IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2019, accepted August 7, 2019, date of publication August 27, 2019, date of current version September 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2937743

The Conundrum of Success in Music: Playing it or Talking About it?

ALBERTO COSIMATO[1], ROBERTO DE PRISCO[1], ALFONSO GUARINO[1], DELFINA MALANDRINO[1], NICOLA LETTIERI[2], GIUSEPPE SORRENTINO[1], AND ROCCO ZACCAGNINO[1]

[1]Department of Computer Science, University of Salerno, 84084 Salerno, Italy
[2]National Institute for Public Policy Analysis (INAPP), 84084 Rome, Italy

Corresponding author: Rocco Zaccagnino (rzaccagnino@unisa.it)

ABSTRACT Nowadays social media are the main means for conducting discussions and sharing opinions. The huge amount of information generated by social media users is helpful for predicting outcomes of real-world events in different fields, including business, politics and the entertainment industry. In this paper, we studied the possibility of forecasting the success of music albums by analyzing heterogeneous data sources spanning from social media (Twitter, Instagram and Facebook) to mainstream American newspapers (e.g., New York Times, Rolling Stones). The idea is to exploit music albums' *pre-release hype* and *post-release approval* to predict the album's rank with reference to the well-known Billboard 200 album chart, which tabulates the weekly popularity of music albums in the USA. To predict the success of a music album, that is its rank in the chart, we identified metrics based on the messages' posting trend, the variation of the sentiment associated to such messages, the number of followers of the album's author, and the importance of the people who talk about it. To evaluate the effectiveness of the proposed metrics we have compared the prediction performances of several models based on supervised learning approaches among those most used in literature. As a result, we obtained that the Random Forest approach is able to predict the music album rank in the Billboard 200 Chart with an expected accuracy of 97%. As a further validation, using this specific model, we also conducted an additional real usage test obtaining an almost matching result (accuracy of 94%).

INDEX TERMS Social media, machine learning, prediction, sentiment analysis, music industry.

I. INTRODUCTION

Social media is a category of online channels, tools and applications where people create, share and bookmark content as well as collaborate around it at an unbelievable pace. The key factors are *co-creation* and *collaboration*. The 3.397 billion active social media users form a network in which they interact, co-create contents, and discuss about them. Thanks to its ease of use and popularity, social media is changing the way people evolve and share their opinions with major impacts in several areas such as politics and the entertainment industry. The digital traces that we leave on social media are like ''footprints'' that describe our behavior, both individually and as groups and, as argued, for example, in [1], allow to predict, to a certain extent, real-world events. Data that propagates through large user communities gives an interesting opportunity: an appropriate analysis of the data allows predictions about particular outcomes, as for example marketing and advertising campaigns [2].

In this paper, we consider the task of predicting the music album success in terms of its rank in the Billboard 200 chart, which evaluate the weekly popularity (including sales, streams and so on) of songs and music albums in the USA.

We have focused on music albums for two main reasons:

- The topic of music, including albums and artists, is highly discussed in the social media user community, and likewise a large amount of news about music is very frequently published in online newspapers.
- The album's success is measured by its rank in a chart. Among the various available chart, we have chosen the Billboard 200 chart, published by Billboard, one of the most followed music newspaper in the USA.

We design a machine learning-based prediction model that we call *Billboard 200 Predictor*, or *BB200P*.

The associate editor coordinating the review of this article and approving it for publication was Zhan Bu.

Machine learning enables a system to analyze data and deduce knowledge. The goal is to identify and exploit hidden patterns in a dataset. This provide several advantages respect to classic approaches and instigates a shift in the traditional programming paradigm, where programs are written to automate tasks. The success of machine learning techniques relies heavily on data. There is a colossal amount of data in todays' networks, which is bound to grow further with emerging social platforms. This encourages the application of machine learning that not only identifies hidden and unexpected patterns, but can also be applied to learn and understand the processes that generate the data.

The BB200P model proposed in this work needs to base its prediction on information about the music albums. So, the first question is: ''what are and where can we find information about music albums that allows to predict their success?''.

According to the current market logic of the music industry, music producers, singers/bands, and record labels spend a lot of effort and money in publicizing their albums, and have also embraced the social media for this purpose. As a consequence, the search for information that is tied to the success of an album starts with the analysis of heterogeneous data source: *(1)* Twitter chatter, *(2)* news from New York Times, Billboard, Entertaiment Weekly, Rolling Stones, and Variety, which are among the most followed online newspapers in USA, and *(3)* quantitative data from the most famous social media such as Instagram, Facebook, YouTube.

To extract the most significant features about the albums from such data, with respect to the chart rank, we identified a set of metrics based on two phenomena: *(1) pre-release hype* on social media and the role that attention plays in forecasting real-world music charts positioning, and *(2) post-release approval* within the opinion sharing and the influencing discourses. Such phenomena have been measured in terms of the number of the music album author's followers on the social media, tweet rate of tweets regarding the music album, and sentiment analysis of such tweets (and news about the album) performed by using the Python *Vader* library [3] in order to distinguish positively oriented tweets from negative ones.

To evaluate the effectiveness of the proposed metrics we have compared the prediction performances of three well-known models based on supervised learning: Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Random Forest (RF). We obtained that the Random Forest approach is the one that works better: it is able to predict the position of the music album in the Billboard 200 rank with an accuracy of 97%. Thus the BB200P is based on the Random Forest approach.

The main contributions of our work can be summarized as:

1) A proposal of a novel system to predict the music album success in terms of its rank in the Billboard 200 chart. We remark that to the best of our knowledge this represents one of the first attempts to predict Billboard rankings.
2) An *absolute* prediction, in the sense that it identifies the exact positioning of the album in one of the following chart rank range: from rank 1 to 10 (*Top10*), from rank 11 to 100 (*Top100*), from rank 101 to 200 (*Top200*), and most importantly, not available in Billboard's Chart. Furthermore, the prediction can be also provided on albums that are still in a pre-release phase.
3) An efficient system to make prediction of musicalbums; the computational complexity of our system, compared with other prediction systems proposed in the literature (see Section II), is more efficient, in the sense that with only 2 days of observation (the starting day and the current day), and without a large time window (e.g., 7 days), it is able to classify the Billboard 200 rank of an album coming out in the USA with high accuracy (see Section IV-B4).

