emerging artists participated in the 2015 edition of the "100 Band" contest promoted by "Tuscan Region" and "Controradio"[6].

For each selected artist, we collect song titles, song's *popularity score*, album titles, Spotify ID, and the list of the genres the artist is associated with. If an artist's genre is not set, the array is empty. Solutions adopted to solve the genre's lack will be explained later. Regarding the *popularity score*, each track on Spotify is characterized by a set of "Popularity bars" that if aggregates indicate how popular the track is. In general, the popularity score of a song is based on two parameters: *a)* the total number of plays compared to other tracks; and *b)* how recent those plays are. In addition to the features listed above, for each song, we collect a set of musical features provided by Spotify: *acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo,* and *valence*. All the features range in [0, 1] with the exception of *duration*, *loudness*, and *tempo*. For this reason, we normalize these latter features to align all the feature scales.

Briefly, *acousticness* is related with how much the track is acoustic; *danceability* describes how suitable a track is for dancing based on the combination of musical values (i.e., tempo, rhythm stability, beat strength, and overall regularity); *duration_ms* indicates track's duration in milliseconds; *energy* is a measure that represents a perceptual measure of intensity and activity, which includes dynamic range, perceived loudness, timbre, onset rate, and general entropy; *instrumentalness* predicts whether a track contains no vocals; *liveness* is related to the presence of an audience in the recording; *loudness* is the measure of the overall loudness of a track in decibels (dB); *speechiness* is related to the presence of spoken words in a track; *tempo* is the tempo of a track in beats per minute (BPM); and *valence* describes the musical positiveness.

Moreover, since we are interested in musical data also from a geographical point of view, we integrate this information with a geographical field. For this

Dataset	#Artist	#Tracks	#Genres	#Lyrics
`ITALY`	2,608	503,202	234	59,330
`TUSCANY`	513 (58)	24,147	28	101

Table 1: Datasets statistics. Within brackets are reported the number of artists for which at least a single song lyric was available.

Lexicon	#Lemmas	#Italian lemmas	#Balanced lemmas
ANEW	1,034	1,034	1,034
SentiWordNet	117,659	45,882	3,932
Bad words	550	500	500

Table 2: Lexicon statistics

purpose, we use the artists' Spotify ID as research keys in the Echonest API[1]. Using this API, we obtain the Italian region where each musician comes from.

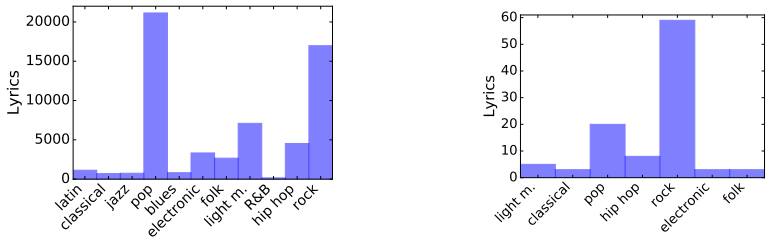
Furthermore, to perform our analysis, we integrate these musical datasets with the lyrics of the songs retrieved for each artist. While the `ITALY-lyric` dataset is collected using Sound-Cloud API[3], the majority of the Tuscany emerging bands' lyrics are not available on public platforms. Due to this reason, the `TUSCANY-lyric` dataset is built by extracting information from the results of a survey. By using the Google Form service[2] we gathered both musical and personal data regarding artists who participated in the Tuscany 100 Band contest of 2015. In Table 1 we describe the details of these datasets.

3.1 Lexical resources

To employ the sentiment spreading algorithm and validate obtained results, we exploit several lexical resources of lemmas annotated with sentiment valences.

Affective Norms for English Words (ANEW). The Affective Norms for English Words[9] provides a set of normative emotional ratings for 1,034 terms. Using Self-Assessment Manikin (SAM), an effective rating system, for each word is provided human subjects' emotional ratings concerning "pleasure", "arousal", and "dominance". Each dimension is represented as a number from 0 to 10. In the sentiment spreading algorithm is used only the "pleasure" dimension as a sentiment valence, as evaluated by both male and female subjects. Were "pleasure" represents positive versus negative emotions. We choose ANEW as seed lexicon because it is *a)* an established dictionary in the literature; *b)* suitable for large-scale texts; *c)* already attested to estimate and add valence scores for musical lyrics [13, 26]; and because *d)* it is the seed lexicon exploited by the authors of the sentiment epidemic spreading algorithm.

SentiWordNet. SentiWordNet[15] is a publicly available lexical resource. Each word in the resource is associated with three numerical scores (*Obj(s)*,



(a) ITALY (b) TUSCANY

Fig. 1: Genres distribution among datasets.

Pos(s) and *Neg(s)*) related to how objective, positive, and negative is, respectively. Following the procedure described in [32] and adopting the method presented in [16], for each word in the resource we compute a unique polarity properly scaled to the interval $[0, 10]$ like ANEW.

Bad words. Conversely of [32], instead of using the Simple Main Bad Words List of the "What Do You Love" (WDYL) Google project[7] we replace it with the Full List of Bad Words Banned by Google. We choose the extended version for better identifying bad words. The Full List is a large list of words banned by Google, and it's composed of 550 items[8]. We manually label each word in the lexicon is with a 0.0 polarity score, hence is considered strongly negative.

