Multimed Tools Appl (2018) 77:26371–26396
https://doi.org/10.1007/s11042-018-5865-0

Assessing the reTweet proneness of tweets: predictive models for retweeting

Paolo Nesi[1] · **Gianni Pantaleo**[1] · **Irene Paoli**[1] · **Imad Zaza**[1]

Received: 30 June 2017 / Revised: 1 March 2018 / Accepted: 4 March 2018 /
Published online: 20 March 2018

Abstract The problem of assessing the mechanisms underlying the phenomenon of virality of social network posts is of great value for many activities, such as advertising and viral marketing, influencing and promoting, early monitoring and emergency response. Among the several social networks, Twitter.com is one of the most effective in propagating information in real time, and the propagation effectiveness of a post (i.e., tweet) is related to the number of times the tweet has been retweeted. Different models have been proposed in the literature to understand the *retweet proneness* of a tweet (tendency or inclination of a tweet to be retweeted). In this paper, a further step is presented, thus several features extracted from Twitter data have been analyzed to create predictive models, with the aim of predicting the degree of retweeting of tweets (i.e., the number of retweets a given tweet may get). The main goal is to obtain indications about the probable number of retweets a tweet may obtain from the social network. In the paper, the usage of the classification trees with recursive partitioning procedure for prediction has been proposed and the obtained results have been compared, in terms of accuracy and processing time, with respect to other methods. The Twitter data employed for the proposed study have been collected by using the Twitter Vigilance study and research platform of DISIT Lab in the last 18 months. The work has been developed in the context of smart city projects of the European Commission RESOLUTE H2020, in which the

✉ Paolo Nesi
paolo.nesi@unifi.it; http://www.disit.dinfo.unifi.it; http://www.disit.org

Gianni Pantaleo
Gianni.pantaleo@unifi.it

Irene Paoli
Irene.paoli@unifi.it

Imad Zaza
Imad.zaza@unifi.it

[1] DISIT Lab, University of Florence, Florence, Italy

capacity of communicating information is fundamental for advertising, promoting alerts of civil protection, etc.

Keywords Social media · Twitter monitoring · Retweet proneness · Virality · Predictive models · Principal component analysis · Classification trees · Machine learning

1 Introduction

In recent years, social media have become an important communication tool and instrument for monitoring preferences of users, as well as making predictions in a number of contexts. Many social media platforms allow rapid multimedia information diffusion, and thus they may be used as a source of information for viral advertising and marketing, early warning, emergency response and, more generally, for promoting and/or informing many users. Among the various platforms, Twitter.com has a very large user base, consisting of 1.3 billion of accounts and hundreds of millions of users per month. Twitter users can produce a post (i.e., a "tweet"), about any topic within the 140-characters limit and can follow other users, in order to receive their tweets/posts on their own twitter web page, as well as on the mobile App. Twitter plays an important role in spreading information, allowing people to communicate and share contents in a fast manner. The posts made by a user are displayed on his/her profile page, and they are also brought to the attention of all his/her followers. It is also possible to send some direct private messages to other users without provoking diffusion. Another solution to enhance the diffusion and the echo of tweets is to include in a tweet including a direct mention of a user; this can be done by using the "@" prefix such as "*@usernickname*". In this case, the *@usernickname* user is stimulated by receiving a notification. Therefore, the information conveyed in a tweet is diffused among the social network users through retweets of the former tweet, thus echoing the original message to the followers, hence producing a chain of messages since the retweets are also echoed. A retweet represents the echo of an original tweet made by one user that has been automatically forwarded by Twitter.com to the followers of the retweeting users (a part for eventual promotions performed by Twitter.com for featuring the most important tweets when they are getting on the list of the most appreciated). In the world of Twitter, the effectiveness of a tweet is frequently measured in terms of retweet count, which is the number of times the tweet has been retweeted [46]. It gives a measure of the number of reached audience and/or appreciation.

There is a growing interest, both in research and commercial fields, for influential strategies and solutions for seeding and diffusing information. Twitter offers to business users the possibility to integrate its analytics with audience measurement tools and services, such as Nielsen Digital Ad Ratings (DAR) and ComScore validated Campaign Essentials (vCE). Overviews of predictive methods exploiting tweets have been proposed in the works of Sikdar et al. [52], Madlberger and Almansour [37], Zaman et al. [61]. In most cases, the predictive capabilities of Twitter data have been identified by using volume metrics on tweets (i.e., the total number of tweets and/or retweets associated with a Twitter user or presenting a certain hashtag). However, in specific cases, a deeper semantic understanding of tweets has been required to create useful predictive capabilities. Thus, algorithms for sentiment analysis computation have been proposed to consider the meaning of tweets by means of natural language processing algorithms. Moreover, the adoption of techniques for segmenting, filtering or clustering by context (e.g., using natural language processing for avoiding the

misclassification of tweets talking about flu), or by users' profiles (e.g., age, location, language, and genre) may help to obtain more precise results in terms of predictability. On the other hand, the aim of this paper is to study the *retweet proneness* of a tweet, which we define and refer in the following of the paper as the capability to be retweeted, including a quantitative measure of the number of retweets a given tweet may get (which can be considered as the potential degree of being retweeted).

This paper is focused on presenting a study on identifying and assessing the most representative metrics which can be used to predict the *degree of retweeting* of a tweet (i.e., the number of retweets a given tweet may get). According to the literature, the tweet features can be related to the tweet, to the author of the tweet and thus to the network of relationships of the tweets' author. The study is grounded on the analysis of tweets datasets collected in different areas in the last 18 months, for a total amount of about 100 million posts. By analyzing the datasets with the aim of identifying the best predicting model allowed us to identify also the main characteristics of tweets to predict the *degree of retweeting*. Please note that, according to the state of the art reviewed and presented in the following section, the identification of models for estimating of the *degree of retweeting* of a tweet has been only partially addressed in the literature; a few efforts are mainly focused on identifying parameters to guess the probability of retweeting, and/or to study the cascading effected through the network.

To our knowledge, the main original contributions brought by the work proposed in the present paper are the following: our work aims not simply predicting the probability for a tweet to be retweeted, rather to go a step further, which is predicting and estimating the *degree of retweeting*. Moreover, the proposed analysis identified additional relevant metrics/features, with respect to those proposed in the reviewed literature, such as the publication time of tweets and the number of users who added a given tweet's author to a list, as discussed later in more detail. The motivation for establishing the probability of prediction of a tweet is related with the value of the tweet itself and the value of the advertising service that may have produced it. The estimation of the probability to be retweeted is a measure of the effectiveness of a tweet and it is somehow a more precise measure of the concept of tweet *virality*, that tend to assess only tweets and their context to create huge volumes of retweets.

The paper is organized as follows. In Section 2, a review of the state of the art and related works found in recent literature is presented. In Section 3, the general architecture of the Twitter Vigilance solution, adopted for collecting Twitter data and making statistical analysis, is reported and discussed. Section 4 provides an overview of the methods and models adopted to explain the metrics that might affect the number of retweets of a tweet and the prediction of the *degree of retweeting*. In Section 5, preliminarily the different classification models are summarized; then the predictive method is presented together with an analysis of features that determine the retweet proneness of tweets. Section 6 provides a comparison among results that can be obtained by using different models. Conclusions are drawn in section 7.

2 Related works

In this section, the predictive capability of Twitter data has been reviewed with the aim of providing a better view of the context in which the research has been developed, and the impact of the obtained results. In the work of Sinha et al. [53], a solution for predicting results of football games has been proposed, taking into account the volume of tweets. Opinions pools

and politic elections predictions have been proposed to be correlated with the volume of tweets by using Sentiment Analysis techniques in O'Connor et al. [43]. Different models based on volume of tweets and other means have been also used for predicting purposes: voting results in Bermingham and Smeaton [3] and in Tumasjan et al. [56], economics [4, 15], marketability of consumer goods [50], public health seasonal flu [1, 34, 51], box-office revenues for movies [2, 36, 38, 54], crimes [58], book sales [26], recommendations on places to be visited [14] and weather forecast information [24, 25]. Moreover, Twitter-based metrics have been used to predict and estimate the number of people in some location, such as airports, the so-called crowd size estimation by the work of Botta et al. [5], as well as to predict the audience of scheduled television programmes, where the audience is highly involved, such as it occurs with reality shows (i.e., X Factor and Pechino Express, in Italy) [17]. Other adoptions of Twitter have been used to perform risk analysis [29].

In general, a Twitter user could find a tweet worth sharing, and therefore he/she may retweet it to followers. There is no upper limit to the number of times a retweet (re-post) operation can be performed. Hence, multiple levels of retweeting can be identified (considering the retweet of an original tweet as the first-level). A user could actually retweet a formerly retweeted post to his/her followers, and his/her followers can do the same again and again. In this way, retweets became a popular mean of propagating information through the Twitter community, as they may get viral propagation when volumes of retweets become high. Most studies about the assessment of the retweeting capability of tweets (proneness of a given tweet to be retweeted) try to analyze retweeting behaviors and, thus, to discover the features that may help Twitter users (i.e., the tweets' authors) in creating tweets which are more effective in collecting retweets. In the literature, different models have been proposed to shed some light on what kind of factors are likely to influence information propagation in Twitter.