This paper is organized as follows: in Section III we describe relevant works about the use of social media analysis for predicting real-world outcomes. Section IV provides the description of the BB200P model, while Section V describes a real-usage of the model. Finally, Section VI provides final considerations and directions for future works.

II. RELATED WORK

Several studies in literature exploit social media analysis to define new models for prediction of real-world outcomes.

For instance, political analysts have already turned to the social media as an indicator of political opinions. As an example, in [4] it is stated that Barack Obama's presidential campaign has heavily exploited social media such as Twitter, Facebook, MySpace, and others. In [5], the authors, in the context of the German federal election, have investigated whether Twitter is used as a forum for political deliberation; they conducted a content-analysis of over 100,000 messages containing references to political parties or to politicians. The conclusion drawn in [5], is that result of the election reflects the number of tweets mentioning a political party.

Other works focus on the impact of Twitter in the context of product marketing. In [2] the authors have found that 19% of a random sample of tweets contained mentions of a brand or product and that a classification system was able to extract statistically significant differences of customer sentiment, as for example the attitude of a writer towards a brand.

In [6] a system able to detect heartquakes based on the analysis of tweets was described. They elaborate their detection system further to detect rainbows in the sky, and traffic jams in cities [7]. Their work consists of an alerting system which could perform so promptly that the alert message could arrive faster than the earthquake waves to certain regions.

Other studies have also faced with the use of social media indicators in other contexts, to have daily predictions on up and down changes in Dow Jones Industrial Average values [8] or to predict the scientific impact of research articles [9]–[11].

One of the works most similar to that described in this paper regards the prediction of box-office outcomes by means of a linear regression method [12]. The authors exploited

the tweet rate and sentiment analysis, by analyzing the rate for 7 days antecedents the movie release for each of the 24 movies chosen and the amount of tehaters where the movie was released, achieving an adjusted R^2 value of 0,973.

A more comprehensive overview of predictive methods exploiting social media analysis can be found in [13]. In this work, the authors have criticised the predictive capabilities of some proposed models adopting specific filtering or classifications based on human assessors, thus reducing the replicability of the solution. We remark that our work, instead, does not rely on humans but is completely automatic.

III. BACKGROUND

Several machine learning models have been used for classification or prediction problems. In this section we briefly describe the ones that we use in this paper. For further details about machine learning we refer the reader to [14]. The models we use are: Random Forest (RF), Support Vector Machines (SVM), and MultiLayer Perceptron (MLP).

- *Random Forest (RF):* is a supervised classification algorithm which consists of an ensemble of methods based on bagging [15]; we used the *scikit-learn* implementation which combines trees by averaging their probabilistic prediction instead of letting each tree vote for a single class, and inherently support multi-class problems.
- *Support Vector Machines (SVM):* is a supervised learning model with associated learning algorithms [16]; an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate classes are divided by a clear gap; new examples are predicted to belong to a class based on which side of the gap they fall. SVM in scikit-learn implements the *one-against-one* approach [17] for multi-class classification.
- *MultiLayer Perceptron (MLP):* is a feedforward artificial neural network [18] which exploits a supervised learning technique called backpropagation [18] for training; MLP supports multi-class classification by applying Softmax [19] as the output function.

There exist several metrics for evaluating scores of machine learning models. In this work we used the following:

- *accuracy:* informally, accuracy is the fraction of predictions our model got right; formally it is defined as

$$accuracy = \frac{(tp + tn)}{(tp + tn + fp + fn)}$$

where tp, fn, fp, and tn are the number of true positives, false negatives, false positives and true negatives, respectively.
- *precision:* intuitively, precision is the ability of the classifier not to label as positive a sample that is negative; formally it is defined as

$$precision = \frac{tp}{(tp + fp)}$$

where tp is the number of true positives and fp the number of false positives; the precision is intuitively the ability of the classifier not to label as positive a sample that is negative.
- *recall:* intuitively, recall is the ability of the classifier to find all the positive samples; formally it is defined as

$$recall = \frac{tp}{(tp + fn)}$$

where tp is the number of true positives and fn the number of false negatives.

TABLE 1. History of album success from 1979 to 2009. The table show a subset of most famous albums of every decade.

Artist	Title	Year	Release date	Entrance in top 10	days
Neil Young	Rust Never Sleeps	1979	1979-6-22	1979-8-25	64
Pink Floyd	The Wall	1979	1979-11-30	1980-1-5	36
Supetramp	Breakfast in America	1979	1979-3-29	1979-4-28	30
John Mellencamp	Big Daddy	1989	1989-5-9	1989-6-10	32
Madonna	Like a Prayer	1989	1989-3-21	1989-4-15	27
Rolling Stones	Steel Wheels	1989	1989-8-28	1989-9-30	33
Backstreet Boys	Millenium	1999	1999-5-18	1999-6-5	18
Britney Spears	Baby One More Time	1999	1999-1-12	1999-1-30	18
Eminem	The real slim shady LP	1999	1999-2-23	1999-3-13	18
Jay-Z	The Blueprint 3	2009	2009-9-8	2009-9-26	18
U2	No Line on the horizon	2009	2009-3-3	2009-3-21	18
Chris Brown	Graffiti	2009	2009-12-7	2009-12-21	14

IV. OUR APPROACH

In this section we provide details about the preliminary study we performed to analyze the trend, in the last 40 years, of the time needed for albums to reach the top of the most famous charts. Then we describe the steps followed to build a new model for the prediction of music albums' chart rank.

A. PRELIMINARY STUDY

To start with, we decided to analyze past data in the Billboard 200 chart to see how much time successful albums took to reach the top 10 positions. We looked at albums in the following specific years: 1979, 1989, 1999, 2009. In Table 1 we report some of these albums, and as we can see, very famous artists like Pink Floyd (winner of the USA best seller album in 1980),[1] despite their popularity in 70-80s, took more time than U2 and Chris Brown in 2009 to get in the Top 10 ranks of Billboard 200 Chart. Specifically, taking account of proper exceptions, while in 1979 famous artists needed an average of 43 days to get in the Top 10 ranks of Billboard 200 chart, in 2009 the average positioning time was 20 days.