PAISA'. The PAISA' corpus[25] is a vast collection of about 380,000 Italian texts taken from the web, for a total of approximately 25 million tokens. Together with lemmas are also provided the lemmas' frequencies.

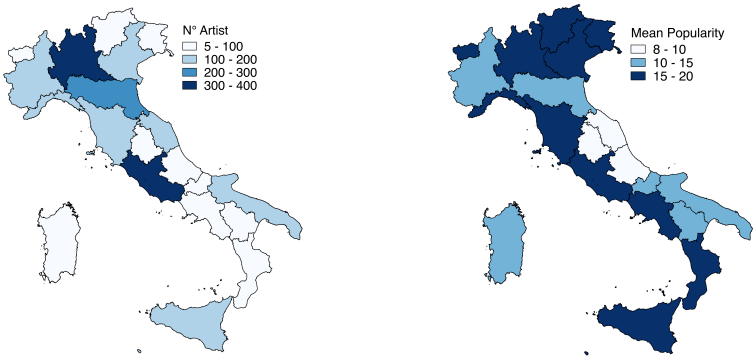
Since we are interested in apply the sentiment spreading algorithm to Italian music lyrics, we translate each word in the lexical resources presented before. To obtain the most accurate translation, we combine two different python libraries and then, we impose some constraints, such a threshold on the translation confidence score. First, each word in resources is translated by using both Googletrans[9] and Goslate[10]. English words can have different meanings, resulting in one or more translations. For this reason, we exploit the confidence score provided by Googletrans by choosing the translation with the highest score. Then we cross-check the translated words detecting the language by using the opposite library. In addition, for SentiWordNet, we also check if the translation is attested in the PAISA' list of frequencies. This step allows us to be more confident in words translation. Since polarity classes in lexicons were strongly unequal, we decide to balance them. To accomplish this task, for each class we select the n most strongly polarized lemmas, where n

[7] The service is now inactive, with the URL resulting in a 404 response.

[8] The two version of the WDYL's list can be public downloaded from Free Web Header

[9] Googletrans is a free and unlimited python library that implemented Google Translate API. For more details see Googletrans

[10] Goslate provides a python API to Google translation service by querying google translation website. More details can be found at Goslate



(a) Artists distribution over Italian regions. (b) Region average popularity score.

Fig. 2: ITALY datasets statistics.

is the number of items in the least represented class (negative). Details about lexicons obtained after the translation are reported in Table 2.

3.2 Preprocessing

The first problem faced is the lack of genres that, once obtained, are normalized and aggregated, as explained later. With this scope, we integrate our data with data from another web resource: Wikipedia[7]. For each artist with an empty genre field, we make a call to the Wikipedia Italian version using the name of the artist.

Once obtained all the genres for the musicians who lack them, we proceed by aggregating them into few major music genres classes. For this purpose, by using a list of popular music genres gathered from AllMusic and Wikipedia we match and assign each song's minor genres to the respective major class. Following this approach, we can distribute 234 different genres in only 11 major classes: blues, hip-hop, latin, electronic, folk, jazz, rock, R&B, and pop, light music and classical music. Genres distribution in datasets is shown in Figure 1.

After the songs' major music genres assignment, we address the lyrics' noise problem. In fact, lyrics show a large number of domain-specific stop words in a wide variety of lexical forms, i.e. "rit.",‘‘RIT", "[Intro]", etc. Therefore, to obtain clean texts, first, we treat lyrics' datasets using a domain-specific rule-based cleaning procedure by employing, for example, regular expressions. Once obtained standardized music lyrics, we can employ a general-purpose pipeline of Natural Language Processing (NLP). Then, all lyrics are lemmatized and Part-Of-Speech are tagged using the POS tagger TreeTagger[37,38]. Finally, we consider only words with a full semantic value, filtering by part of speech only nouns, verbs and adjectives. Therefore, for each text we obtain only significant words from the sentiment points of view.

To allow experiment reproducibility the final dataset is made publicly available on *Zenodo* at https://bit.ly/2MUUwEx.

3.3 Problems Faced

Mixing different resources referred to a unique specific domain is a not such an easy task, and this is harder when starting with user-generated contents. In addition to the difficulty in obtaining the data, processing these information presents a dual face problem. From a part, the non-standardized data variety does not allow to use standard rules. From the other part, a standard pipeline is not effective with domain-specific data. As explained before, data requires a dual-phase cleanup: an "ad hoc" supervised step to identify particular lexical forms attested, and then, an automatic standard cleaning phase. Another problem encountered is the difficulty of correct identification of the Italian authors in the international music scene. To curtail this problem, we explore only the Italian Wikipedia version of the platform. Anyway, in the musical APIs, many Italian musicians are unknown or have residual positions. Moreover, due to the commercial orientation of the APIs, less known artists are at the bottom of their lists and this can leads to songs mis-allocation. A preferential attachment phenomenon may be noted: the more famous an author is, the more likely a searched information is credited to him and vice-versa.

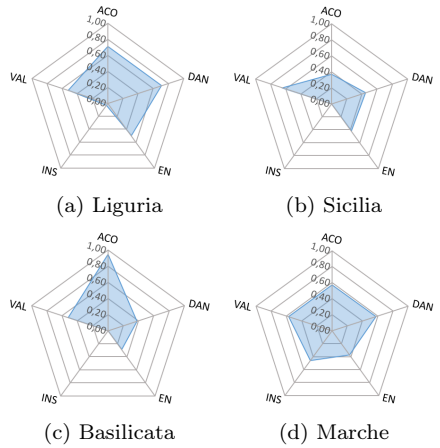
4 Italian Regional Profiles

In this Section, we describe the novel application of the algorithm presented in [32] to Italian musical lyrics, and our approach targeted to compute and evaluate Italian music regional profiles.