Various motivations for retweeting behaviors have been explored in the paper of Golder [22]. They found that the most influential users can retain significant influence over several different topics. In the works of Kwak et al. [33] and Cha et al. [13], the relationships between the number of followers of Twitter users and their influence and lists of the most influential Twitter users, compiled according to a variety of metrics (including retweet count), have been investigated. Kwak et al., have ranked users by the number of followers and by PageRank, and found the two rankings to be similar. They have analyzed the tweets of top trending topics and reported on the temporal behavior of trending topics and user participation. Cha et al. [13] have examined three types of influential users, performed in propagating popular news topics. Hansen et al. investigated the features of tweets that garner large numbers of retweets, analyzing a dataset of 210,000 tweets about the 2009 United Nations Climate Change Conference, as well as a random sample of about 350,000 tweets from 2010 [27]. Hong et al. [28], studied the dynamics of user influence across topics and time, as well as the problem of predicting the popularity of messages as measured by the number of future retweets. The study was conducted by classifying tweets in four categories according to the number of retweets they received (0, < 100, $[100, 9999]$, $\geq 10{,}000$), formulating the prediction task as a classification problem. Moreover, they used a multi-class classifier, training it on one week and testing it on the next week for creating a short-term prediction. Naveed et al. [41] used a similar technique to predict the probability that a tweet receives any retweets. They proposed a predictive model to forecast the likelihood for a given tweet of being retweeted, based on its contexts; furthermore, they deduced what are the most influential features that contribute to the likelihood of a retweet on the basis of the parameters learned by the model. In the work of Suh et al. [55], a number of features that might affect the probability of tweets to be

retweeted ("retweetability", e.g., retweet proneness of a tweet) have been examined by using the principal component method and logistic regression models. The aim was the assessment of the probability of a tweet to be retweeted without assessing the *degree of retweeting*. Amongst the features that can be computed for each tweet, the presence of URLs and hashtags in the tweet body have been proved to present a strong relationship with retweetability. The experiment has been computed on a small dataset of 10 K observations, and the achieving prediction accuracy is not reported. Pezzoni et al. [46] have defined the "*influence*" as the ability of a user to spread information in a network, assuming that the retweet count may measure the popularity of a message on Twitter. The influence of a user could be also estimated by the average number of retweets collected by all tweets of the user. In that paper, the authors demonstrated by simulation that the probability to be retweeted is modeled by a power law function and the capacity of the most influential authors depends on their number of followers. Peng et al. [45] have proposed a model called *retweet patterns* (i.e., the retweet propagation trend). In that case conditional random fields have been used, taking into account three types of features: tweets features, users features and relationship features (which incorporates the perspectives whether the tweet may be simultaneously retweetable for two users). They have constructed the network relations for retweet prediction, and have demonstrated that conditional random fields can improve prediction effectiveness by incorporating social relationships, compared to those baselines that do not take into account such feature. Morchid et al. [39] have computed both Naïve Bayes and Support Vector Machine models considering two classes: tweets retweeted less than 30 times and tweets retweeted more than 100 times (massively retweeted tweets). The aim of their study was to detect those tweets that are *massively retweeted* in a short time, however without addressing the problem of predicting the potential number of retweets. They also used the principal component analysis to evaluate relevant features that could have an impact in detecting some retweeting proneness, without proposing a model for assessing the degree of retweeting, thus presenting only an exploratory descriptive approach. Zaman et al. [60] have measured the popularity of a tweet through the time-series path of its retweets, by using a Bayesian probabilistic model. They have used the user ID of the original tweet and retweet authors, the number of followers and the word contained in tweets to predict the future retweets. Uysal and Croft [57] proposed a predictive model for estimating the likelihood of retweeting for a given user and tweet by using a logistic regression model.

Yang and Counts [59], used a factor graph model to investigate the retweeting behavior focusing on those features related to the user profile and to the content of a tweet. Can et al. [10] focused their research on predicting the expected retweet count of a tweet by studying three types of features: content based features (presence or absence of hashtags), structure based features (as followers count, friends count, statuses count), as well as multimedia and image based features (the distribution of color intensities, perceptual dimensions, responses of individual object detectors). They have used the logarithm of retweet count for a given tweet as the response variable, and three different types of regression: linear, SVM with a Gaussian kernel, and Random Forest. The experiments produced better results with Random Forest, providing a RMSE score of 1.297 in log scale, very close similar performances have been obtained with SVM. They identified the Followers counts to be the most correlated feature. A common drawback found in content-based predicting tools reviewed in the literature, is represented by the 140-character constraint imposed by Twitter, which makes it difficult to identify and extract content-based predictive features [10].

Pálovics et al. [44] have treated the retweet prediction as a binary classification problem. They have used a multi-class classification for ranges of cascade sizes, in order to directly

predict the logarithm of the retweets volume. For each day in the testing period, they have trained a Random Forest classifier to predict the future volume of retweets for tweets appearing on the day. The experiments have been compared by using the AUC (area under the precision-recall curve) demonstrating the dependency of the model with respect to the user feature (e.g., followers counts), hashtag used popularity, user network features. Bunyamin and Tunys [9], have provided a comparison of the performance for different learning methods and features, in terms of retweet prediction accuracy and feature importance, to understand what kind of tweets would be retweeted, by using as response variable a dummy variable representing the two states of being retweeted or not retweeted. They have found that Random Forests method archives the best performance. Moreover, they have found and included among the best features the following ones: number of times the user is listed by other users, number of followers, and the average number of tweets posted per day. On the same line, Jiang et al. [30] and Zhang et al. [62] have treated the retweeting behavior prediction as a binary classification problem, achieving an accuracy of 0.85 and 0.789 respectively. Liu et al. [35] have proposed a two-phase model to predict how many times a tweet can be retweeted in Sina Weibo microblog. In the first step, they have built a multi-classification model, while in the second step a regression model on each class has been constructed. They have achieved a high Mean Absolute Error of 58.22%, using the combination of Random Forest model and Least Median Squared Linear Regression model. However, discussion about the importance of each considered features is not reported. Firdaus et al. [19] have tried to consider user's different behaviors in different roles for the purpose of retweet prediction. They argue that the retweet prediction model might give better prediction accuracy results when the difference between the behavior of the author and retweetters is considered, determining the topic of interest of a user based on his past tweet and retweet.

3 Twitter vigilance architecture

The Twitter Vigilance platform (http://www.disit.org/tv/) has been designed and realized by the DISIT Lab of University of Florence as a multipurpose comprehensive tool providing different tasks and metrics suitable for Twitter search API and streams, their monitoring and analysis, for research purpose [12]. The architecture is depicted in Fig. 1.

In Twitter Vigilance, a distributed crawler performs data gathering and extraction by using Twitter Search API. The data acquisition approach is based on the concept of *Twitter Vigilance Channel*, consisting in a set of simple and complex search queries which can be defined by a registered user by combining keywords, hashtags, user's IDs, citations, etc., in a structured logical syntax, according to the search syntax of Twitter. The search queries associated with each *Twitter Vigilance Channel* are posed to the Twitter platform via a crawler. Both configuration parameters and statistical results are accessible from the front-end interface for the user. Collected tweets are made accessible to the back-office processes, which implement statistical analysis, natural language processing (NLP) and sentiment analysis (based on distributed NLP on Hadoop [42]), as well as general data indexing. The metrics resulted by the back-office processes are stored on a dedicated database and made accessible to the front-end graphical user interface (see Fig. 2 as an example), which allows visual analytics, temporal trends and time series visualizations, data results navigation, Twitter users statistics and analysis.

All these kinds of analysis are performed at both *Twitter Vigilance Channel* level and at single search level. In the specific, the following information and metrics can be retrieved:

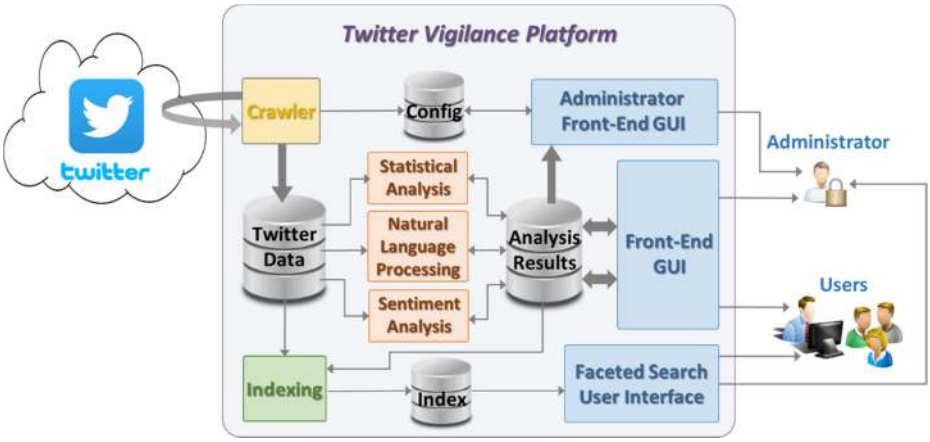


Fig. 1 Twitter vigilance architecture

number of tweets and retweets; user citations (to detect potential influencers, pushers, emerging citations, etc.); hashtags (to understand which are the most used, emerging, evolving, etc.); keywords tagged with their part-of-speech (that is, their grammatical function), in terms of nouns, verbs, and adjectives; sentiment analysis; relationships among users; etc.

The derived metrics and information can be useful to understand which are the most widely used or emerging hashtags, as well to detect which are the most influential in determining the positive/negative signature and polarity detection in the sentiment analysis, and thus for better tuning the tweet collected and for precomputing basic metrics that can be useful for the researcher to make further analysis in different domains and generically for communication and media, predictive models [24, 25]. It can be a useful tool for identifying reasons for positive/negative tweets, as well as the reaction of the community.