B. MACHINE LEARNING METHOD

As explained before, we study how data coming from social media can influence the USA music discographic market. We focus on the success of music albums in terms of rank on the Billboard 200 chart. We consider data coming from heterogeneous sources such as Twitter data, news from Billboard, Entertainment Weekly, Rolling Stones, and Variety,

[1] https://bit.ly/2Srd0SH

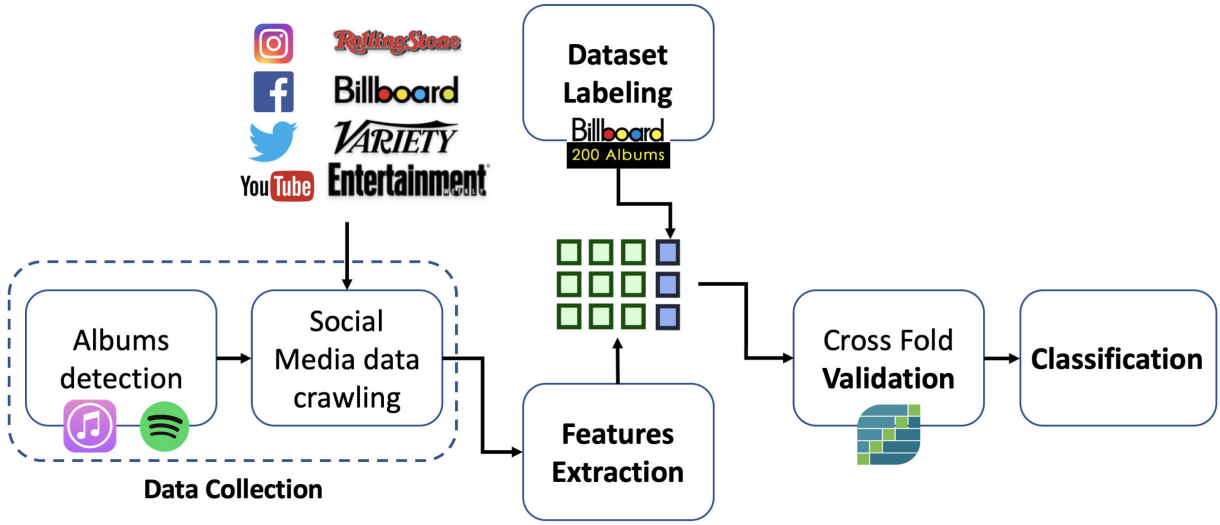

FIGURE 1. The main steps of the proposed approach. *(1) Data collection:* Information from the most important social media have been crawled to build the raw dataset; *(2) Features extraction:* A set of features was engineered from the data collection in order to describe the input sample; *(3) Dataset labeling:* The features have been labeled to build the dataset used by machine learning models; *(4) Cross fold validation:* We perform a 5-fold cross-validation in order to obtain the best parameters for each classifiers considered; *(5) Classification:* We used the best parameters obtained in the previous step to train and test several machine learning classifiers. Details of each step can be found in the respective sections (from Section IV-B1 to Section IV-B4).

New York Times, quantitative data from Instagram, Facebook, YouTube. To achieve it, a methodology based on the guidelines provided in [20] has been defined (see Figure 1 for an overview of the adopted methodology).

Algorithm 1 shows the pseudocode for BB200P. The algorithm takes as input a set of albums *Album*, the set of considered social media *Social* and the set of considered newspaper *News*. It returns the accuracy obtained.

Algorithm 1 BB200P pseudocode

Input : *Album*, *Social*, *News*, *K fold*.
Output: *accuracy*.
1 *rawDataset* ← *DataCollection*(*Album*, *Social*, *News*);
2 *samples* ← *FeaturesExtraction*(*rawDataset*);
3 *MLDataset* ← *DatasetLabeling*(*Album*, *samples*);
4 (*TrainingSet*, *TestingSet*) ← *Split*(*MLDataset*);
5 *MLParameters* ← *CrossValidation*(*TrainingSet*, *K fold*);
6 *accuracy* ← *Classification*(*TestingSet*, *MLParameters*);
7 return *accuracy*;

As we can see in Algorithm 1, our methodology consists of the following steps:

1) *data collection*: a set of forthcoming music albums has been identified; for each album we have collected heterogeneous data related to the album itself and the albums' authors (Section IV-B1).
2) *features extraction*: a set of (virtually) suitable features has been defined and engineered from the data collected, in order to build the dataset (Section IV-B2);
3) *dataset labeling*: the dataset was labeled according to the Billboard 200 chart (Section IV-B2);
4) *validation*: the dataset was splitted into a *training set* and a *testing set*; a *k*-fold cross-validation was performed on the training set to validate different machine learning models (Section IV-B3);
5) *classification*: most used prediction models in literature have been tested on the test set with the best parameters found during the previous step (Section IV-B4).

For the implementation of the BB200P model we have used the Python *scikit-learn* library.

1) DATA COLLECTION

In order to build the dataset, we gathered information about all incoming music albums in the period ranging from 30 Nov 2018 to 31 Dec 2018, from iTunes and Spotify. As a result, we have considered 86 music albums and all information about such albums are related to this period of time. Several types of information have been crawled from the most important social media, and in particular *quantitative* information about the degree of participation of the album's author on the Social themselves. Specifically:

- *Twitter*: for each album, we have considered *(1)* the set of tweets related to the album or the album's author, *(2)* the number of fans, *(3)* the number of favorites and *(4)* the tweets of the album's author (information about the degree of participation of the album's author on Twitter); to obtain such information we built a search query list in order to crawl only Twitter data regarding the selected album; the data have been downloaded by means of Twitter API Standard. We downloaded 11,2 GB of tweets and 50,2 GB of retweets. To collect

TABLE 2. An example of query performed on twitter for some music album considered.