4.1 Regional profiles: the sound point of view

The music is one of the most ancient cultural and national expression. It can also be said that is one of the most valuable expressions of national identities, since, through its performer, the music tells the history, events, and transformations of cultures and nations. However, nowadays is particularly rare that a country can be described through a singular "cultural" entity, like a singular music genre. Conversely, as already underlined in [30] a national music scene is characterized by different aspects. Indeed, regarding the Italian music scene, each region can be characterized by the music it produces.

To obtain Italian regional profiles, we first exploit geographical information to partition the `ITALY` dataset into different subsets. Using the artists' region of provenance, for each Italian region, is built a regional Italian lyrics dataset. Figure 2 shows regional dataset statistics. In particular, Figure 2a shows the artists distribution among Italian regions while Figure 2b shows the regions average popularity. This latter score is computed through two steps. At the first step, for each artist, we calculate the average popularity over his all discography by using the popularity scores provided by Spotify. Then, we aggregated artists based on his regional provenience and, finally, we calculate each regional



(a) Liguria (b) Sicilia

(c) Basilicata (d) Marche

Fig. 3: The four Italian "super-profiles" represented by (a) Liguria, (b) Sicilia, (c) Basilicata, and (d) Marche.

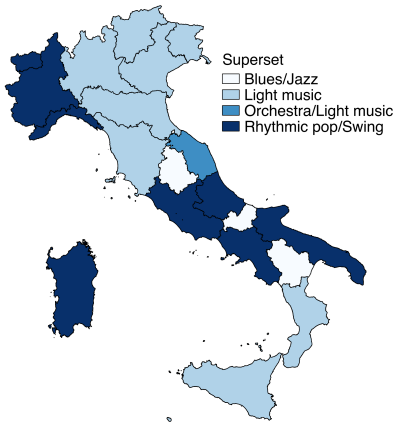


Fig. 4: Melodic profiles.

average popularity. As be can note by comparing Figures 2a and 2b, artists distribution and popularity are not correlated. Indeed, the number of artists belonging to a region has no effect on the region popularity score. Regions having a relatively low number of artists, such as Calabria and Campania show a high popularity score. To compute regional profiles, as a first step, songs in the `ITALY` dataset are grouped by artist and, for each of them, we extract a profile, as the medoid song computed over the entire artist' discography. A medoid song isn't a real song attested in the artist discography, but a "sample song" which combines the main artist's musical characteristics. Indeed, we extract each medoid song as the most artist's representative song identified by minimizing the sum of the Euclidean distances between the artist's tracks' musical features gathered from Spotify. Given the results obtained in [30] and [31], the medoids computation doesn't take into account all the ten musical features obtained from Spotify. Indeed, previous results display that features like *speechiness, liveness, loudness,* and *tempo* present similar values in each type of data aggregation and comparison, while others features are conversely high discriminant. Following a relevant music parameters selection phase similar to the one presented in [17], when computing the artists' medoid song we select only some features: *acousticness*, *energy*, *danceability*, and *instrumentalness*.

Once profiled all the Italian artists we focused our attention on the regional level by aggregating artists' medoids based on their geographical provenience. Following the approach previously explained, we extract a profile for each region, as the most representative artist among all those who come from the selected region. Our results suggest on a side that each Italian region has its music peculiarities, and on the other side that regions share several common characteristics. This is more interesting if we consider that these joint features are visible independently from the regions' geographical locations.

Focused our attention on similarities among them, we are able to group regional melodic profiles based on their analogies. This brings us to identify four "super-profiles" (Figure 4) composed as follow:

(a): Abruzzo, Valle D'Aosta, Campania, Lazio, Liguria, Piemonte, Puglia, and Sardegna;

(b): Emilia Romagna, Calabria, Lombardia, Sicilia, Trentino Alto Adige, Veneto, Toscana, and Friuli Venezia Giulia;

(c): Basilicata, Molise, and Umbria;

(d): Marche.

We underline that these clusters are the result of the obtained melodic profiles aggregation. Therefore, we cannot find in these clusters genres showed before, as for example those of Figure 1.

Figure 3 shows the profiles obtained (one for each super-profile identified). Each axis in radar graphs represents one of the musical features gathered from Spotify and described in Section 3. Super-profiles are computed based on region instead of genres; therefore, they contain a vast and heterogeneous range of kind of music. However, by observing the regional profiles characteristics, each super-profile can be described through sufficiently specific music genres. For a more in-depth comprehension, we describe each super-profile and some of their representative artists.

(a) *Rhythmic pop/Swing*: is described by upbeat pop and the most rhythmic easy listening music, and by the swing and ska. Tracks have a good beat strength and claim the listener to move, like songs played by Max Gazze (Lazio), Fred Buscaglione (Piemonte) and 99 Posse (Campania).

(b) *Light music*: is the superset characterized by the most relaxing part of pop, by the light music and by tracks played by singer-songwriters. Songs have a slow rhythm and are suitable only for slow dance. Artists that fall into this category can be recognized in Sergio Cammariere (Calabria), Carmen Consoli (Sicilia) and Laura Pausini (Emilia Romagna).