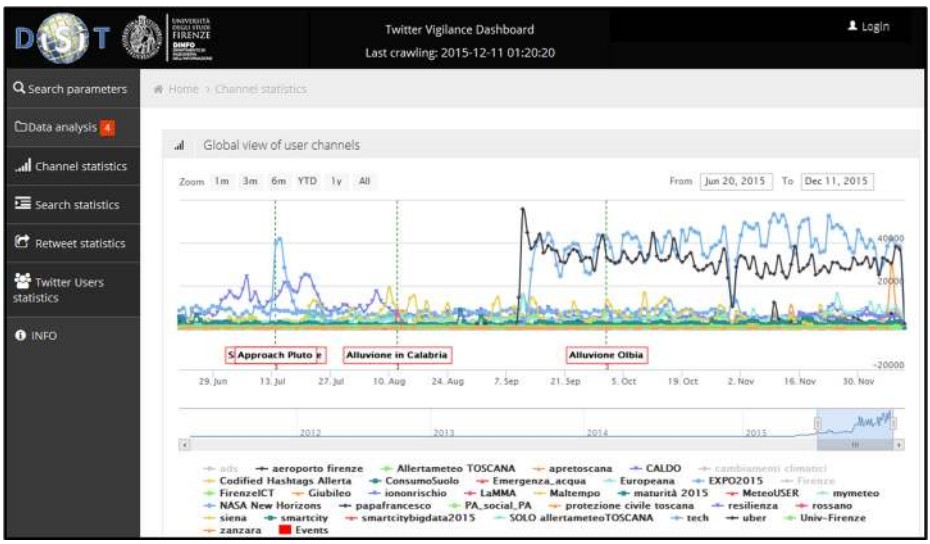


Fig. 2 Twitter vigilance front-end graphic user interface, showing temporal trends volume based metrics calculated for different user defined channels

4 Assessment framework for retweet modeling by using Twitter vigilance outcomes

According to the above presented state of the art, retweeting is a powerful mechanism to diffuse information on Twitter. The number of retweets of a tweet can be considered as a measure of how much the produced tweet has been effective in propagating the information, which is one of the major motivations for tweeting on Twitter.com. The proposed study aims at identifying the values of tweets' features which may determine the *degree of retweeting* and, as a side effect to understand the mechanisms which may determine retweeting in Twitter. The main goal is to create a predictive model for assessing the *degree of retweeting*, and thus to classify tweets in terms of certain classes for their *degree of retweeting*. The computational process at the end is performed through the following steps as depicted in Fig. 3, and better described in the rest of the paper:

I. Collection of the data from Twitter.com by crawling them by using Twitter Vigilance platform and tools on the basis *searches* and *channels*. The platform allows computing simple metrics for counting tweets/retweets for search and channel, extracting relationships among users, etc.
II. Selection of predictors/features from collected data and metrics.
III. Computation of potential predictors: a statistical criterion is applied to identify the statistically significant features. The use of an exploratory method is a crucial issue not only for ranking the variables before the construction of a prediction model, but also to give the phenomenon's first interpretation and to understand the underlying data structure.
IV. Computation of a predictive model for the assessment of the binary probability to be retweeted or not.
V. Computation of a model to predict the *degree of retweeting.* The results have been obtained by comparing several different computational alternatives and approaches and selecting the better ranked and the most relevant metrics as described in the following.

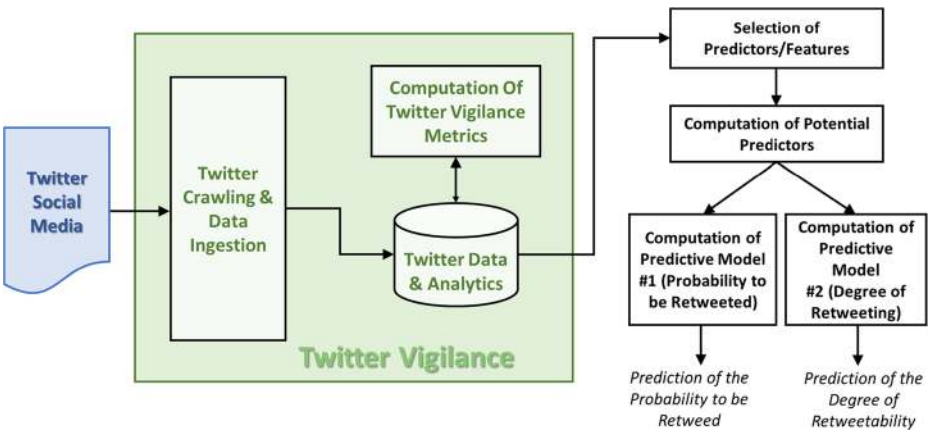


Fig. 3 Workflow of the overall process carried on by the proposed framework, from Twitter data ingestion to the computation of the predictive model

According to the previous statements, we have adopted Classification And Regression Tree (CART) models to understand the relevance of variables and to construct a model for predicting the probability to be retweeted and the *degree of retweeting*.

4.1 Collection of the datasets

Three datasets have been considered for the analysis. The first includes 100 Million of tweets (100 M dataset) related to 45 different *Twitter Vigilance Channels* covering many different topics but collected on the basis of a large number of search keys on Twitter.com API (which can be mainly related to terrorism, weather, mobility and transport, politics, city services, health and drugs, tourism and city, TV events, etc., see Fig. 4 for details) from a larger set of 200 million dataset (as defined in Section 3, from April 2015 to June 2016). The second set includes 100,000 randomly selected tweets (100 K dataset) from the 200 million dataset. The third includes 500,000 randomly selected tweets (500 K dataset) from the 200 million dataset. All datasets have been used to perform an exploratory analysis, a classification and a regression tree model. From the 100 M dataset, the 61% of the tweets are in English, the 12% in Italian, the 9% in Spanish and the remaining tweet are in many other languages. In Fig. 4, details of the distribution of collected posts are illustrated, showing the most numerous (covering almost 90% of the whole collected dataset) search queries used for data ingestion (i.e. hashtags, citations, keywords etc.) grouped in their pertaining Twitter Vigilance channels; actually, as described in Section 3, a Twitter Vigilance channel can be considered as a thematic categorization of a set of semantically similar search queries. However, it is worthy to be noticed that the analysis and estimation of the *degree of retweeting* performed in this work is not dependent from the topic or subject.

4.2 Identification of potential features/metrics

As a second step, a set of features/metrics has been identified from the literature, by considering the information available on Twitter data, and by performing a qualitative analysis of

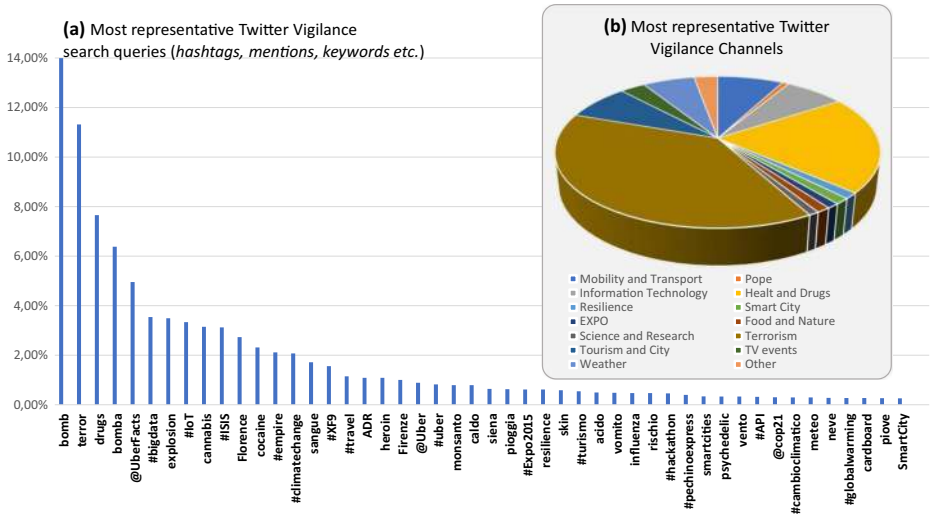


Fig. 4 Distribution of collected posts dataset, showing the most frequent search queries (a), grouped by their pertaining Twitter Vigilance channels (b)

twitter mechanisms by using a metric identification approach and methodology, such as GQM (Goal, Question, Metric). Such an approach has been followed considering that it would be desirable to identify metrics that may have some predictive capabilities in explaining the *degree of retweeting*. The identified metrics are reported in Table 1, in which some metrics can directly refer to data and information contained in the single tweet, while other ones are derived from the author that has produced the tweet. A first set of metrics concerns the content of the tweet, and includes the number of Hashtags, Mentions and URLs contained in the message, the number of Favorites obtained by a tweet. A second set of metrics is about the tweet authors, and includes information regarding the user who posted the tweet: the number of days since the author created the Twitter account and the number of tweets posted since the creation of its own account (Statuses). A third set of metrics is related to network connected to the author: the number of users who follows the author of a tweet (Followers), the number of friends that author is following (Followees) and the number of other users that have listed the author in some of their own lists (Listed Count). A part of the identified metrics has been also used in [55], where a simple descriptive and Principal Component Analysis have been provided without deriving a predictive model. In the paper of Bunyamin and Tunys [9], a comparative analysis of several methods has been proposed without considering all metrics we identified, and without addressing the prediction of the degree of retweeting.

In the proposed analysis, we have specifically addressed metrics such as: Publication Time and Listed Count. The Publication Time metric should consider the classical claim stating that a higher probability of retweeting could be achieved if the tweet is published when the audience is on-line. The Listed Count metric should consider the reputation of the author, which is an additional level with respect to be just followed by another user. In addition to the metrics reported in Table 1, we also collected the Retweet Count (i.e., # of retweets obtained by the tweet), which can be considered, in our case, the target of our prediction models and not a real metric.

4.3 Computation and understanding of potential predictors

In a third phase, all the metrics have been extracted for the above-mentioned datasets. Figure 5 reports the percentage of the distribution of *Retweet Count* for the 100 million dataset.