Artist	Album Title	Query
The 1975	"A Brief Inquiry Into Online Relationships"	*The1975* *The 1975* *A Brief Inquiry Into Online Relationships*
Alessia Cara	"The Pains of Growing"	*Alessia Cara* *The Pains of Growing* *AlessiaCara*
Clean Bandit	"What is Love"	*Clean Bandit* *cleanbandit* *What is Love album*
21 Savage	"I Am >I Was"	*21Savage* *Shayaa Bin Abraham-Joseph* *iamiwas* *I Am I Was album*
Coldplay	"The Butterfly Package (Live in Buenos Aires / Live in Sao Paulo / A Head Full of Dreams)"	*coldplay* *theButterflyPackage* *AHFODFilm* *AHeadFullofDreams*
ZAYN	"Icarus Falls"	*ZAYN* *zaynmalik* *ICARUSFALLS* *Icarus Fall Album*

the data, as a general rule, we used keywords containing the album title, the name of the album's author and the hashtags associated to the album. For instance, given the album "The Butterfly Package" by Coldplay we used search queries with the following keywords: *coldplay*, *theButterflyPackage*, *AHFODFilm*, and *AHeadFullofDreams* (see Table 2).

- *Instagram, Facebook*: we considered the number number of fans of the album's author on Instagram and Facebook (degree of participation of the author on Instagram and Facebook); Data have been downloaded from SocialBlade through automatic scraping with Selenium[2];
- *YouTube*: we considered the number of fans and the total videos' views of the album's author on YouTube; Data have been downloaded from SocialBlade through automatic scraping with Selenium;
- *Newspapers*: the news on the most read magazines could influence people's behavior [21]; thus, we considered the news from Billboard, Entertaiment Weekly, New York Times, Rolling Stones and Variety; we remark that Billboard, Entertainment Weekly, Rolling Stones, Variety are in the top 4 of the most-read online music magazines in the USA,[3] while New York Times is the most read news website in the USA;

[2]https://socialblade.com/
[3]https://bit.ly/2Srd0SH

2) DATASET: FEATURES AND LABELS

In this section we provide details about the features engineered from the data collection (raw data) described in Section IV-B1 and we explain how samples have been labeled.

The choice of the set of features is a crucial step during the definition of a machine learning system because it must describe the input sample. In our case, given a music album, the set of features has to describe the information about the album itself on the chosen social media in a specific time t belonging to the considered time window. Specifically, the set of features has to represent: *(1)* the number of fans on the social media (i.e., the reached people), *(2)* how much the album's author is discussed on the net ("are people interested in this album?"), and *(3)* the collective opinion about the album or the album's author on social media ("is it well-talked?"). In addition, we considered the fact that a famous music artist could influence the success of a new artist by appearing in one or more of his songs. So we decided to include also the *featurings*[4] for each music album considered. The following features (Section IV-B2.b) are then calculated for every day in a time window of 21 days. Specifically, we considered 7 days before the album's release, for the analysis of *pre-release hype* phenomenon, and 14 days after the album's release, for the *post-release approval* phenomenon analysis.

a: USEFUL DEFINITIONS

In this section we provide some useful notions used for the formal definition of the engineered features.

We have considered the following sets of entities: *(1) Album*, the set of the albums (86 in total) considered; *(2) Author*, the set of the authors of the albums, *(3) Featuring*, the set of the "featured" artists, *(4) Artist*, the set of all artists (authors and featured), and finally *(5) Writer*, the set of users which have written a tweet. Observe that the sets are defined in such a way that $Author, Featuring \subseteq Artist$.

Let $a \in Artist$ be a music artist, with Y_a we indicate the set of *YouTube channels* of a. Formally, $Y_a = \{y | y$ is a YouTube channel of $a \in Artist\}$.

Let $a \in Author$ be an album's author, we define:

- $N_a = \{n | n$ is a news about $a \in Author\}$, i.e., the set of news about a;
- $T_a = \{t | t$ is a downloaded tweet about $a \in Author\}$, i.e., the set of downloaded tweets about a;
- $R_a = \{r | r$ is a retweet about $a \in Author\}$, i.e., the set of retweets about a.

We remark that to define the sets T_a and R_a we have used *search queries*, as exemplified in Table 2 that shows the queries performed on Twitter for some of the music albums. As we have seen, in addition to the tweets and retweets

[4]In the context of the production of music albums, a "featuring" is the participation of a guest artist in a music album. The guest artist is called the "featured" artist.

regarding the author, these sets include also the tweets and retweets regarding the album.

We will denote Twitter, Instagram and Facebook, respectively, as *tw*, *ig* and *fb*. Let $sn \in \{tw, ig, fb\}$ be a social media, let $s \in Artist$ be an artist, then with $F_{s,sn}$ we denote the set of followers of s on sn, and with $V_{s,sn}$ we indicate whether s has a *verified* account on sn or not. To give more importance to verified accounts, we set $V_{s,sn} = 1$ if s has a verified account on sn, $V_{s,sn} = 0.5$ otherwise.

Let $w \in Writer$ be a tweet writer, we indicate with TR_w the set of tweets and retweets written by w. We remark that we only considered writers that have written at least one tweet or one retweet about either an author or an album; formally we have considered only writers w such that $TR_w \cap (\bigcup_{a \in Author} T_a \cup R_a) \neq \emptyset$. Furthermore, F_w denotes the set of followers of w and we set $V_w = 1$ if w has a verified account, $V_w = 0.5$ otherwise. Also, given a tweet $t \in TR_w$, with R_t we indicate the number of retweets for t.

We denote with $S = \{-, 0, +, C\}$, the set of sentiments where $-$, 0, $+$ and C stand for, respectively, negative, neutral, positive and compound. We have used the Python *Vader* library to perform sentiment analysis.

Let $a \in Author$ be an author, let $n \in N_a$ be a news about a, and $z \in S$ be a sentiment, then $S_{n,z}$ represents the measure of sentiment z in the new n. Furthermore, with $S_{t,z}$ we indicate the measure of the sentiment z for the tweet(s) t. Let us observe that $S_{n,z}, S_{t,z} \in [0, 1]$.

