Assessing User Privacy on Social Media: The Twitter Case Study

GIOVANNI LIVRAGA, Università degli Studi di Milano, Italy
ALESSANDRO MOTTA, Università degli Studi di Milano-Bicocca, Italy
MARCO VIVIANI, Università degli Studi di Milano-Bicocca, Italy

At the time of writing, nearly four billion people worldwide employ social media platforms such as Facebook, Instagram, WeChat, TikTok, etc. to share content of various kinds, which may also include personal data. In addition to this, users interact with members of the virtual community, leaving behind important behavioral traces. In most cases, people do not have a full understanding of who will be able to access and use such a body of information, and for what purposes. Although social platforms provide users with some tools to protect their privacy, the very nature of these technologies and the psychological characteristics of users often lead them to ignore such solutions.

To address this issue, in this paper we aim to propose a model for assessing the privacy of users on social media by identifying the critical aspects associated with their content and interactions generated on such platforms. This model, in particular, considers distinct features, of different kinds, that capture the level of users' exposure with respect to privacy. These features, dropped into a vector space, are used to derive a score that expresses, in a measurable way, the privacy risk of users compared to the information available on social media about them. The proposed model is instantiated and tested on data collected from the microblogging platform Twitter, on which the results of the experimental evaluation are analyzed. Specifically, the model is tested by considering both a binary scenario, i.e., where users' privacy is evaluated as at risk or not, a multi-class scenario, i.e., where their privacy is evaluated against different risk ranges, and a ranking scenario, i.e., where the users are ranked according to their privacy assessment.

CCS Concepts: • **Security and privacy → Software and application security**; **Social network security and privacy**;

Additional Key Words and Phrases: Privacy, Confidentiality, Vector Space Model, Social Media

ACM Reference Format:
Giovanni Livraga, Alessandro Motta, and Marco Viviani. 2022. Assessing User Privacy on Social Media: The Twitter Case Study. In *Open Challenges in Online Social Networks (OASIS'22), June 28, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3524010.3539502

1 INTRODUCTION

Social media platforms are used every day to share content of multiple types through interactions of different nature with friends and acquaintances, but, in many cases, with unknown people. The reasons that push users to spread such a mass of content are varied, from meeting new people, to feeling an active part of a community [17], to increasing their social capital [23], etc. To achieve these disparate goals, people leave behind a significant amount of information that can affect their privacy, from their personal or sensitive data (e.g., date of birth, sentimental status, political and religious beliefs, sexual preferences, health data, family data, etc.), to the inevitable behavioral traces associated with social interactions. Sometimes this is exacerbated by the fact that, to use digital services and apps, users accept that this body of data is subjected to complex analyses for economic and commercial purposes. Users are often unaware of the

Manuscript submitted to ACM

exact use made of such data, having not read the privacy disclaimer due to either laziness or difficulty in understanding it [25]. In this scenario, therefore, a serious privacy issue arises, to which users may not give sufficient importance.

To tackle this issue, in this article we aim to propose a model that can be used to assess user privacy in the social media context by assigning a *privacy score* to users based on the analysis of privacy-critical aspects associated with the information available about them on a considered social platform. To define a comprehensive model, we identify several features that may contribute to that privacy score, ranging from the (personal) data that are disclosed by users on the social media platform, to their behavior in terms of engagement with the platform, to the network of users that may access their body of information. These features, whose associated values represent how much the user's privacy is more or less at risk with respect to each of them, come to form a *privacy risk vector* in an n-dimensional space, where n represents the number of the privacy features considered. The idea would be, at this point, to evaluate the distance between the privacy risk vector of each user and an *ideal privacy vector* modeling a (possibly fictitious) user who is fully protected, to obtain the privacy score. The greater the distance between the two vectors, the higher the privacy score must be, representing a greater privacy risk for the user. In reality, such a privacy risk vector could consist of a very large number of features, depending on the information available in social platforms and how one intends to represent it in the form of features; this could lead to difficulty in defining the ideal privacy vector, and in understanding the actual contribution of each individual feature considered with respect to the overall privacy risk. Firm therefore remaining the idea of the privacy risk vector, our model is based on the classification of the considered features into different *feature categories*, with respect to specific aspects that concern social platforms and their characteristics (e.g., the information contained in the user profile, the behavioral traces left by users, the content disseminated, etc.). In this way, within each category, the features are combined to produce a single privacy risk value for the category. These feature category values at this point come to constitute the *category-based privacy risk vector*, which represents a dense vector of dimension m, $m << n$, where m represents the number of categories considered. In this way, the ideal vector to be constructed will be more easily interpreted semantically, as will the privacy risk value associated with each category.