(c) *Blues/Jazz*: despite the low number of Italian region that falls into this superset, the profile is well characterized. Indeed, it is represented by the most "commercial" portion of the jazz and blues music. In particular, in this set, we found the few Italian *croones*[11], like Fred Bongusto (Basilicata).

(d) *Orchestra/Light music/Light pop*: this is a separate superset composed by only a region. Indeed, this region show mean values for each musical feature. Probably artists coming from this region are well balanced from the genres point of view. As the most famous representative artist who comes from Marche we can found only Jimmy Fontana.

To understand if emerging bands are characterized as well as the not emerging artists of the same region, we compute a melodic profile also for them. Following the method applied to not emerging Italian artists, we group the `TUSCANY` dataset by artist and, for each of them, we extract a profile. Once

[11] "Crooner" is an American term given to male singers of jazz standards, accompanied by either a full orchestra, a big band or a piano. The most famous America crooner is Frank Sinatra, despite the fact that he does not consider himself a crooner.

(a) TUSCANY (b) Toscana

Fig. 5: Comparison between Tuscany emerging youth bands and the Toscana profile computer for not emerging artists.

profiled all emerging bands, we compute the regional profile as the most representative artist among all the emerging bands. The obtained profile is shown in Figure 5a in comparison with the profile computed for the Toscana region (Figure 5b). As can be seen, the two profiles regarding Toscana are perfectly aligned. Our results show that emerging bands and famous artists present similar musical characteristics. This led us to affirm that youth bands tend to mimic characteristics of known artists, probably to receive more attention from the public instead pursue their own musical way and risk to fail.

4.2 Regional Profiles: The Lexical Point of View

Once identified and inspected prototypical types of regional music based on artists' musical features, we move on the lexical and emotional content of the songs. Indeed, we aim to investigate if using two completely different approaches it is possible to identify consistent and correlated regional profiles. To compile regional lexical profiles, we choose to apply the algorithm presented in [32] to the `ITALY-lyric` dataset.

Following the sentiment spreading algorithm method, we build a network of lemmas for each regional dataset, where each lemma corresponds to one node. Hence, the network is an unweighted co-occurrence graph based on the target lyrics to be classified. Once each regional network of lemmas is obtained, valences are added to each node in the network.

To be able to evaluate obtained dictionaries, we use cross-validation on the translated ANEW dictionary. In particular, the ANEW dictionary is randomly split into two halves. One half is used as a *seed dictionary* during the spreading epidemic process, merged with SentiWordNet and Bad words translated dictionaries. The second half is used as a test dataset, to compare the valence obtained through the algorithm to the original valence in the ANEW dictionary. The Pearson correlation is even then used to quantify the similarity and to obtain an indication of whether the process produces valid sentiment valences. As explained by the authors, the algorithm requires two parameters (range and entropy), since "*agents are better able to influence their neighbors as a consensual group rather than isolated; hence a heterogeneous group will not influence its neighbors*"[32]. To obtain optimal values for the entropy and

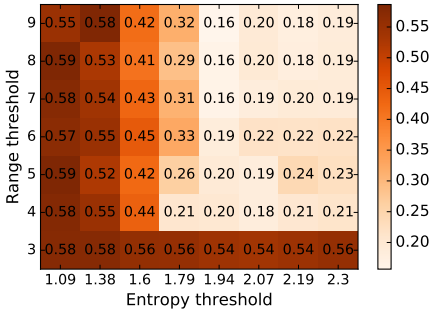


Fig. 6: Average correlation between modeled and real word valence (`ITALY-lyric` dataset).

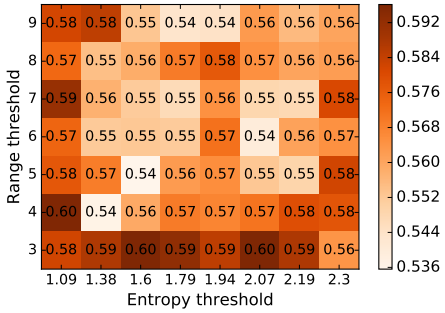


Fig. 7: Average correlation between modeled and real word valence (`TUSCANY-lyric` dataset).

range thresholds, ten runs are repeated for each couple of thresholds values. In this case, even though we will compute singular regional profiles, we decide to extract optimal threshold values by computed the runs on the entire `ITALY-lyric` dataset. This choice aims to make better compare the obtained results. Therefore, before applying the method to Italian regions, we extract the optimal range and entropy values from the entire `ITALY-lyric` dataset. We applied the same procedure to compare the `TUSCANY-lyric` dataset with `ITALY-lyric` dataset and once again we extract the optimal range and entropy values for the `TUSCANY-lyric` dataset. For transparency and not for comparison, Figure 6 and Figure 7, show the average correlation obtained for each threshold combination for both the `ITALY-lyric` and `TUSCANY-lyric` datasets respectively. As already observed by the algorithm, the range parameter is more important in obtaining higher correlations. Even then, small ranges resulting in better results. The optimal performance of the `ITALY-lyric` dataset is obtained for a range threshold of 5, and entropy threshold of 1.09. Indeed, we consider the distribution described by 10 bins of equal size. Hence the maximum entropy obtainable is approximately 2.3. For the `TUSCANY-lyric` dataset, the optimal performance is obtained for a range threshold of 3, and entropy threshold of 1.6. Note that is this case, range and entropy threshold values differ less. This result is due to the small number of lyrics in the dataset.