Let $a \in Author$ be an author and $y \in Y_a$ be a YouTube channel of a, then $subs_y$ indicates the number of subscribers at y and V_y indicates the total number of visualizations for y. Then, the total number of subscribers for a is calculated as $Tsubs_a = \sum_{y \in Y_a} subs_y$, while the total visualization for a is calculated as $TV_a = \sum_{y \in Y_a} V_y$.

b: FEATURES

Now, by using the definitions provided before, we give the list of engineered features (see Table 3). Let $d \in Album$ be an album and $a \in Author$ be the author of d, we distinguish between *quantitative* features and *sentiment* features. As we can see in Table 3: the number of post published by a ($\#postInstagram_a$), the number of tweets and reweets about a ($\#tweets_a$), the total number of subscribers for a ($YouTubeFans_a$), the number of Twitter fans of a weighted according to the number of verified account ($TwitterFans_a$), the number of Facebook fans of a weighted according to the number of verified account ($FacebookFans_a$), the number of Instagram fans a weighted according to the number of verified account ($InstagramFans_a$), the average number of Twitter fans of featuring artists on d weighted according to the number of verified account ($Feats_d^{tw}$), the average number of Facebook fans of featuring artists on d weighted according to the number of verified account ($Feats_d^{fb}$), the average number of Instagram fans of featuring artists on d weighted according to the number of verified account ($Feats_d^{ig}$), the importance of the writers of tweets regarding a ($Importance_{tweet}$), the average number of retweets about a ($Retweets^{tw}$), the importance of the writers of retweets regarding a ($Importance_{retweet}$), the number of news about a ($NewsRate$) and the number of tweets about a ($TweetRate$).

TABLE 3. Quantitative and sentiment features.

Feature Class	Formula (Meaning)
Quantitative	
$\#postInstagram_a$	number of post published by a on Instagram
$\#tweets_a$	TR_a
$YouTubeFans_a$	$Tsubs_a$
$YouTubeViews_a$	TV_a
$TwitterFans_a$	$F_{a,tw} * V_{a,tw}$
$FacebookFans_a$	$F_{a,fb} * V_{a,tw}$
$InstagramFans_a$	$F_{a,ig} * V_{a,ig}$
$Feats_d^{tw}$	$\frac{\sum_{x \in Featuring_d} F_{x,tw} * V_{x,tw}}{\lvert Featuring_d \rvert}$
$Feats_d^{fb}$	$\frac{\sum_{x \in Featuring_d} F_{x,fb} * V_{x,fb}}{\lvert Featuring_d \rvert}$
$Feats_d^{ig}$	$\frac{\sum_{x \in Featuring_d} F_{x,ig} * V_{x,ig}}{\lvert Featuring_d \rvert}$
$Feats_d^{YouTube}$	$\frac{\sum_{x \in Featuring_d} Tsubs^x}{\lvert Featuring_d \rvert}$
$Importance_{tweet}$	$\frac{\sum_{t \in T_a} F_t * V_t}{\lvert T_a \rvert}$
$Retweets^{tw}$	$\frac{\sum_{t \in T_a} R_t}{\lvert T_a \rvert}$
$Importance_{retweet}$	$\frac{\sum_{rt \in R_a} F_{rt} * V_{rt}}{\lvert R_a \rvert}$
$NewsRate$	$\lvert N_a \rvert$
$TweetRate$	$\lvert T_a \rvert$
Sentiment	
$NSen_a^-$	$\frac{\sum_{x \in N_a} S_{x,-}}{\lvert N_a \rvert}$
$NSen_a^0$	$\frac{\sum_{x \in N_a} S_{x,0}}{\lvert N_a \rvert}$
$NSen_a^+$	$\frac{\sum_{x \in N_a} S_{x,+}}{\lvert N_a \rvert}$
$NSen_a^C$	$\frac{\sum_{x \in N_a} S_{x,C}}{\lvert N_a \rvert}$
Sen_{tw}^-	$\frac{\sum_{t \in T_a} S_{t,-}}{\lvert T_a \rvert}$
Sen_{tw}^0	$\frac{\sum_{t \in T_a} S_{t,0}}{\lvert T_a \rvert}$
Sen_{tw}^+	$\frac{\sum_{t \in T_a} S_{t,+}}{\lvert T_a \rvert}$
Sen_{tw}^C	$\frac{\sum_{t \in T_a} S_{t,C}}{\lvert T_a \rvert}$

Regarding the sentiment features, as explained before, for each tweet and for each news, the sentiment analysis was performed by using the Vader library for Python [3], a simple rule-based model for general sentiment analysis, which improves the accuracy of the sentiment analysis across several domain contexts (social media text, NY Times editorials, movie reviews, and product reviews). The objective of the sentiment analysis was to study how sentiments are created, how positive, neutral and negative opinions propagate and how they influence people. In our case, the set of sentiment features includes: the average negative sentiment score on news about a ($NSen_a^-$), the average neutral sentiment score on news about a ($NSen_a^0$), the average positive sentiment score on news about a ($NSen_a^+$), the average compound sentiment score on news about a ($NSen_a^C$), the average negative sentiment score on tweets about a (Sen_{tw}^-), the average neutral sentiment score on tweets about a ($NSen_{tw}^0$), the average positive sentiment score on tweets about a ($NSen_{tw}^+$) and the average compound sentiment score on tweets about a ($NSen_{tw}^C$).

Observe that all the features have a dynamic behavior that changes day by day. For this reason we also engineered for

TABLE 4. The labeling strategy and labeled music albums. For example an album ranked as 1st is labelled with 0. Label 3 is for albums that had not ranked in billboard 200, i.e. the least successful ones.

Label	Chart rank range	No. Albums
0	Top10 - from rank 1 to 10	10
1	Top100 - from rank 11 to 100	9
2	Top200 - from rank 101 to 200	5
3	Not found in Billboard's Chart	62

each feature two additional features, i.e., the first one calculated by means of *Finite Difference (FD)* and the second one calculated using the *Rate of Change (ROC)*. Both are methods to represent a value shaped as a difference between two function values. In our case the difference has been calculated compared to the first day of the analysis (7 days before the album release date). Our interest is towards the change in both absloute and relative terms. Formally, let f_0 be the value of a feature f in the first day, i.e., day 0, of the analysis, f_i is the feature's value in the i^{th} day. Then $f_{ROC,i} = \frac{f_i - f_0}{f_0} * 100$ is the feature represented as ROC, and $f_{FD,i} = f_i - f_0$ is the feature represented as FD. An example could be an author a with the following values of f: for day 0, $f_0 = 1000$ and for day 1 $f_1 = 1100$. Then the FD feature is $f_{FD,i} = 100$ while the ROC feature is $f_{ROC,i} = 10$. For another author $a' \neq a$ the $f_0 = 1000000$ and $f_1 = 1000100$ we obtain $f'_{FD,i} = f_{FD,i} = 100$ but, instead, $f_{ROC,i} = 0.01$.

c: LABELING

The labeling has been performed by crawling the Billboard's top 200 chart for exactly 14 days after the album release date. Chart rankings and the number of collected music albums in our dataset for each identified label are shown in Table 4.

d: DATASET

By using the features described in Section IV-B2.b we built the dataset, referred as *CollectedData*. Each sample in *CollectedData* represent one day of analysis and it is built by concatenating *quantitative* features and *sentiment* features, with *FD* features and *ROC* features. Figure 2 shows the timeseries trend in the number of tweets for a subset of music albums over the critical period. We can observe that the time of greatest interest for a music album is around the time of its release, and in the following days the tweet rate invariably fades. As we can see, for "Meek Mill" and his album titled "Championships" the trend shows a renewing interest after the album release with two spikes in the following two weeks.