Table 1 Considered features/metrics from the tweet information

Tweet metrics	Description
URLs count	# of URLs in the tweet
Mentions count	# of mentions/citation of Twitter users in the tweet
Hashtags count	# of hashtags included in the tweet
Favorites count	# of favorite obtained by the tweet
Publication time	Local hour H24 in which the tweet has been published in the day according to the author' local time.
Author of tweet metrics	Description
Days count	# of days since the tweet's author created its Twitter account
Statuses count	# of tweets made by the tweet's author since the creation of its own account
Author network metrics	Description
Followers count	# of followers the author of the tweet
Followees count	# of friends the tweet's author is following
Listed count	# of people added the tweet's author to a list

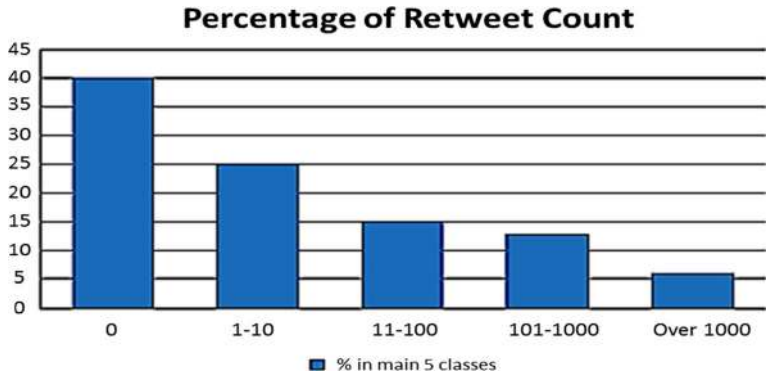


Fig. 5 Percentage of the retweet count distribution in main 5 classes

Then, Principal Component Analysis (PCA) has been applied. PCA is an exploratory technique for multivariate data, applied as a structure analysis method typically used to reveal the underlying structure that maximally accounts for the variance in datasets. The basic goal of PCA is to describe variations in a set of correlated variables, $x^T = (x_1, \ldots, x_q)$, in terms of a new set of uncorrelated variables, $y^T = (y_1, \ldots, y_q)$, each of which is a linear combination of x variables. The new variables are derived in decreasing order of importance in the sense that y_1 accounts for as much as possible of the variation in the original data amongst all linear combinations of x. Then y_2 is chosen to be uncorrelated with y_1 and to account for as much as possible of the remaining variation, and so on. The new variables defined by this process, $y_1, \ldots, y_q$, are the principal components [18]. The first few components will account for a substantial proportion of the variation in the original variables, and they can be used to provide a lower-dimensional summary of these variables. To identify the optimal number of factors, several informal and more formal techniques are available [31]. The most common procedures to choose the number of components/metrics to retain are the following:

- Retain just enough components to explain some specified large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as q or n (the sample size) increases [18].
- The Kaiser criterion [32] recommends retaining only factors with eigenvalues greater than one.
- The screen test of Cattell [11], recommends plotting the eigenvalues and finding a place where the smooth decrease of eigenvalues appears to level off to the right of the plot. The number of components selected is the value corresponding to an “elbow” in the curve, i.e., a change of slope.

PCA provides a first general idea about the internal structure of the data in a way that best explains the variance. PCA is performed on a representative random sample of 100 K observations with the eleven features (see Table 1), also including in this case the retweet count as performed by [55] on smaller number of variables. Table 2 reports the importance of factors extracted by PCA in descending order of variance. In the second column of Table 2, the eigenvalues that represent the variance for each factor are reported. The corresponding percentage of the variance is shown in the third column of the table. With respect to our analysis on a 100 K tweet dataset, according to the Kaiser Criterion and to the screen test (see

Table 2 Importance of principal components

Factors	Eigenvalue	% variance	% Cumulative variance
1	1.9545	17.7681	17.7681
2	1.3748	12.4979	30.2659
3	1.0777	9.7976	40.0636
4	1.0335	9.3959	49.4594
5	1.0248	9.3164	58.7758
6	0.9623	8.7485	67.5243
7	0.9523	8.6576	76.1819
8	0.9339	8.4899	84.6717
9	0.7679	6.9808	91.6526
10	0.5976	5.4325	97.0851
11	0.3206	2.9149	100

Fig. 6), the right number of principal components to be considered as relevant is five. The first five factors account for the 58.77% of the total variance. In Suh et al. [55], only 3 main PCA with an eigenvalue greater than 1 have been identified, explaining the 44,34% of the variance (Kaiser criterion), and considering only 10.000 tweets. In the work of Morchid et al. [39], 4 main components have been identified, explaining the 56.34% of the variance considering 6 million of tweets, not sampled from a larger dataset.

In Table 3, the principal components loading for the features of Table 1 (plus *Retweet Count*) are reported. The component correlations of the original metrics are graphically depicted in Figs. 7, 8, 9, and 10. Each feature in Table 2 is mapped into a vector in the factor map. The vector represents the correlation between the feature and the principal components (the axis of the graph).

Factor 1 carries more than 17% of the total variability of the dataset (Table 2), and this variability is mainly explained by the covariates *Favorite Count*, *Followers Count* and *Listed Count*. This first factor is strongly different with respect to the one identified by the Kaiser criterion [32], since the *Listed Count* metric (which is dominant) was taken into account in that article. The variability of Factor 2 (12.5%) is carried by the negative correlation of *Hashtags Count* (−0.5661) and *URLs Count* (−0.5483), while Factor 3 explains about 9.7% of the total variability, and it is represented by *Followees Count* feature. Component 4 explains almost

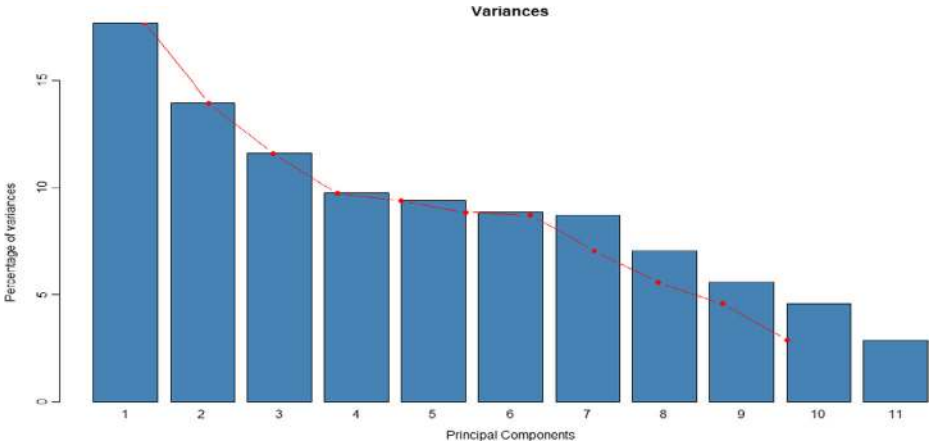


Fig. 6 Distribution of the percentage of variance from PCA analysis

Table 3 Principal component loadings

Metrics	PC1	PC2	PC3	PC4	PC5
Retweet count	−0.1623	**0.4346**	0.1635	−0.0026	−0.1009
Favorites count	**−0.6294**	0.3908	0.1922	−0.1128	−0.1880
Followers count	**−0.7599**	0.2736	0.0522	−0.0983	−0.0857
Followees count	−0.1336	−0.0907	**−0.4627**	−0.2494	0.1182
Listed count	**−0.8431**	−0.1549	−0.0498	0.1500	0.1871
Statuses count	−0.4256	**−0.5016**	−0.3781	0.2795	0.2410
Hashtags count	−0.1585	**−0.5661**	0.4377	−0.0517	0.0309
Mentions count	0.0394	0.2194	0.0786	−0.1607	**0.7697**
URLs count	−0.1288	**−0.5483**	0.2539	−0.3388	−0.3248
Publication time	0.0076	−0.0728	0.3639	**−0.5186**	0.3707
Days count	−0.0370	0.0070	−0.5072	**−0.6604**	−0.1691

Data reported in bold are the most relevant in the context

9.3% of the total variability, and it is negatively correlated with the *Publication Time* of a tweet and the age of the author account (*Days Count*). Please note that also the *Publication Time* was not considered in [32]. The *Mentions* feature (0.7696) is mainly carried by Factor 5, and it explains the same proportion of variability of Component 4. PCA allowed to sort the features according to the impact on total variability, as well as to understand the correlation among the metrics and the number of retweets.