5 Regional Profiles Evaluation

To evaluate the procedure, we perform two different analyses. As the first criterion, we focus on the valences obtained after the sentiment spreading process using cross-validation on the translated ANEW dictionary. So, once extracted optimal threshold values from the entire dataset, we use them as algorithm's parameters. Therefore, for each Italian region, we compute ten different runs. For each run, the *seed dictionary*, composed by the 50% of the translated ANEW together with the Italian translation of SentiWordNet and

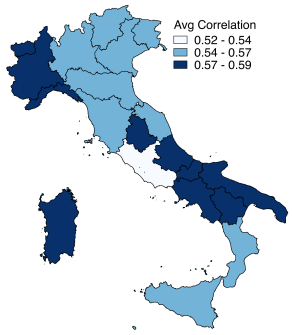
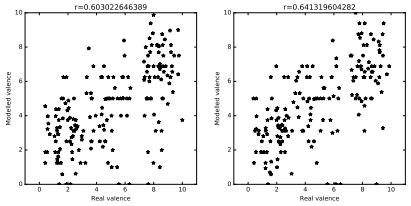
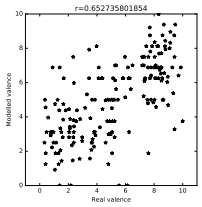


Fig. 8: Lexical profiles based on regional average correlation over ten runs.



(a) Lazio (0.60) (b) Basilicata (0.64)



(c) `TUSCANY-lyric` (0.65)

Fig. 9: Modelled and real word valence for a selected run with best parameters.

Bad words lexicon, is randomly recomputed. Following this approach, for each region, we obtained a lexicon of words labeled with a polarity score[12].

To provide more comprehensive results, Figure 8 shows for each region the average correlation over ten runs, so the regional lexical profiles. Focusing our attention on closeness among their correlation values, we can group regional lexical profiles based on their analogies. This brings us to identify three "super-profiles" (Figure 4) composed as follow:

(a): Valle D'Aosta, Piemonte,Liguria, Sardegna, Campania, Liguria, Puglia, Basilicata, Molise, Umbria, and Sardegna;

(b): Emilia Romagna, Calabria, Lombardia, Sicilia, Trentino Alto Adige, Veneto, Toscana, and Friuli Venezia Giulia and Marche;

(c): Lazio.

Moreover, Figure 9 displays for two sample regions (Lazio and Basilicata, Figure 9a and 9b respectively) the modeled and real valences on test data for selected runs with best parameters. Within brackets are shown best correlation values for the selected run. In particular, the Lazio is the Italian region with the lowest average correlation value (0,53), while the Basilicata is one of the regions with the highest average correlation value (0,59). The two examples plots show that also in the "worst" case (Lazio 9a) the valences obtained by applying the spreading epidemic sentiment algorithm to Italian music lyrics align well with human-tagged data.

[12] For example, the dictionary we obtained for Lazio for a selected run with best parameters (the same showed in Figure 9a) is composed of 11,243 Italian lemmas, each labeled with a polarity score in the range [0,10].

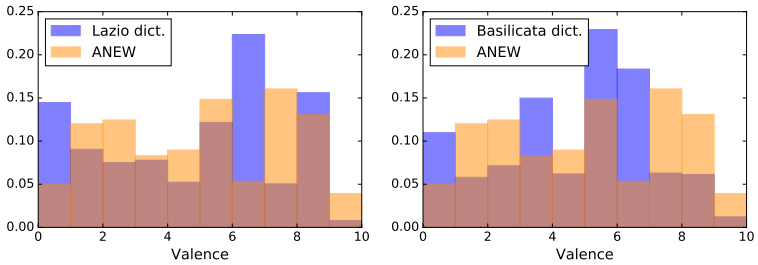


Fig. 10: Histograms of valences for ANEW and two obtained dictionaries: Lazio (left), Basilicata (right).

Finally, Figure 9c shows the modeled and real valances on the `TUSCANY-lyric` dataset for a selected run with best parameters. Here too, the obtained valances align well with human tagged data. Unfortunately, the huge difference of dataset sizes between the `TUSCANY-lyric` dataset and Toscana's lyrics won't let us compare correlation results.

Obtained results, besides validate the original method itself, validate its usage for a different language and a different kind of initial data. In fact, as already underlined, until now, the algorithm presented in [32] was tested and validated only on Twitter tweets exclusively in English. Our results bear witness the algorithm efficiency also for the Italian language and longer texts, like an entire song text. For a further comparison, we also display (Figure 10) the distribution of valences in the ANEW dictionary, compared with two obtained regional dictionaries related to Lazio and Basilicata. The results confirm that the spreading algorithm can identify different relations between lemmas by their usage in texts. With respect to the region, valences distributions differ widely. The only likeness is that in all regional lexicons is the higher number of lemmas tagged with a score in the range [0,1], and, conversely, the less number of lemmas tagged in the range [9,10] than ANEW.