About the sentiment analysis on Twitter, our expectation is that there would be stronger sentiments after the music album has released, than before. We expect tweets prior to the release to be mostly anticipatory, and stronger positive/negative tweets to be disseminated later following the release. As we can see in Figure 2 and Figure 3 there exists a significant difference between "Meek Mill" and "Zayn". The former went 1st place in Billboard 200 chart (label 0), the latter after the top 10 (label 1).

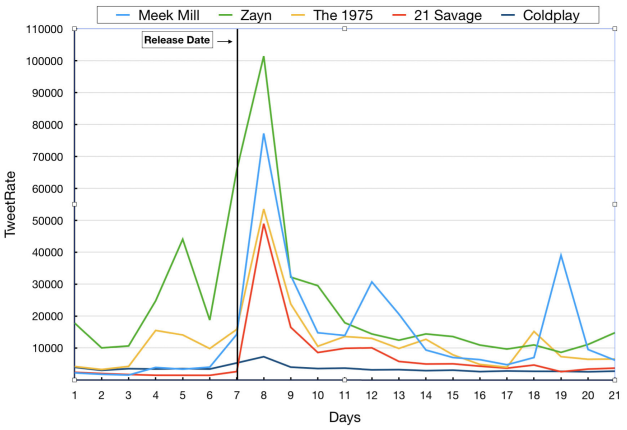

FIGURE 2. Time-series of tweet rate over the critical period for different music albums (shown with different colors). Hype increases in the pre-release period and the attention reaches the maximum one day after the album's release.

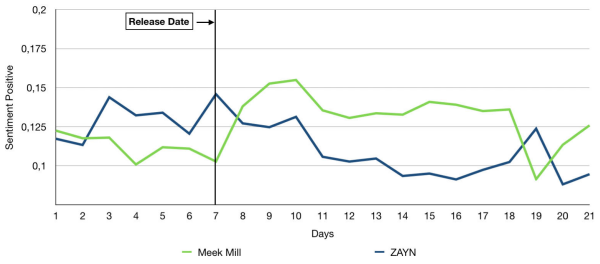

FIGURE 3. Time-series of tweets sentiment positive over the critical period for two music albums of interest. On y-axis, the average positive score returned by vader library. Tweeets related to ZAYN and his album (ranked top 100 in the billboard 200 chart) lose positive score after the album's release, while the ones related to meek mill (ranked top 10 in the billboard 200 chart) show an opposite trend.

In order to individuate the the most relevant features we also performed a feature importance task with the Random Forest classifier (the task involved the Gini Impurity). We found that the tweet rate, the negative sentiment of a tweet and the variation in percentage of retweet rate are the most important features. This confirms that: *(a)* even a minimal shift of the sentiment could result into the success or "failure" of a music album (or a higher/lower chart rank), and *(b)* the tweet rate ("how much is a album/author talked on Twitter?") is the forerunner of success in the music field.

3) VALIDATION

First we split *CollectedData* into: *(1)* the training set, obtained by including the 80% of the elements (randomly chosen), and *(2)* the testing set, obtained including the remaining 20% of the elements. Initially, we found that the *CollectedData* dataset has some unbalance between classes 0,1,2, and 3 (see Table 4). So, the *RandomOverSampler* technique is used to balance the set. Then, we scaled the data (for both testing and training set) by using the *MinMaxScaler* technique. Finally, to obtain the best parameters for each classifiers considered, and in order to validate machine learning models, we perform a 5-fold cross-validation by using the *GridSearchCV* method. We recall that *k-fold*

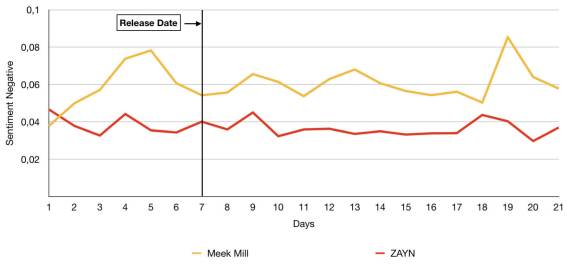

FIGURE 4. Time-series of tweets sentiment negative over the critical period for two music albums of interest. On y-axis, the average negative score returned by vader library. Tweeets related to ZAYN and his album (ranked top 100 in the billboard 200 chart) show an almost flat trend over the critical period, while the ones related to meek mill (ranked top 10 in the billboard 200 chart) show different spikes due to the lively online discussion.

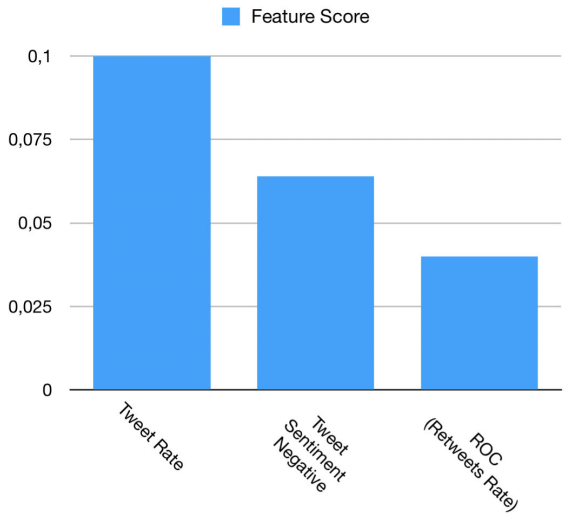

FIGURE 5. Importance of first three features ranked by a random forest classifier.