In order to implement and subsequently test the proposed model, we considered a case study constituted by the Twitter microblogging platform and the data available on it. The results obtained were evaluated with respect to three distinct tasks: the *binary classification* of users as *at risk* and *not at risk* with respect to their privacy; the *multi-class classification* of users as at *low risk*, *moderate risk*, *intense risk*, or *high risk*; the *ranking* of users with respect to their privacy scores. It was possible to carry out these assessments because human assessors were involved in the evaluation phase. They assessed target users against the information they released on Twitter to evaluate their exposure to privacy.

The reminder of the article is organized as follows: Section 2 discusses related works; Section 3 introduces and discusses the model proposed in this article, the high-level choices behind the building of privacy features in social media, their possible classification into feature categories, and the use of the *Vector Space Model* to generate an overall privacy score for each user; Section 4 illustrates the instantiation of our model on the Twitter microblogging platform; Section 5 discusses the experimental evaluation and analyzes the results obtained with respect to the considered case study; finally, Section 6 summarizes this work and illustrates some possible future research directions.

2 RELATED WORK

The line of work closest to ours is represented by models and approaches that aim at assessing users' privacy on social media with respect to the release of personal/sensitive data and other behavioral information released in the interactions on social platforms (e.g., [1, 5, 21]). One of the first approaches addressing this problem is the one proposed by Liu and Terzi [21], which is based on a combination of the *sensitivity* of data items and of their *visibility*, and leverages *Item*

Response Theory [11] for computing users' privacy scores. While sharing the same goal of our work of computing a score quantifying user privacy, their approach considers the disclosure entailed by the release of specific data items (which may resemble, to some extent, our profile-based and content-based features detailed in Section 3), while we also consider additional factors (i.e., the behavior-based and network-based features) that can impact the overall privacy assessment. The approach proposed in [1], focusing on the problem of evaluating the privacy level of a user across multiple social media platforms, builds on some of the notions introduced in [21], and does not accommodate our behavior-based and network-based features. In [5], the authors propose an approach based on *Natural Language Processing* solutions to build features that can identify whether a text contains private information. Similarly to our approach, this solution is instantiated on the Twitter case study, and produces a score that characterizes the privacy of Twitter users; unlike our work, though, it focuses on the presence of private information in textual content (i.e., tweets) while our approach is more general and considers also other aspects that can contribute to the overall privacy assessment.

Besides the quantification of the release of personal/sensitive information, the scientific community has investigated other related but different problems connected to modeling, quantifying and representing privacy-related risks that users may face when operating in social media. For example, the approach proposed in [36] addresses the problem of mapping users' preferences to the default privacy policy of social platforms. Similarly to our proposal, the approach described in [13] considers different aspects characterizing the release of information to online platforms, but aims at defining a scoring system for evaluating private data vulnerabilities in case of data breaches. In [2], the authors propose a solution for evaluating the risk entailed by establishing new interactions with social media users: while related, the problem addressed differs from ours and the solution is based on interactions with the users to consider their subjective assessments. The solution illustrated in [18] focuses on the related but orthogonal problems of identifying, characterizing, and assessing possible privacy violations in social media. Other approaches have investigated the possibility to support users in comprehending social media privacy policies and the corresponding visibility of their profiles (e.g., [24]) and to predict and suggest privacy policies (e.g., [12, 29, 30, 39]).

A related line of work has addressed different problems in the context of self-disclosure (verbal expressions by which individuals reveal aspects of themselves to others [4]), focusing for example on Twitter conversations (e.g., [3]), on the impact that the COVID-19 outbreak and its subsequent social distancing and quarantines had on self-disclosure (e.g., [32]), on the automatic identification of linguistic markers indicating the occurrence of self-disclosure (e.g., [15]), on the macro-societal factors that can contribute to private information disclosure and self-disclosure (e.g., [20]).