Once obtained both melodic and lexical profiles, displayed in Figure 3 and Figure 8 respectively, we focus on their similarities. Starting from melodic ones, merging regional medoids, we identify four different "super-profiles" characterizable based on their genres. We obtain two major and two minor super-profiles: *Rhythmic pop/Swing* and *Light music*; *Blues/Jazz* and *Orchestra/Light music*, respectively. The two majors regional blocks are composed of 8 regions each. The *Rhythmic pop/Swing* profile includes the regions at north-west (Valle d'Aosta, Piemonte, and Liguria), the Sardegna island, and regions placed between Central Italy and south Italy, excluded Calabria, Basilicata, and Molise. The *Light music* profile is composed of a major cohesive block including all the North and North-East regions, until the Toscana, plus two southern regions, Calabria and Sicilia. Finally, observing the two minor super-profiles, the *Blues/Jazz* profile is composed of three regions of Central Italy, while the *Orchestra/Light* music profile includes only a region of Central-East Italy.

Moving to lexical profiles (Figure 8) we found three "super-profiles" instead four. However, the distribution of regions in blocks is highly related

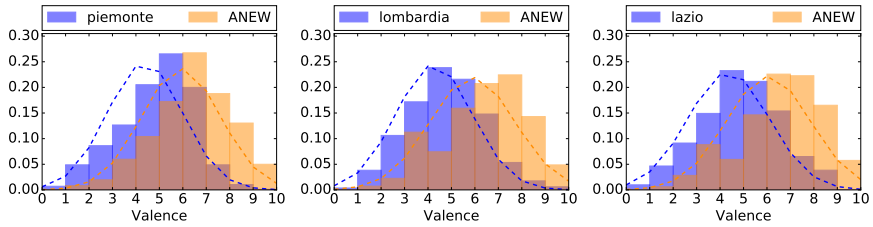
to the one previously observed. Indeed, another time we identify two major super-profiles, plus a minor super-profile composed of an alone region. The two major profiles enclose regions having highest and middle average correlations, which include ten and nine regions respectively. As can be seen, regions having highest average correlations are the same regions that fall into the *Rythmic pop/Swing*, together with regions of the *Blues/Jazz* profile. From the other side, regions having the middle average correlations are the same that fall in the Light music profile together with Marche. Alongside this allocation, Lazio shows the lowest average correlation, going out of the relative cohesive block. It is probably caused by the high number of very heterogeneous artists or by a different usage of the language by its artists.

In light of these considerations, we observe that using different features over the geographical dimension leads to two similar, comparable and coherent results. In practice, through the language as a set of trans-idiomatic practices, and specific musical features, we are able to highlight discriminant characteristics that violate the regional political boundaries, reconfiguring them following the actual musical communicative practices. In details, by computing lexical profiles we obtain a coarse-grained representation of the superdiversity. A motivation can be found in the lyrics' style. Indeed, it is common that in a single can be lyrics found both positive and negative parts of the text. For example, in a romantic lyrics, the main part of the text could speak about the "positive" aspects of love, while the chorus could be focused on the love's painful part. This lyrics' characteristic could lead to flattening and to standardization of valence scores. Indeed, since the lemmas' valence scores spread among a network of words that co-occurs in texts, it is frequent that strong positive lemmas include in their networks strong negative ones, and vice-versa. On the other side, computing melodic profiles, we obtain fine-grained details.

Unfortunately, we cannot align our lexicons evaluation with one presented in [32]. The leading causes are *(a)* the lack of an already existing Italian musical tagged dataset and *(b)* the high costs required to manually tag a larger set of Italian songs. Due to this, on one side we cannot train a sentiment classifier based on Support Vector Machines (SVM), on the other side, also applying another sentiment analysis method, we cannot compare our results with others already evaluated. Finally, to build a tagged dataset to use as a training dataset was not taken into account the method applied in [13]. Indeed, this latter approach exploits ANEW as seed dictionary, so we believe that an SVM performances comparison with ANEW would be misleading.

5.1 Music Sentiment Analysis

Due to the lack of an already tagged and evaluated Italian musical dataset, to further evaluate the goodness of obtained lexicons, we compare the behavior of ANEW and each regional lexicon in classifying our untagged regional lyrics datasets. We choose to follow the method proposed in [13] because it is already attested to estimate the overall valence score for musical lyrics.



(a) Piemonte (b) Lombardia (c) Lazio

Fig. 11: Histograms of lyrics' valences in selected regional lyrics subsets using ANEW and the related regional lexicon.

For each regional subset in the `ITALY-lyric` dataset, we apply the proposed approach by using both ANEW and the respective regional lexicon obtained applying the sentiment spreading epidemic model (see Section 4.2).

$$v_{text} = \frac{\sum_{i=1}^{n} vi \cdot fi}{\sum_{i=1}^{n} fi} \tag{1}$$

where vi is the ANEW average valence for the lemma i.

Figure 11 shows lyrics' valences in selected regional lyrics subsets (Piemonte, Lazio, and Lombardia) using ANEW and the related regional lexicon. As can be noted, valences obtained with the related regional lexicon are aligned. In particular, histograms show a common behavior. ANEW tends to assign scores higher of 1 or 2 points compared to the regionals lexicons. We observe that regional lexicons tend to assign to lyrics values in the range [4,6], while ANEW in the range [6,8]. These behaviors are aligned with those observed by in [32], where ANEW tend to evaluate tweets as positive, while the obtained dictionary balance better negatives and positives classes, but tends on assign mean valences.