According to the analysis results, the most relevant metrics are: *Mentions Count* (76.9% of Factor 5 total variability); *Listed Count* (explains the main variability of Factor 3 sharing it with *Followers* and *Favorite*); *Hashtags* (that explains the main variability of Factor 2, sharing

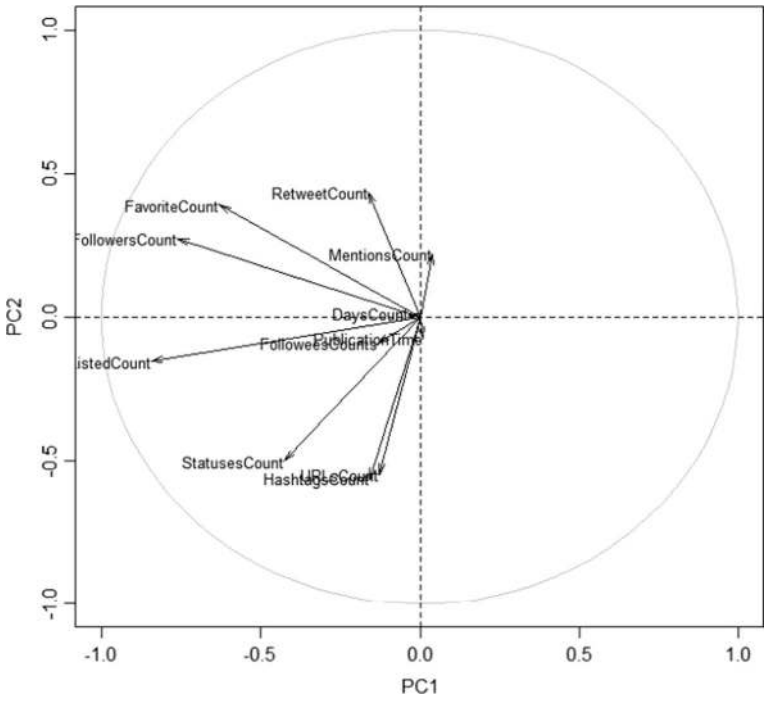


Fig. 7 PCA factor map with factor 1 and factor 2

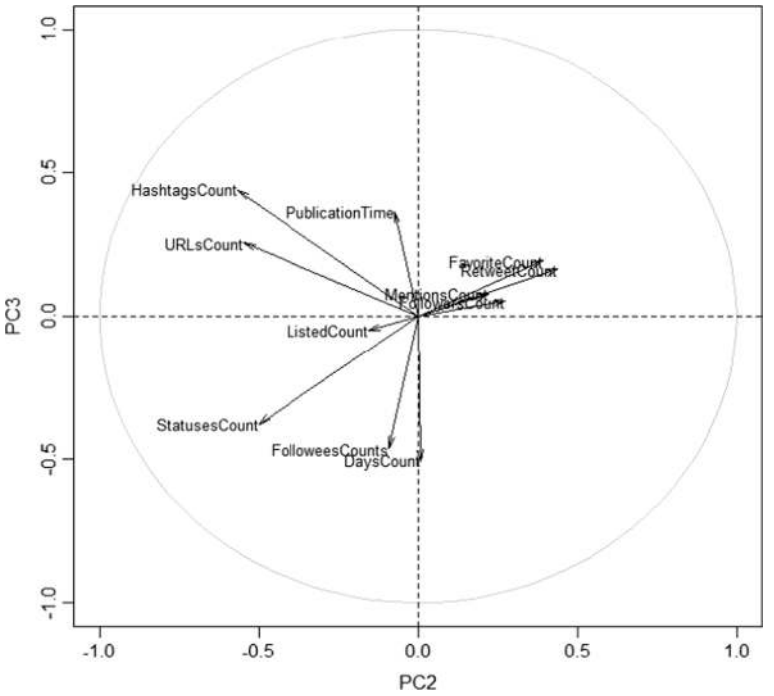


Fig. 8 PCA factor map with factor 2 and factor 3

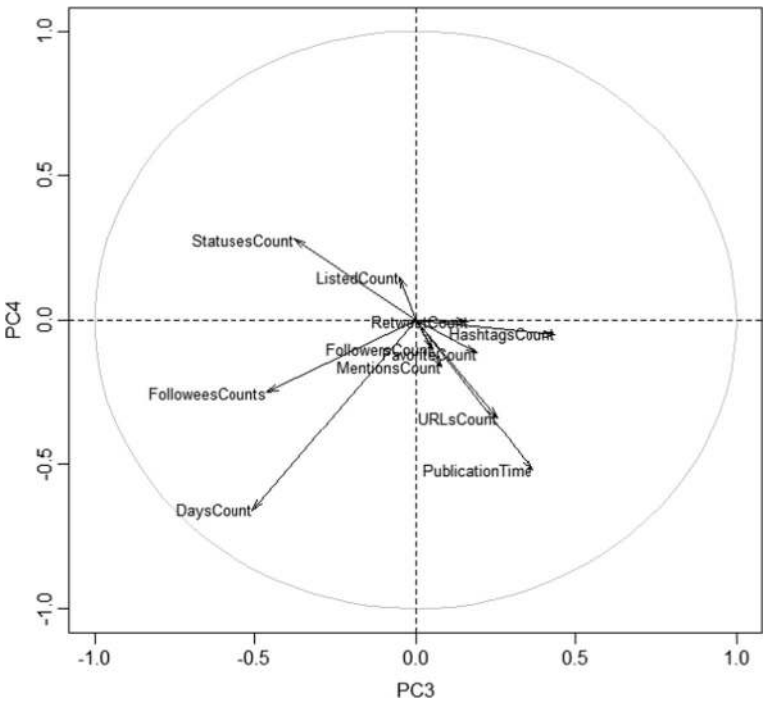


Fig. 9 PCA factor map with factor 3 and factor 4

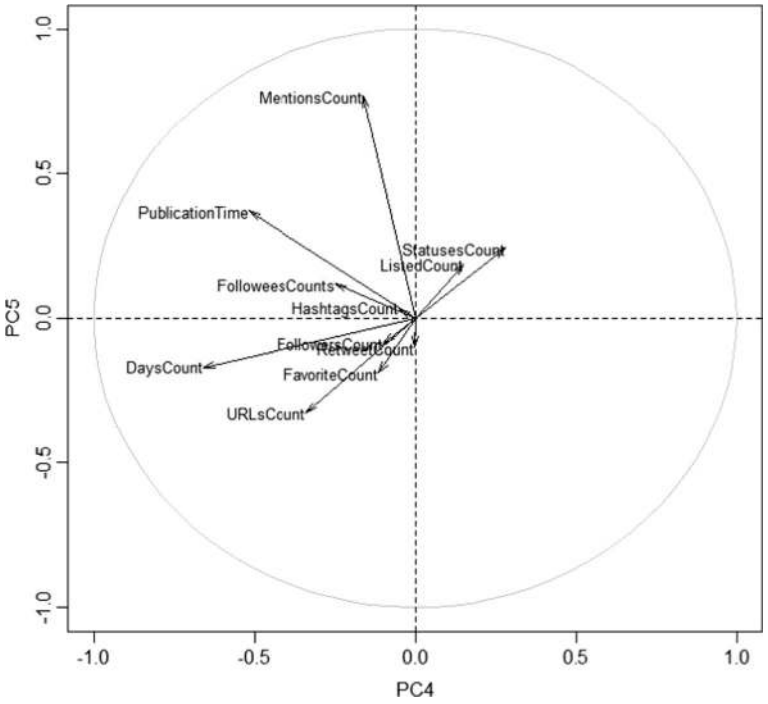


Fig. 10 PCA factor map with factor 4 and factor 5

it with *URLs Count, Statuses Count* and *Retweets Count*); *Days Count* (that explains the main variability of Factor 4, sharing it with *Publication Time*).

5 Predicting the probability to be retweeted and the degree of retweeting of a tweet

In this section, before to present the analyses performed, a presentation of the considered classifications methods is provided. Then, the different analyses are reported. As a first phase, as reported in Section 5.2, a binary classification has been performed to create a model to identify tweets that have a probability to be retweeted, and thus the most relevant features that may determine the model. As a second phase, Section 5.3 presents the model for predicting the degree of retweeting of tweet. Also in this case, the most relevant features for the prediction have been identified.

5.1 Analysis of the considered classification methods

Classification Trees are machine-learning methods for constructing prediction models from data, and they have been widely used for the data exploration, description and prediction purposes. Trees have many properties, including their ability to handle various types of response such as numeric, categorical, censored, multivariate, and dissimilarity matrices; trees are invariant to monotonic transformations of the predictors; complex interactions are modeled in a simple way; besides, missing values in the predictors are managed with minimal loss of

information. Thanks to these properties, the use of classification and regression trees (i.e., a recursive partitioning method that is free from distributional assumptions), has potential advantages to construct predictive models.

In this section, a short recall of the methods considered and compared for creating a suitable predicting model to estimate the *degree of retweeting* for single and/or groups of tweets is reported.

Recursive partitioning procedure models (RPART) are defined by recursively partitioning the data space, and defining a simple local prediction model for each resulting partition. This can be represented graphically as a decision tree, with one leaf per partition [6]. The model can be written in the following form (1):

$$f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}] = \sum_{m=1}^{M} w_m \mathbb{I}(\boldsymbol{x} \in R_m) = \sum_{m=1}^{M} w_m \phi(\boldsymbol{x}; \boldsymbol{v}_m) \tag{1}$$

where R_m is the m-th partition, w_m is the response in this partition, and $\boldsymbol{v}_m$ encodes the choice of variable to split on, together with the threshold value, on the path from the root to the m-th leaf. The best feature and the best value for that feature have been chosen by the split function (2):

$$\left(j^*, t^*\right) = \arg \min_{j \in \{1,\ldots,\mathrm{D}\}} \min_{t \in T_j} \operatorname{cost}\left(\left\{\boldsymbol{x}_i, y_i : x_{ij} \le t\right\}\right) + \operatorname{cost}\left(\left\{\boldsymbol{x}_i, y_i : x_{ij} > t\right\}\right). \tag{2}$$

In the classification setting, a multinoulli model has to be fitted to the data in the leaf satisfying the test $X_j < t$ by estimating the class-conditional probabilities $\hat{\pi}_c = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbb{I}(y_i = c)$, where $\mathcal{D}$ is the data in the leaf. Given the class-conditional probabilities, we have used the Gini index [23] to evaluate the partition: $\sum_{c=1}^{C} \hat{\pi}_c (1 - \hat{\pi}_c) = \sum_C \hat{\pi}_c - \sum_c \hat{\pi}_c{}^2 = 1 - \sum_C \hat{\pi}_c{}^2$.