TABLE 5. *5*-fold cross-validation results for *CollectedData*.

Classifier	Avg. Accuracy	Variance
MLP	0,887	0,02
SVM	0,996	0,005
RF	0,997	0,003

cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into: in our case, due to the size of the dataset, we set $k = 5$. Cross-validation is primarily used to estimate the skill of a machine learning model on unseen data. As clearly explained in [22] *"this approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds"*. Best results of performed 5-fold cross-validation can be found in Table 5.

4) CLASSIFICATION

At the end of the validation step (see Section IV-B3) we have obtained the best parameters to train and test (on the testing set) our classifiers. Results of classification tests are shown in Table 6. We remark that our goal was to develop a forecast model that had high precision and recall scores for all classes and consequently high accuracy. As we can see, SVM in this case is a very bad classifier. Although the percentage of relevant retrieved samples is high (i.e., precision 1,00), with such low recall the classifier loses most of the samples. For example, SVM is not able to find 93% of the *Top10* albums and 95% of the of the *Top100* albums. This behavior is not desiderable when building a forecast model. This behavior is not desiderable when building a forecast model. As result, the best classifier is Random Forest that reaches 97% of accuracy in forecasting music albums chart positioning using *CollectedData*. This means that, by exploiting only social media-based information and by tracking them for very few days, it is possible to predict with 97% of accuracy the Billboard 200 chart rank of an incoming music album in USA. Moreover the Random Forest is the classifier that shows highest Precision and Recall measures for each of the four classes. This because the unbalanced dataset still represent a problem for the other classifiers.

TABLE 6. *CollectedData* result in performing one-shot test. MLP = Multilayer Perceptron, SVM = Support Vector Machine, RF = Random Forest. Label *No* is for album not found in billboard's chart. *Acc.* is abbreviation for accuracy score.

Classifier	Acc.	Precision				Recall			
		Top10	*Top100*	*Top200*	*No*	*Top10*	*Top100*	*Top200*	*No*
MLP	0,86	0,77	0,6	0,67	0,93	0,79	0,55	0,76	0,93
SVM	0,75	1	1	1	0,74	0,07	0,05	0,24	1
RF	**0.97**	**1**	**1**	**0,95**	**0,97**	**0,93**	**0,87**	**0,90**	**1**

C. ARTIST'S MUSICAL REPUTATION

The BB200P model introduced in this paper, does not take into consideration the *reputation* of the artists, which represent an aspect that could somehow influence the sales of the albums and therefore the ranking of the Billboard 200. As an example, one should notice that albums from famous artists, like Michael Jackson or Rolling Stones, always float in stable manner in the top part of the chart. To assess the prediction potential of the artist's musical reputation, we used both a model that uses only the artist's reputation to predict the rank and an augmented version of the BB200P model. For the easy of explanation, let's dub these two variants BB200P-Artist and BB200P-Augmented.

For BB200P-Artist we study how considering only information about the artist's reputation can affect an album's success. In other words, we built a model which is able to predict Billboard 200 chart rank of a newly released album in USA by only using reputation-based information. We collected 1846 of the albums released between 2017 and 2018, and for each album we calculated the album's artist reputation until its release. The reputation is calculated by using both Billboard 200 chart and Spotify Chart USA. We selected the average rank from both and from Spotify Chart USA the number of streamings, while from Billboard the number

TABLE 7. List of music album (released from 8 March 2019 to 15 March 2019) considered to test our model.

No.	Author	Album title
1	Benjamin Francis Leftwich	Gratitude
2	Chief Keef & Zaytoven	GloToven
3	Dan Sultan	Aviary Takes
4	Elizabeth Colour Wheel	Nocebo
5	Karen O & Danger Mouse	Lux Prima
6	Jack Savoretti	Singing to Strangers
7	Joanne Shaw Taylor	Reckless Heart
8	Matmos	Plastic Anniversary
9	Rich the Kid	The World is Yours
10	Snarky Puppy	Immigrance
11	Stephen Malkmus & The Jicks	Groove Denied
12	Steve Adamyk Band	Paradise
13	The Brian Jonestown Massacre	The Brian Jonestown Massacre
14	The Cinematic Orchestra	To Believe
15	The Comet is Coming	Trust in the Lifeforce of the Deep Mystery
16	The Faint	Egowerk
17	Tim O'Brien Band	Tim O'Brien Band
18	Todd Snider	Cash Cabin Sessions, Vol. 3
19	Turning Jewels Into Water	Map of Absences

of times in which the album's artist was ranked in. The experiment followed the same steps described in Section IV, therefore *(1)* labeling through Billboard 200 Chart, *(2)* k-fold cross-validation, and then *(3)* testing classifiers. The Random Forest model reaches an average accuracy of 82% in the prediction of Billboard 200 chart rank with this data.

For BB200P-Augmented we repeated the tests performed with BB200P adding the additional features regarding the artist's reputation. The problem with such approach was that behavior of overfitting occurred for each tested classifiers.

Thus, the conclusion is that the relationship between reputation and success is not very helpful in predicting the rank of the forthcoming album. This might seem surprising, but it is not: indeed the artist's reputation is a static information in the realtive short interval of time considered while in the same interval social media informations were generated at high pace.

V. REAL-USAGE TEST

As a "road test" we decided to use the B200P system to check its prediction accuracy. So, we used the system from March 8, 2019 through March 15, 2019. We collected information for 19 music albums (7 days before the album release and 14 days after, as explained in Section IV). The albums considered are shown in Table 7.

We found that in this road test the BB200P system has achieved a accuracy of 94%, which is close to the expected 97%. We remark that in this specific test, for the album titled *GloToven* by *Chief Keef & Zaytoven* we have not been able to extract all the need features, because the BB200P sources did not have enough information.

VI. CONCLUSION

This paper presented an effective and efficient model to predict the Billboard 200 Chart rank of a forthcoming music album by exploiting (heterogeneous) data coming from social media. Specifically, by analyzing the Twitter chatter (sentiment, tweet rate, and so on), by exploiting quantitative data from Facebook, Instagram and YouTube, and by including sentiment of news published in the main online newspapers and magazines, we built a prediction model, based on the Random Forest approach, which predicts the chart rank of music albums with an accuracy of 97%. The model performed very well also with data not belonging to the original dataset collected during the Christmas period, achieving 94% of accuracy. Furthermore, we showed that that most significant features come from Twitter, i.e., the tweet rate, the negative sentiment of tweets and the Rate Of Change (ROC) of the retweet rate are the most important features. Thus, Twitter is the most effective social media followed by Instagram and Youtube.