Our work is also related to privacy metrics, which quantify how much a privacy requirement is satisfied by a data collection (e.g., k-anonymity [28] and its variations [9], differential privacy [10]). Recently, k-anonymity has been investigated to quantify the risk that nodes of the published social network graph can be associated to specific users [7]. While related, these proposals focus on different problems and application scenarios [22, 34].

3 A MODEL FOR USER PRIVACY ASSESSMENT

In this section, we illustrate our general model for calculating a *privacy score* for distinct target users on social media. The model considers different *privacy features* characterizing users, built on top of their data and their interactions on social media platforms. We first discuss such features (Section 3.1) and then how these features can be formally represented and combined to compute the user's privacy score (Section 3.2).

3.1 Through the Identification of Privacy Features

We posit that a comprehensive assessment of a privacy score characterizing a user interacting with a social media platform should reflect and hence jointly consider, in a unified approach, a variety of aspects. For example, the privacy of a user u can be impacted by whether the date of birth u provided when registering to the social media platform is then made publicly available on u's profile, by the content (e.g., posts, images, and videos) u generates and publishes online, but also by the audiences that access u's data and information (e.g., u's friends and/or followers) and by u's personal behavior online (e.g., the frequency with which u comments and likes others' content). Restricting the consideration to specific aspects would produce only partial insights related to users' privacy (e.g., their attitude towards self-disclosure). Note that, in principle, all aspects that can contribute to a more comprehensive evaluation are dependent not only on the data and information implicitly and explicitly released by u on the social media platform, but also on the privacy policies and related actions of the social media platform, as well as by the data and information implicitly and explicitly released by the other users on the social media platform (e.g., by friends publishing contents related to u).[1]

Clearly, different social media platforms require and/or permit users to provide different kinds of data and information and to interact with the platform in different ways. To provide for a general model that can be instantiated on different platforms, we then propose to identify distinct *privacy features* that can be built on top of the data released by a user on a given social media platform as well as of u's interactions on it. For example, in Section 4 we illustrate a possible instantiation of the model on Twitter data, where natural features can model the size of u's network (e.g., the number of *followers*), the frequency with which u interacts with other users' content (e.g., the number of *likes* to tweets), or the release of personal/sensitive data in u's textual content (e.g., in the description and/or tweets). In general, privacy features can either directly correspond to the data of a user u in a social media platform, can be somehow derived from them, or can be somehow derived from the interactions of u on the social media platform. With reference to the example above, while u's network size can be directly available by counting u's friends, the release of personal/sensitive information requires some more elaboration as a sensitive data item may be released in different ways by u.

3.1.1 Feature classification. Our first contribution in this work is the definition of four main *feature categories*, based on different aspects related to the social media context, which can contribute to the privacy assessment of a user.

- *Profile-based features*: they model data that a user u discloses to the social media platform, possibly upon registration, which relate to u's profile. Features often available can model users' date of birth, gender, occupation, education level, and address. These data items are often visible in u's profile, and impact on u's privacy as they can typically be personal and/or sensitive. Some of these data items are required by some social media platform for accepting user registrations (e.g., Facebook requires to provide name, surname, date of birth, and telephone number), and other ones can be freely added. Often users tend to release more data than those strictly required by the platform to accept the registration: this can be easily noticed with a look at an average Facebook profile, where it is not uncommon to see users who publicly disclose heterogeneous and non-mandatory data;
- *Behavior-based features*: they model the privacy-relevant behavior of a user u in interacting with the social media platform. Possible features that fall in this category can model the number of *likes* assigned, or the number of created and commented posts. Intuitively, considering the features of this example, the higher their numerosity, the higher the possible impact on u's privacy, as more data points are available to an observer for inferring (personal and/or sensitive) characteristics of the user. Although more precise assessments can of course be made,

[1] In this work, in particular in the instantiation of the model on Twitter, we are not concerned with inferences that may be made with respect to personal data and information posted by others. Even assuming we have them available, however, it does not change the formal design of the model.