This index is the expected error rate $\hat{\pi}_c$ is the probability that a random entry in the leaf belongs to class c, and $(1-\hat{\pi}_c)$ is the probability that it would be misclassified. To prevent overfitting, we have stopped the growth of the tree performing a pruning. This is performed by using a scheme that prunes the branches giving the least increase in the error [6]. A problem introduced by using recursive partitioning procedure is the fact that trees are unstable. One way to reduce the variance of an estimate is to average together many estimates using the bagging (bootstrap aggregating) technique.

In the Random Forests approach [8] each tree is constructed using a different bootstrap sample from the original data. For each tree of the collection, a random subset of predictors is chosen to determine each split. In this way, the correlations between predictions of the individual trees are reduced. In other words, Random Forests try to decorrelate (each tree has the same expectation) the base learners by learning trees based on a randomly chosen subset of input variables, as well as a randomly chosen subset of data cases. In general, Random Forests procedure is better than bagging.

Stochastic Gradient Boosting [21] is another way to reduce the variance. The algorithm for Boosting Trees evolved from the application of boosting methods. Boosting method (Freund and Schapire [20]) fits many large or small trees to reweighted versions of the training data, and performs classifications by weighted majority vote. In Stochastic Gradient Boosting, many small classification (or regression) trees are built sequentially from "pseudo"-residuals (the gradient of the loss function of the previous tree). At each iteration, a tree is built from a random sub-sample of the dataset (selected without replacement) producing an incremental

improvement in the model. An advantage of Stochastic Gradient Boosting is that it is not necessary to pre-select or transform predictor variables. It is also resistant to outliers. In general, boosting procedure outperform the Random Forests.

In the multinomial approach, trees are formulated as statistical models, alike generalized linear and additive models [16]. In this approach, splits are based on an explicit statistical model, the deviance of which defines the dissimilarity measure. For classification trees the use of a multinomial model is equivalent to the information index, with the deviance defined by the multinomial log-likelihood.

5.2 The probability to be retweeted

By following the line of Suh et al. [55] and Naveed et al. [41], we have transformed the variable *Retweet Count* into a binary variable (0: no retweets, 1: one or more retweets). Suh et al., fitted a Generalized Linear Model (GLM) to 10 K dataset, and used the results in a logistic equation to predict the probability of a retweet. Naveed et al., trained a prediction model to forecast the likelihood, for a given tweet, of being retweeted based on its contents. From the parameters learned by the model, they deduced which are the influential content features that contribute to the likelihood of a tweet to be retweeted. Our aim is to evaluate the relevant metrics associated to the action of retweeting in a predictive perspective: we used a learning approach to predict the probability for a tweet to be retweeted. The binary classification model provides us a general picture of the most important features (Table 1) related to retweeting. Given the finding that some features have strong relationship associated with the *degree of retweeting*, we have fitted the predictive models, presented in Section 5, on 500 K dataset.

In order to verify and validate the learned model parameters, we measure the accuracy of retweet prediction. Therefore, we split the set of tweets into a training and a test set. We have used about 80% of data for the training set, and 20% for the validation set. According to the results reported in Table 4, Random Forests is the best model in terms of accuracy (91.5%) and F_1score (90.61%). *Mentions Count* is the most relevant metric associated to retweeting in Random Forests, Recursive Partitioning and Gradient Boosting, while *Favorites Count* is the second one in all three models. In Multinomial (Logistic) Model, *Favorites Count* is the most important metric, followed by *Mentions Count.*

5.3 Predicting the degree of retweeting of a tweet

For the analysis of collected tweets, we conducted a 10-fold cross-validation evaluation on the complete **100 Million dataset** and the features reported in Table 1. After the assessment of the above-mentioned approaches (as shown in the following), we have considered a CART model with Recursive Partitioning procedure (RPART model) as the best learning algorithm. In the

Table 4 Retweet binary classification models comparison on 500 K data

Classification methods	Accuracy	Precision	Recall	F_1score
Recursive partitioning	0.9071	0.9926	0.8157	0.8955
Random forests	**0.9150**	0.9826	0.8407	**0.9061**
Gradient boosting	0.9061	0.9936	0.8127	0.8941
Multinomial/Logistic model	0.9021	0.8115	0.9853	0.8899

next section, a comparison of the above-mentioned methods is provided. In the considered predictive models the response variable *Retweet Count* has been transformed in a categorical variable, namely *Retweet Class*, having classes: "0", "1–100", "101–1000", "1001–10,000", and "Over 10,000", with the evident meaning of classifying the *degree of retweeting*, in 0 retweets, from 1 to 100 retweets, etc. Please note that the chosen classes are different from those of Fig. 5. Actually, classes "1–10" and "11–100", as depicted in Fig. 5, have been merged into a single size class "1–100". In addition, we have created two new classes "1001–10,000" and "Over 10,000", with the aim of understanding the *degree of retweeting* especially when the retweet count is high. As it will be described in the following, compacting classes "1–10" and "11–100" allowed us to obtain a higher accuracy (a better prediction model).

Note that, the training set has been extracted as the 80% of 100 million data and the validation of the predictive capability has been performed on a test set of 20% of the total observations.

According to the RPART approach, the CART models use a two-stage procedure. The resulting model can be represented as a binary tree. It should be noted that the resulting quality of most of the machine learning techniques is highly dependent on the calibration parameters. In our model, no optional classification parameters are specified, the Gini rule has been used for the splitting [49], according to which the prior probability is proportional to the observed data frequencies and the 0/1 losses are used. We used a cross-validation to choose the best value for the complexity parameter (CP). The 1-SE rule has been used to find the lowest cross-validation error as the sum between the smallest cross-validation error and the corresponding standard error. The results of RPART model statistics by class and the overall statistics are reported in Tables 5 and 6, respectively. The resulting accuracy of the predictive model is 68.15% and the precision is 85.64%, obtaining a satisfactory model for predicting the *degree of retweeting*. The kappa coefficient suggests that the level of agreement between the raters is discrete (see Table 6). The balanced accuracy (see Table 5) is very high for the first two classes, while it tends to decrease with the increasing degree of the retweeting classes. The accuracy decrease is probably due to a lack of numerosity in the higher classes of retweet (Class: "1001–10,000", Class: "Over 10,000") (see Fig. 5). Moreover, very high numbers of retweets are sporadic to be obtained, depending on many other factors, and less interesting for advertising and day by day activity of Twitter users. In fact, only the 6% over 100 Million of tweets obtain more than 1000 tweets. Typically, advertising campaigns are grounded on a large number of former tweets that collected less than 1000 retweets each. The classification performed also allows identifying when a tweet has low or null probability to be retweeted.

Table 5 Predicting class of degree of retweeting of the RPART procedure

Assessment drivers	Degree of retweeting classes				
	0	1–100	101–1000	1001–10,000	Over 10,000
Sensitivity	0.7737	0.8105	0.3142	0.0208	0.0136
Specificity	0.9132	0.6694	0.9199	0.9996	1.0000
Positive predictive value	0.8564	0.6256	0.3752	0.7345	0.8488
Negative predictive value	0.8579	0.8382	0.8975	0.9485	0.9915
Prevalence	0.4007	0.4053	0.1328	0.0526	0.0086
Detection rate	0.3100	0.3285	0.0417	0.0011	0.0001
Detection prevalence	0.3620	0.5251	0.1112	0.0015	0.0001
Balanced accuracy	**0.8435**	**0.7399**	**0.6170**	0.5102	0.5068

Table 6 Overall statistics in predicting class of degree of retweeting

Assessment parameters	Values
Accuracy	0.6815
Accuracy 95% confidential interval (min, max)	(0.6813, 0.6817)
Recall	0.7737
Precision	0.8564
Kappa	0.4922

Figure 11 reports the features in order of importance in the prediction. The histogram suggests that the variable *Mentions Count* is the most correlated with the *degree of retweeting*. Furthermore, it has demonstrated to be the metric that better explains the volume of retweets. On the other hand, by eliminating the covariate *Mentions Count* from the model, the overall accuracy decreases to 0.5378, the precision to 0.5243, the recall equals to 0.6610 and Kappa index 0.2395. Table 7 reports the confusion matrix among the classes considered for the classification. From the table, it is possible to understand how well the first two classes have been identified.

6 Comparison among different approaches

The choice of the RPART model has been justified by the fact that the accuracy obtained was higher than other ensemble learning techniques as Random Forests, Stochastic Gradient Boosting and Penalized Multinomial Regression. The comparisons have been performed by using the datasets of 100 K and 500 K tweets, due to the computational costs of some of the compared algorithms. Moreover, the recursive partitioning procedure is also the result of a compromise between goodness in terms of accuracy, simplicity in terms of interpretation (each tree derives from a series of logical rules [47]) and the ability to take into account of millions of data within a reasonable timeframe.