6 Conclusion

In this work, we have proposed a data-driven investigation of a music-specific domain, the Italian scene. We have built our dataset by exploiting heterogeneous online resources. Peculiarities of Italian music have been analyzed leveraging the geographical and emotional dimensions, through melodic and lexical features, to evaluate the Italian music *superdiversity*. Our results show that melodic and lexical features lead to coherent profiles. The applied sentiment spreading algorithm has allowed us to highlight both the lexical and emotive Italian music's characteristics. It can be argued that being two part of the same phenomenon, lexical and melodic features can be combined in favor of a better understanding of the analyzed Italian scene. Moreover, since there are no attested publicly annotated Italian music datasets, the free availability of our data represents important contribution to start to fulfill this lack.

As future research directions, due to the sparsity of contribution regarding the Italian music panorama, we would like to consolidate the cross-domain dataset to enrich the actual one. Furthermore, we are planning to identify the worldwide musical superdiversity through profiles of different nations.

Acknowledgements This work is supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement, http://www.sobigdata.eu, GS501100001809, 654024 "SoBigData: Social Mining & Big Data Ecosystem".

References

1. Echonest web api (2018). URL `http://docs.echonest.com.s3-website-us-east-1.amazonaws.com/`
2. Google form service (2018). URL `https://www.google.com/forms/about/`
3. Soundcloud web api (2018). URL `https://developers.soundcloud.com/docs/api/guide`
4. Spotify (2018). URL `https://www.spotify.com/`
5. Spotify web api (2018). URL `https://developer.spotify.com/web-api/`
6. Toscana100band contest (2018). URL `http://toscana100band.it/`
7. Wikipedia - ita version (2018). URL `https://it.wikipedia.org/wiki/Pagina_principale`
8. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification-a hybrid approach. In: ISMIR, pp. 657–662 (2009)
9. Bradley, M.M., Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Tech. rep., Citeseer (1999)
10. Çano, E., Morisio, M.: Moodylyrics: A sentiment annotated lyrics dataset. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, pp. 118–124. ACM (2017)
11. Çano, E., Morisio, M.: Music mood dataset creation based on last. fm tags (2017)
12. Celma, O.: Music recommendation. In: Music recommendation and discovery, pp. 43–85. Springer (2010)
13. Dodds, P.S., Danforth, C.M.: Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of happiness studies **11**(4), 441–456 (2010)
14. Downie, X., Laurier, C., Ehmann, M.: The 2007 mirex audio mood classification task: Lessons learned. In: Proc. 9th Int. Conf. Music Inf. Retrieval, pp. 462–467 (2008)
15. Esuli, A., Sebastiani, F.: Sentiwordnet: a high-coverage lexical resource for opinion mining. Evaluation pp. 1–26 (2007)
16. Guerini, M., Gatti, L., Turchi, M.: Sentiment analysis: How to derive prior polarities from sentiwordnet. arXiv preprint arXiv:1309.5843 (2013)
17. Helmholz, P., Siemon, D., Robra-Bissantz, S.: Summer hot, winter not!–seasonal influences on context-based music recommendations
18. Hu, X., Downie, J.S.: Exploring mood metadata: Relationships with genre, artist and usage metadata. In: ISMIR, pp. 67–72 (2007)
19. Hu, X., Downie, J.S.: When lyrics outperform audio for music mood classification: A feature analysis. In: ISMIR, pp. 619–624 (2010)
20. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. American music **183**(5,049), 2–209 (2009)
21. Lamere, P., Pampalk, E., Schmitz, C., Bello, J., Chew, E., Turnbull, D.: Social tags and music information retrieval. In: ISMIR, p. 24 (2008)
22. Laurier, C., Sordo, M., Serra, J., Herrera, P.: Music mood representations from social tags. In: ISMIR, pp. 381–386 (2009)
23. Lee, J.H., Hu, X.: Generating ground truth for music mood classification using mechanical turk. In: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, pp. 129–138. ACM (2012)

24. Li, T., Ogihara, M.: Music artist style identification by semi-supervised learning from both lyrics and content. In: Proceedings of the 12th annual ACM international conference on Multimedia, pp. 364–367. ACM (2004)
25. Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'Orletta, F., Dittmann, H., Lenci, A., Pirrelli, V.: The paisa'corpus of italian web texts. In: 9th Web as Corpus Workshop (WaC-9)@ EACL 2014, pp. 36–43. EACL (European chapter of the Association for Computational Linguistics) (2014)
26. Malheiro, R., Panda, R., Gomes, P., Paiva, R.P.: Classification and regression of music lyrics: Emotionally-significant features. 8th International Conference on Knowledge Discovery and Information Retrieval (2016)
27. Mihalcea, R., Strapparava, C.: Lyrics, music, and emotions. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 590–599. Association for Computational Linguistics (2012)
28. Perna, S., Guarasci, R., Maisto, A., Vitale, P.: Il linguaggio del rap. possibilità di un'analisi multidisciplinare. In: A. editrice (ed.) XXVI Convegno Internazionale Ass.I.Term. Terminologia e organizzazione della conoscenza nella conservazione della memoria digitale, vol. 34, p. 209–217. AIDAinformazioni, Rende (CS) (2016)
29. PODIUC, R.E., GRATIE, D., VOICU, O.: Inferring song moods from lyrics
30. Pollacci, L., Guidotti, R., Rossetti, G.: "are we playing like music-stars?" placing emerging artists on the italian music scene. In: 9th International Workshop on Machine Learning and Music (2016)
31. Pollacci, L., Guidotti, R., Rossetti, G., Giannotti, F., Pedreschi, D.: The fractal dimension of music: Geography, popularity and sentiment analysis. In: International Conference on Smart Objects and Technologies for Social Good, pp. 183–194. Springer (2017)
32. Pollacci, L., Sîrbu, A., Giannotti, F., Pedreschi, D., Lucchese, C., Muntean, C.I.: Sentiment spreading: An epidemic model for lexicon-based sentiment analysis on twitter. In: Conference of the Italian Association for Artificial Intelligence, pp. 114–127. Springer (2017)
33. Rawlings, D., Ciancarelli, V.: Music preference and the five-factor model of the neo personality inventory. Psychology of Music **25**(2), 120–132 (1997)
34. Rentfrow, P.J., Gosling, S.D.: The do re mi's of everyday life: The structure and personality correlates of music preferences. Journal of personality and social psychology **84**(6), 1236 (2003)
35. Russell, J.A.: A circumplex model of affect. Journal of personality and social psychology **39**(6), 1161 (1980)
36. Schedl, M., Orio, N., Liem, C., Peeters, G.: A professionally annotated and enriched multimodal data set on popular music. In: Proceedings of the 4th ACM Multimedia Systems Conference, pp. 78–83. ACM (2013)
37. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: In proceedings of the acl sigdat-workshop. Citeseer (1995)
38. Schmid, H.: Probabilistic part-ofispeech tagging using decision trees. In: New methods in language processing, p. 154 (2013)
39. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: ISMIR, vol. 8, pp. 325–330 (2008)
40. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech, and Language Processing **16**(2), 467–476 (2008)
41. Vertovec, S.: The emergence of super-diversity in Britain. Centre of Migration, Policy and Society, University of Oxford (2006)
42. Vertovec, S.: Super-diversity and its implications. Ethnic and racial studies **30**(6), 1024–1054 (2007)