the Facebook-Cambridge Analytica scandal showed that the Facebook *likes* of a user can be easily mapped to the psychological traits of the user, and it has been shown that a few Facebook likes can be sufficient to infer several personal information pertaining to the liking user [19];

- *Network-based features*: they model information on the user's network. Depending on how specific the privacy assessment should be, following the same reasoning applying to counting the number of likes, they could be based on simple observations such as the size of the user's social network, as well as on more complex inferences on the network members, following the homophily theory observation that individuals are more likely to interact with individuals similar to themselves w.r.t. given characteristics or perceived qualities [26];
- *Content-based features*: they model the content published by the user on the social media platform, in textual and visual form such as photographs or video/audio content. Such content can indeed reveal personal/sensitive information about the user, and could reasonably be considered as the main gateway to disclosing personal/sensitive information, as also testified by established observations that have linked self-disclosure to several benefits perceived by the disclosing user [15].

3.1.2 Feature building. Once identified, the features identified as relevant to user privacy in the social media platform, along with a numerical assessment for each feature for each target user under analysis, are used to model the user's *privacy profile*.

Definition 3.1 (Privacy profile). Given a social media platform P, the set $F_P = \{f_1, \ldots, f_n\}$ of *privacy features* relevant to P, and a user u, the *privacy profile* $\Pi_{F_P}(u)$ of u w.r.t. F_P is a set $\{\langle f_1 : \phi_{f_1}(u)\rangle, \ldots, \langle f_n : \phi_{f_n}(u)\rangle\}$ of pairs, with $\phi_{f_i}(u) \in [0, 1]$ the *privacy risk value* of u w.r.t. f_i, $\forall f_i \in F_P$.

As it emerges from Definition 3.1, each *privacy feature* f_i relevant to a social media platform is associated, when building the *privacy profile* of a user u, with an assessment $\phi_{f_i}(u)$, namely a *privacy risk value*. For simplicity and to ensure consistent and comparable assessments for all relevant features, we define such an assessment in the $[0, 1]$ interval, with 0 the minimum assessment and 1 the maximum assessment. We interpret these values in the following way: the higher the assessment for a feature f_i, the more the privacy-connected risk for u connected with f_i, and vice versa. Given the extreme variability of data items available on social platforms, it is necessary to model the privacy features and their associated risk values with respect to the value domain of the data item on which the feature is built:

- Some data items are naturally defined on the *Boolean domain* (indicating whether the data item is released or not): in this case, the privacy risk value can be evaluated 1 if the data item is released, 0 otherwise;
- Other data items, such as the counts of *likes* and friends, are naturally defined on *numerical domains* that could exceed the $[0, 1]$ interval: in this case, such numerical data can be normalized to assume values in $[0, 1]$ to represent the privacy risk values to be associated with the features built on top of them. For example, the number of friends of user u could be normalized in $[0, 1]$ by scaling u's count with respect to the maximum number of friends of all social media platform users. Several normalization schemes can be used for this purpose, as illustrated in Section 5.3 w.r.t. the instantiation of our model on Twitter;
- In addition, other data items could be naturally defined on *non-numerical domains* and/or may require the application of some analysis or inference process to actually model their impact on the privacy assessment of the user. To this end, different approaches could then be used to compute the normalized privacy risk value associated with a specific feature. Our model can be instantiated with different metrics: for example, in the

considered Twitter case study, we leverage an approach to quantify the probability of a user's gender and age from the chosen nickname, and entity-based solutions to classify a user's declared address as existing or not.

Intuitively, the privacy profile $\Pi_{F_P}(u)$ can be used as a building block to assess the overall *privacy score* of u w.r.t. P. In the following section, we illustrate how privacy profiles can be used to compute such a score.

3.2 Privacy Score Assessment

Our proposal to evaluate for each target user a *privacy score* consists in quantifying the difference between the target user's privacy profile and an *ideal privacy profile*, modeling a (possibly fictitious) user who is scored 0 (i.e., who enjoys the lowest privacy-related risk) for all features.