Furthermore, RPART models can easily handle mixed discrete and continuous inputs, they are insensitive to monotone transformations of the inputs (because the split points are based on ranking the data points), they perform automatic variable selection, and they are relatively

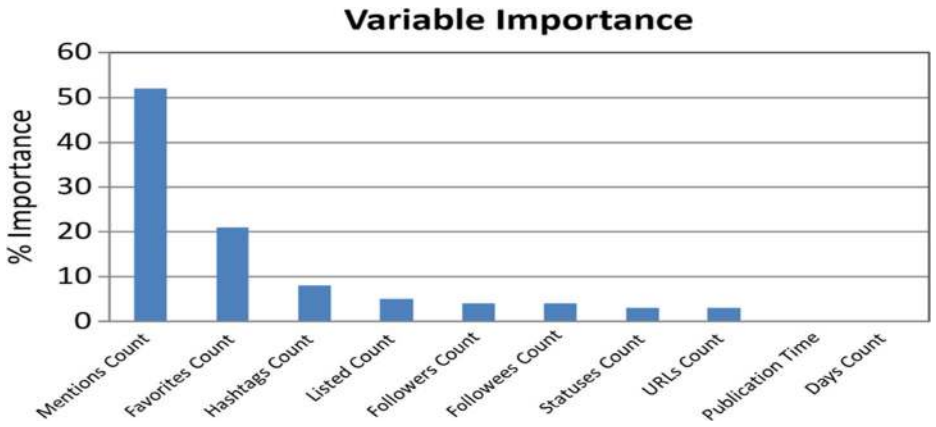


Fig. 11 Variable Importance from the RPART model

Table 7 Confusion matrix of the RPART procedure

Degree of retweeting classes	Reference degree of retweeting classes				
	0	1–100	101–1000	1001–10,000	Over10000
0	31.0009	4.7219	0.3055	0.1487	0.0232
1–100	7.3885	32.8530	8.7785	2.9702	0.5240
101–1000	1.6765	2.9545	4.1732	2.0247	0.2941
1001–10,000	0.0005	0.0055	0.0258	0.1092	0.0077
Over10000	0.0000	0.0000	0.0000	0.0021	0.0117

robust to outliers [40]. However, RPART model trees can produce models with high variance in the estimators. Two ways to reduce the variance of predictions could be adopted, for instance by using a bagging approach [7] or a boosting technique [48]: models like Random Forests often provide very good predictive accuracy. Actually, such an approach [8] aims at decorrelating the base learners by learning trees on the basis of a randomly chosen subset of input variables. Typically, the running time of classical Random Forests technique is not viable for millions of observations. On the other hand, applying it on a 100 K tweet dataset does not provided relevant improvements in term of accuracy with respect to the recursive partitioning procedure.

The F_1score has been used to measure the models performance, and four approaches have been followed to build the model. Table 8 presents the results of the classification model with Recursive Partitioning procedure (RPART), the Random Forests techniques, the Stochastic Gradient Boosting model and the Multinomial Regression model on 100 K observations dataset. Also in these cases, we have used about 80% of data for the training set, and 20% for the validation set. In the fourth column, the F_1score is reported. This is a measure to evaluate the robustness of a model for making predictions, as a compromise between precision and recall:

$$F_1 score = 2 \times (Precision \times Recall)/(Precision + Recall) \tag{3}$$

$$Precision = \frac{\#\text{tweets classified into class } i}{\#\text{tweets classified as class } i}, Recall = \frac{\#\text{tweets classified into class } i}{\#\text{tweets belonging to the class } i} \tag{4}$$

According to results reported in Table 8, the differences among the first three methods in terms of F_1score (3) are minimal. Moreover, we should remark that the *Mentions Count* is the most relevant metric in all the models. Then, the second more relevant metrics in the models are *Favorites Count* for Recursive Partitioning, *Hashtag Count* for Multinomial Model, *Followers Count* for Random Forests, and *Favorites Count* for Gradient Boosting (see

Table 8 Models comparison on 100 K observations. The recursive partitioning resulted as the better ranked in terms of accuracy

Classification methods	Accuracy	Precision	Recall	F_1score
Recursive partitioning	**0.6827**	0.8436	0.7806	0.8108
Random forests	0.6812	0.8509	0.7761	**0.8117**
Gradient boosting	0.6764	0.8547	0.7715	0.8110
Multinomial model	0.6480	0.8423	0.7275	0.7807

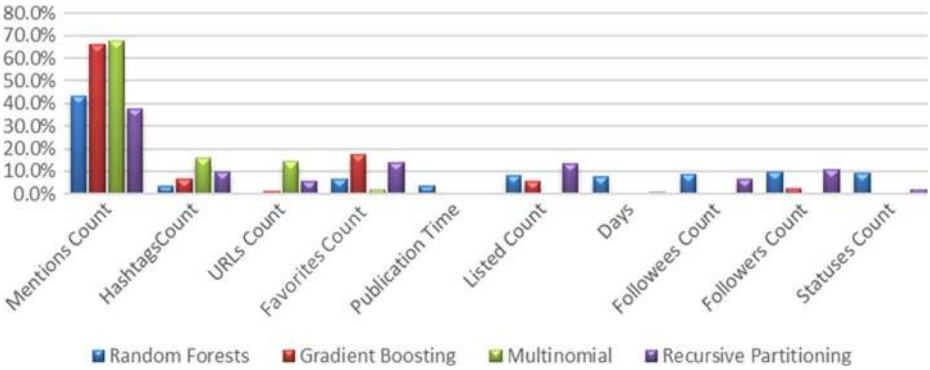


Fig. 12 Variable Importance between models on 500 K data

Fig. 12). Please note that the only first two metrics are the same in the RPART model on 500 K and RPART model on 100 M.

On the other hand, Table 9 shows the comparison among the models working on a 500 K dataset in terms of processing time for training. The higher value of overall accuracy among the models, as well as the constraint of working with millions of observations (which, consequently, conveys fast execution times as a requirement), have led us to choose the recursive partitioning technique as the better ranked (see Table 9). The experiments have been performed for the evaluation of the predictive models on a computational node with 98 GB Ram and 4 octa core CPUs (32 total cores, at 2.5 Ghz), using R which exploited only one core at time. Despite the lack of parallelization, the Recursive Partitioning approach resulted to be the most suitable to work on large datasets, as 100 M or more.

7 Conclusions and future perspectives

The work presented in this paper started with the aim of better understanding the correlation of features associated to tweets with respect to the action of retweeting. Most of the proposed papers in the literature proposed analysis without deriving models for predicting the degree of retweeting, in others they limited to identify the probability to be rewetted or not. The proposed analysis identified additional relevant metrics with respect to those proposed in the literature, namely, *Publication Time* and *Listed Count*. This approach resulted in obtaining a more effective principal component analysis and coverage of the phenomena. Therefore, on the basis of such an analysis, in this paper we proposed a method to predict the *degree of retweeting* through a classification trees model with recursive partitioning procedure applied on a dataset of 100 Million of tweets. We have shown that the choice of the RPART model is justified by the fact that the accuracy is better with respect to Random Forests, Stochastic

Table 9 Retweet models comparison on 500 K data in terms of computation time in model estimation

Classification methods	Accuracy	Precision	Recall	F_1score	Processing time in sec.
Recursive partitioning	0.6807	0.8512	0.7767	0.8122	180
Random forests	0.6884	0.8601	0.7866	0.8217	198,968
Gradient boosting	0.6796	0.8534	0.7731	0.8113	64,448
Multinomial model	0.6411	0.8367	0.7245	0.7765	31,576

Gradient Boosting and Penalized Multinomial techniques, compared on a viable sample of 100 K observations. The Recursive Partitioning procedure is the result of a compromise between goodness in terms of accuracy, simplicity in terms of interpretation and the ability to take into account millions of observations within a reasonable timeframe. By analyzing the results obtained with the Recursive Partitioning procedure, *Mentions Count* is the most correlated metric with the degree of retweeting, and the accuracy of the predictive model is about 68%.

The model produced can be used for assessing the degree of retweeting of each single tweet produced by some author or those prepared for advertising and/or for information campaign. Potential applications fields are many, including marketing and advertising, early monitoring, emergency response and, more generally, promoting and diffusing information; and the related raking and pricing of the actions performed in advertising. The work has been developed in the context of smart city projects in which the capacity of communicating information is fundamental for diffusing information about changes in the city, and/or directives for alerts of civil protection, as weather forecast, and in general for early warning, and thus for communicating. In fact, when a tweet is structurally more likely to be retweeted is more effective in propagating information.

As a perspective for future research, the analysis for predicting the *degree of retweeting* could be focused at a deeper and more specific level, for instance considering narrower domains (e.g., selecting tweets on the basis of their topics or subjects in terms of hashtags, as well as considering specific Twitter Vigilance channels) such as politics, healthcare, weather, healthcare, city services, emergency, etc. This could be made in order to understand if it is possible to identify more specific metrics and models, with respect to the ones analyzed in the present work, which could lead to higher values of prediction accuracy.

Acknowledgements This work has been supported by the RESOLUTE project (www.RESOLUTE-eu.org) and has been funded within the European Commission H2020 Programme under contract number 653460. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

References

1. Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B (2012) Twitter improves seasonal influenza prediction. Healthinf 61–70
2. Asur S, Huberman BA (2010) Predicting the future with social media. CoRR abs/1003.5699
3. Bermingham A, Smeaton A (2011) On using twitter to monitor political sentiment and predict election results. Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), Chiang Mai, Thailand, p 2–10
4. Bollen J, Mao H, Zeng XJ (2011) Twitter mood predicts the stock market. J Comput Sci 2(1)
5. Botta F, Moat HS, Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. Roy Soc Open Sci 2:150–162
6. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press
7. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
8. Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231