Authors' Biographies

Laura Pollacci is a PhD Student at the Computer Science Department of the University of Pisa and member of the Knowledge Discovery and Data Mining Laboratory. She was born in 1988 in Viareggio (LU), Italy. She received the degree in 2014 and the master degree at the University of Pisa in Digital Humanities cum laude. Both theses have been developed in the Computational Linguistics area at Computational Linguistic Laboratory. Her research interests lie in the area of Social Network Analysis for Social Mining, Migration flows, Sentiment Analysis & Emotion.

Riccardo Guidotti was born in 1988 in Pitigliano (GR) Italy. In 2013 and 2010 he graduated cum laude in Computer Science (MS and BS) at University of Pisa. He received the PhD in Computer Science with a thesis on Personal Data Analytics in the same institution. He is currently a post-doc researcher at the Department of Computer Science University of Pisa, Italy and a member of the Knowledge Discovery and Data Mining Laboratory (KDDLab), a joint research group with the Information Science and Technology Institute of the National Research Council in Pisa. He won the IBM fellowship program and has been an intern in IBM Research Dublin, Ireland in 2015. His research interests are in personal data mining, clustering, explainable models, analysis of transactional data.

Giulio Rossetti earned his PhD in Computer Science at University of Pisa in 2015 with the thesis "Social Network Dynamics". He is a researcher at ISTI-CNR and a member of the KDD Lab (Knowledge Discovery and Data Mining Laboratory). His research interests involve the analysis of Big Data to study and model the heterogeneous aspects of human behaviours, with a special focus on: dynamic networks analysis, diffusion and epidemic spreading, data-driven model for the science of success.

Fosca Giannotti is a senior researcher at the Information Science and Technology Institute of the National Research Council at Pisa, Italy, where she leads the Knowledge Discovery and Data Mining Laboratory – KDD LAB – a joint research initiative with the University of Pisa, founded in 1995, one of the earliest European research groups specifically targeted at data mining and knowledge discovery. Her current research interests include data mining query languages, knowledge discovery support environment, web-mining,

spatio-temporal reasoning, spatio-temporal data mining, and privacy preserving data mining. She has been involved in several research projects both at national and international level, holding both management and research positions. She has been the coordinator of various European and national research projects and she is currently the co-ordinator of the FP6-IST project GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery. She is responsible for the Working Group on Privacy and Security in Data mining of the KDUBIQ network of excellences. She has taught classes on databases and data mining at universities in Italy and abroad. She is the author of more than one hundred publications and served in the scientific committee of various conferences in the area of Logic Programming, Databases, and Data Mining. In 2004 she co-chaired the European conference on Machine Learning and Knowledge Discovery in Data Bases ECML/PKDD 2004. She is the co-editor of the book "Mobility, Data Mining and Privacy", Springer, 2008.

Dino Pedreschi is a Professor of Computer Science at the University of Pisa, and a pioneering scientist in mobility data mining, social network mining and privacy-preserving data mining. He co-leads with Fosca Giannotti the Pisa KDD Lab - Knowledge Discovery and Data Mining Laboratory, a joint research initiative of the University of Pisa and the Information Science and Technology Institute of the Italian National Research Council, one of the earliest research lab centered on data mining. His research focus is on big data analytics and mining and their impact on society. He is a founder of the Business Informatics MSc program at Univ. Pisa, a course targeted at the education of interdisciplinary data scientists. Dino has been a visiting scientist at Barabasi Lab (Center for Complex Network Research) of Northeastern University, Boston (2009-2010), and earlier at the University of Texas at Austin (1989-90), at CWI Amsterdam (1993) and at UCLA (1995). In 2009, Dino received a Google Research Award for his research on privacy-preserving data mining. He is also Director of the Master in Big Data.