Future steps of the project include the using of:

- more accurate ML techniques to perform sentiment analysis, for example for the recognition of a 5 or 7-scale sentiments (not only positive, neutral, negative), trained on dataset containing tweets about music [23],
- a multi-view learning model [24]–[26] to improve our classifiers' performance; the idea is that of using a Random Forest on the single view regarding reputation data, and another Random Forest on the social data single view; then experiments will be performed to using intermediate and late integration techniques in order to assess which one achieve best prediction results,
- premium APIs of every social media accounted in order to include more information such as Facebook posts, comments, reactions (*like*, *love*, *laughing*, and so on), and Instagram stories.

REFERENCES

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann, "Computational social science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[2] B. J. Jansen, M. Zhang, K. Sobe, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, Nov. 2009.

[3] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Social Media*, 2014, pp. 1–16.

[4] K. Skemp, "All a-twitter about the massachusetts senate primary," *Retrieved December*, vol. 15, pp. 1–2, Dec. 2009.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, vol. 10, no. 1, pp. 178–185, 2010.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 851–860.

[7] M. Okazaki and Y. Matsuo, "Semantic twitter: Analyzing tweets for real-time event notification," in *Recent Trends and Developments in Social Software*. Berlin, Germany: Springer, 2010, pp. 63–74.

[8] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[9] T. Brody, S. Harnad, and L. Carr, "Earlier Web usage statistics as predictors of later citation impact," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 8, pp. 1060–1072, Jun. 2006.

[10] G. Eysenbach, "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact," *J. Med. Internet Res.*, vol. 13, no. 4, p. e123, 2011.

[11] X. Shuai, A. Pepe, and J. Bollen, "How the scientific community reacts to newly submitted preprints: Article downloads, twitter mentions, and citations," *PloS One*, vol. 7, no. 11, 2012, Art. no. e47523.
[12] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2010, pp. 492–499.
[13] L. Madlberger and A. Almansour, "Predictions based on Twitter—A critical view on the research process," in *Proc. Int. Conf. Data Softw. Eng. (ICODSE)*, Nov. 2014, pp. 1–6.
[14] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. Newton, MA, USA: O'Reilly Media, 2017.
[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[16] L. Wang, *Support Vector Machines: Theory and Applications*, vol. 177. New York, NY, USA: Springer, 2005.
[17] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in *Neurocomputing*. New York, NY, USA: Springer, 1990, pp. 41–50.
[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognit. Modeling*, vol. 5, no. 3, p. 1, 1988.
[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[20] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, "Challenges and opportunities with big data visualization," in *Proc. 7th Int. Conf. Manage. Comput. Collective Intell. Digital Ecosyst.*, 2015, pp. 169–173.
[21] H. G. Boomgaarden and R. Vliegenthart, "How news content influences anti-immigration attitudes: Germany, 1993–2005," *Eur. J. Political Res.*, vol. 48, no. 4, pp. 516–542, 2009.
[22] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.
[23] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proc. 11th Int. Workshop Semantic Eval.*, 2017, pp. 502–518.
[24] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," 2013, *arXiv:1304.5634*. [Online]. Available: https://arxiv.org/abs/1304.5634
[25] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.
[26] M. Fratello, G. Caiazzo, F. Trojsi, A. Russo, G. Tedeschi, R. Tagliaferri, and F. Esposito, "Multi-view ensemble classification of brain connectivity images for neurodegeneration type discrimination," *Neuroinformatics*, vol. 15, no. 2, pp. 199–213, Apr. 2017.

ALBERTO COSIMATO received the master's degree in computer science from the University of Salerno, in 2019. His research interests include social media and data analytics.

ROBERTO DE PRISCO received the Laurea from the University of Salerno, Italy, in 1991, the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1997 and 2000, respectively, and the Dottorato from the University of Napoli, Italy, in 1998, all in computer science. In 2000 and 2001, he was a Research Scientist with Akamai Technologies. He is currently a Professor of computer science with the University of Salerno, Italy, where he is also a Co-Founder of eTuitus, an academic spin-off. His research interests include algorithms, distributed systems, cryptography, network security, and computer music. He is responsible for the Musimathics Laboratory, Dipartimento di Informatica, University of Salerno. He serves as a Referee for scientific journals and conferences and involved in the organization of scientific conferences.

ALFONSO GUARINO received the master's degree in computer science from the University of Salerno, in 2016, where he is currently pursuing the Ph.D. degree with the Department of Computer Science. His research interests include techno-regulation, privacy, machine learning, and visualization.

DELFINA MALANDRINO received the Ph.D. degree in computer science from the University of Salerno, Italy, in 2004, where she has been an Associate Professor in computer science with the Dipartimento di Informatica, since April 2019. Her research interests include distributed systems on the world wide Web and intermediaries, collaborative and learning systems, social and network analysis, privacy, green computing and power-aware software, usability studies, visualization, and open data.

NICOLA LETTIERI is currently a Researcher with the National Institute for Public Policy Analysis, Rome, and a Professor of legal informatics and computational social sciences with the University of Sannio, Benevento. He is also a co-editor of the international series *Law, Science, and Technology*, an Associate Editor or a member of the editorial board of the international journals Frontiers in Artificial Intelligence, Law and Technology, Future Internet, and Frontiers in Evolutionary Sociology. He was an Associate of the Laboratory of Agent-based Social Simulation (Istc / Cnr) and a Secretary of the Italian Association of Cognitive Sciences.

GIUSEPPE SORRENTINO received the master's degree in computer science from the University of Salerno, in 2019. His research interests include business intelligence and social media.

ROCCO ZACCAGNINO received the Ph.D. degree in computer science from the University of Salerno, Italy, in 2012, where he is currently a Postdoctoral Researcher with the Dipartimento di Informatica. His research interests include computational intelligence: theory and applications, computer music, formal languages, and bioinformatics.