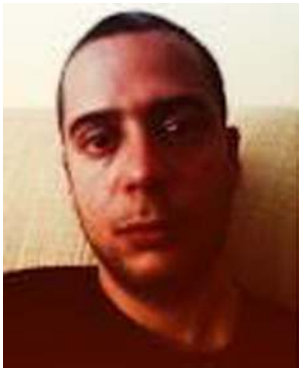
9. Bunyamin H, Tunys T (2016) A comparison of retweet prediction approaches: the superiority of Random Forest learning method. Telkonika (Telecommun Comput Electron Control) 14(3):1052–1058
10. Can EF, Oktay H, Manmatha R (2013) Predicting retweet count using visual cues. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, San Francisco, California (USA), p 1481–1484
11. Cattell RB (1966) The screen test for the number of factors. Multivar Behav Res 1(2):245–276
12. Cenni D, Nesi P, Pantaleo G, Zaza I (2017) Twitter Vigilance: a multi-user platform for cross-domain Twitter data analytics, NLP and sentiment analysis. IEEE international Conference on Smart City and Innovation, San Francisco, California (USA)
13. Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: the million follower fallacy. Proceedings of the International Conference on Weblogs and Social Media (ICWSM 10), Washington DC (USA), p 10–17
14. Chauhan A, Kummamuru K, Toshniwal D (2017) Prediction of places of visit using tweets. Knowl Inf Syst 50(1):145–166
15. Choi H, Varian H (2009) Predicting the present with Google Trends. Official Google Research Blog. Available at: http://bit.ly/h9RRdW
16. Clark LA, Pregibon D (1992) Tree-based models. In: Chambers JM, Hastie TJ (eds) Statistical models in S, Chapman & Hall/CRC, p 377–420
17. Crisci A, Grasso V, Nesi P, Pantaleo G, Paoli I, Zaza I (2017) Predicting TV programme audience by using Twitter based metrics. Multimed Tools Appl 1–30
18. Everitt B, Hothorn T (2011) An introduction to applied multivariate analysis with R. Springer Science & Business Media
19. Firdaus SN, Ding C, Sadeghian A (2016) Retweet prediction considering user's difference as an author and retweeter. Proceedings of the IEEE/ACM International Conference Advances in Social Networks Analysis and Mining (ASONAM), p 852–859
20. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning (ICML'96), Bari (Italy), p 148–156
21. Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367–378
22. Golder S (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: Proceedings of the 43rd International Conference on System Sciences Sciences (HICSS '10), Hawaii (USA), p 1–10
23. Gini C (1921) Measurement of inequality of incomes. Econ J 31(121):124–126
24. Grasso V, Zaza I, Zabini F, Pantaleo G, Nesi P, Crisci A (2016b) Weather events identification in social media streams: tools to detect their evidence in Twitter. PeerJ Preprints 4:e2241v1
25. Grasso V, Crisci A, Nesi P, Pantaleo G, Zaza I, Gozzini B (2016a) Public crowd-sensing of heat-waves by social media data. In: Proceedings of the 16th EMS Annual Meeting & 11th European Conference on Applied Climatology (ECAC), Trieste, Italy
26. Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. In: Proceedings of the 11th ACM International Conference on Knowledge discovery in data mining (SIGKDD), Chicago, Illinois (USA), p 78–87
27. Hansen LK, Arvidsson A, Nielsen FA, Colleoni E, Etter M (2011) Good friends, bad news - affect and virality in Twitter. CoRR, abs/1101.0510
28. Hong L, Dan O, Davison BD (2011) Predicting popular messages in Twitter. In: Proceedings of the 20th International Conference companion on World wide web (WWW), Hyderabad (India), p 57–58
29. Jansen B, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(1532):2169–2188
30. Jiang B, Liang J, Sha Y, Li R, Liu W, Ma H, Wang L (2016) Retweeting behavior prediction based on one-class collaborative filtering in social networks. In: Proceedings of the 39th ACM International Conference on Research and Development in Information Retrieval, Pisa (Italy), p 977–980
31. Jolliffe I (2002) Principal component analysis. John Wiley & Sons, Ltd
32. Kaiser HF (1960) The application of electronic computers to factor analysis. Educ Psychol Meas 20(1):141–151
33. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, New York, NY (USA), p 591–600
34. Lampos V, Bie TD, Cristianini N (2010) Flu detector - tracking epidemics on Twitter. Mach Learn Knowl 6323:599–602
35. Liu G, Shi C, Chen Q, Wu B, Qi J (2014) A two-phase model for retweet number prediction. In: Proceedings of the International Conference on Web-Age Information Management. Springer, Cham, p 781–792
36. Lu Y, Kruger R, Thom D, Wang F, Koch S, Ertl T, Maciejewski R (2014) Integrating predictive analytics and social media. In: Proceedings IEEE Conference on Visual Analytics Science and Technology (VAST), Paris (France), p 193–202

37. Madlberger L, Almansour A (2014) Predictions based on Twitter - a critical view on the research process. In: Processing of the International Conference on Data and Software Engineering (ICODSE), p 1–6
38. Mishne G, Glance N (2006) Predicting movie sales from blogger sentiment. In: Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI CAAW), p 155–158
39. Morchid M, Dufour R, Bousquet PM, Linarès G, Torres-Moreno JM (2014) Feature selection using principal component analysis for massive retweet detection. Pattern Recogn Lett 49:33–39
40. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press
41. Naveed N, Gottron T, Kunegis J, Alhadi AC (2011) Bad news travel fast: a content-based analysis of interestingness on Twitter. In: Proceedings of the 3rd ACM International Conference on Web Science Conference (WebSci), Koblenz (Germany)
42. Nesi P, Pantaleo G, Sanesi GM (2015) A Hadoop based platform for natural language processing of web pages and documents. J Vis Lang Comput 31:130–138
43. O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM), Washington, DC (USA), p 122–129
44. Pálovics R, Daróczy B, Benczúr AA (2013) Temporal prediction of retweet count. In: Proceedings of the IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), Budapest (Hungary), p 267–270
45. Peng HK, Zhu J, Piao D, Yan R, Zhang Y (2011) Retweet modeling using conditional random fields. In: Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), Vancouver, BC (Canada), p 336–343
46. Pezzoni F, An J, Passarella A, Crowcroft J, Conti M (2013) Why do I retweet it? An information propagation model for microblogs. In: Proceedings of the 5th International Conference on Social Informatics, Kyoto (Japan), 8238, p 360–369
47. Quinlan JR (1990) Learning logical definitions from relations. Mach Learn 5(3):239–266
48. Schapire RE, Yoav F (2012) Boosting: foundations and algorithms. MIT press
49. Shih YS (1999) Families of splitting criteria for classification trees. Stat Comput 9(4):309–315
50. Shimshoni Y, Efron N, Matias Y (2009) On the predictability of search trends. Available at: http://doiop.com/googletrends
51. Signorini A, Segre AM, Polgreen PM (2011) The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 6(5):1–10
52. Sikdar S, Adali S, Amin M, Abdelzaher T, Chan KL, Cho JH, Kang B, O'Donovan J (2014) Finding true and credible information on Twitter. In: Proceedings of the 17th IEEE International Conference on Information Fusion (FUSION), Salamanca (Spain), p 1–8
53. Sinha S, Dyer C, Gimpel K, Smith NA (2013) Predicting the NFL Using Twitter, arXiv:1310.6998v1
54. Sitaram A, Huberman BA (2010) Predicting the future with social media. Social Computing Lab, HP Labs, Palo Alto
55. Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: Proceedings of the 2nd IEEE International Conference on Social computing (SOCIALCOM), Washington, DC (USA), p 177–184
56. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the International Conference on Weblogs and Social Media, (ICWSM 10), Washington DC (USA), p 178–185
57. Uysal I, Croft WB (2011) User oriented tweet ranking: a filtering approach to microblogs. In: Proceedings of the 20th ACM International Conference on Information and knowledge management (CIKM), Glasgow, Scotland (UK), p 2261–2264
58. Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from Twitter posts. Social computing, behavioural-cultural modeling and prediction, p 231–238
59. Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in Twitter. In: Proceedings of the International Conference on Weblogs and Social Media, (ICWSM 10), Washington DC (USA), p 355–358
60. Zaman TR, Herbrich R, Van Gael J, Stern D (2010) Predicting information spreading in Twitter. In: Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds, NIPS
61. Zaman T, Fox EB, Bradlow ET (2014) A Bayesian approach for predicting the popularity of tweets. Ann Appl Stat 8(3):1583–1611
62. Zhang Q, Gong Y, Wu J, Huang H, Huang X (2016) Retweet prediction with attention-based deep neural network. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM), Indianapolis, Indiana (USA), p 75–84

Paolo Nesi is a full professor at the University of Florence, Department of Information Engineering, chief of the Distributed Systems and Internet Technology lab and research group. His research interests include massive parallel and distributed systems, physical models, semantic computing, object-oriented, real-time systems, formal languages, and computer music. He has been the general Chair of IEEE ICSM, IEEE ICECCS, DMS, WEDELMUSIC, AXMEDIS international conferences and program chair of several others. He is and has been the coordinator of several R&D multipartner international R&D projects of the European Commission such as RESOLUTE, ECLAP, AXMEDIS, WEDELMUSIC, MUSICNETWORK, MOODS and he has been involved in many other projects. He is the ICARO Cloud project coordinator. He has been co-editor of MPEG SMR.

Gianni Pantaleo has taken his degree on Computer Science from the University of Florence, and the PhD on Computer Science. Presently he is a researcher and aggregate professor of computer at the University of Florence, affiliated with the Distributed System and internet Technology Lab. His main competences are on signal processing, natural language processing, data analysis, audio processing, parallel architecture. He worked on a number of international research and development projects such as: IMAESTRO, AXMEDIS, ECLAP, Sii-Mobility, RESOLUTE.

Irene Paoli has taken her degree on Statistics from the University of Florence. Presently she is a PhD student on Information and Communication Technology at the University of Florence. Her main competences are on statistical analysis, predictive models and machine learning algorithms. At the Distributed System and internet Technology Lab her research interests are on statistical analysis of social media.



Imad Zaza has taken his degree on Computer Science from the University of Florence. Presently he is a PHD student on Information and Communication Technology at University of Florence. His main competences are systems and network administration and object-oriented programming. At the Distributed System and internet Technology Lab his research interest is distributed systems, data analysis, railways interlocking modelling and ontology engineering. He worked on a number of international research and development projects such as Trace-It, RAISSS, Sii-Mobility, RESOLUTE.