Definition 3.2 (Ideal privacy profile). Given a social media platform P and the set $F_P = \{f_1, \ldots, f_n\}$ of features relevant to P, the *ideal privacy profile* $\Pi^*_{F_P}$ w.r.t. F_P is a set $\{\langle f_1 : \phi_{f_1}(*)\rangle, \ldots, \langle f_n : \phi_{f_n}(*)\rangle\}$ of pairs, with $\phi_{f_i}(*) = 0, \forall f_i \in F_P$.

Given a social media platform P, it is intuitive that the higher the difference between the privacy profile $\Pi_{F_P}(u)$ and the ideal privacy profile $\Pi^*_{F_P}$, the more the privacy-related risks to which u is subject based on the relevant $f \in F_P$.

To compute such a difference, we rely on the *Vector Space Model* (VSM) [27] and, hence, leverage a spatial representation of the user privacy profile and of the ideal privacy profile. In particular, we propose two possible representations. A first intuitive representation interprets a (user/ideal) privacy profile Π as a *privacy vector* in a n-dimensional space, with n the number of relevant features in F_P and each feature $f \in F_P$ representing a dimension in the space. The position of the privacy vector in the space is given, for each dimension (i.e., each privacy feature), by the privacy risk value associated in the profile with the feature.

Definition 3.3 (Privacy vector). Given a social media platform P, the set $F_P = \{f_1, \ldots, f_n\}$ of features relevant to P, and a user u with profile $\Pi_{F_P}(u)$ (Definition 3.1), the *privacy vector* $\mathbf{pv}_u$ of u is a vector $\mathbf{pv}_u = [f_1 \ \ldots \ f_n]$ such that $\forall f_i \in F_P : \mathbf{pv}_u[f_i] = \phi_{f_i}(u)$ in $\Pi_{F_P}(u)$.

Clearly, it follows that the *ideal privacy vector* $\mathbf{pv}^*$ for an ideal privacy profile Π^* will have all n privacy risk values for all features equal to 0 (i.e., $\mathbf{pv}^* = [0 \ \ldots \ 0]$).

A second representation interprets a (user/ideal) profile Π as a *catebory-based privacy vector* in an m-dimensional space, with one dimension for each of the m feature categories (as illustrated in Section 3.1.1, in this work we consider $m = 4$ categories). Each category-based privacy vector is then represented by m values, which are computed by combining the privacy risk values associated, in the privacy profile, to the privacy features belonging to the considered category. To combine them, it is possible, for example, to refer to a suitable *aggregation operator* [6].

Definition 3.4 (Category-based privacy vector). Given a social media platform P, the set $F_P = \{f_1, \ldots, f_n\}$ of features relevant to P, and a user u with profile $\Pi_{F_P}(u)$ (Definition 3.1), the *category-based privacy vector* $\mathbf{cpv}_u$ of u is a vector $\mathbf{cpv}_u = [c_1 \ \ldots \ c_m]$ such that: (*i*) c_i represents the i^{th} feature category; (*ii*) $\mathbf{cpv}_u[c_i] = \text{aggr}(\phi_{f_1}(u), \ldots, \phi_{f_k}(u))$, with $f_1, \ldots, f_k$ the features in F_P of category c_i, constitutes the *category-based privacy risk value*.

Like for ideal privacy vector, the *ideal category-based privacy vector* $\mathbf{cpv}^*$ for an ideal privacy profile Π^* will have all m values for all categories equal to 0 (e.g., for $m = 4$, as in our case, $\mathbf{cpv}^* = [0\ 0\ 0\ 0]$).

Both approaches for representing profiles as vectors have their pros and cons. The "flat" privacy vector representation (Definition 3.3) constitutes a direct mapping with features and associated privacy risk values constituting the user profile. However, when the number of features increases significantly, it can be complex for a human (as part of the

recent focus on the explainability of algorithms) to interpret such representation, due to the difficulties in visualizing n-dimensional spaces for large values of n. On the other hand, the category-based vector representation (Definition 3.4) is defined on $m << n$ dimensions only, which can help interpretation and visualization, but can clearly lose specificity in the mapping to the original profile, as some feature combination is needed beforehand.

The vector-based modeling of user privacy profiles permits a natural assessment of the *privacy score* of users entailed by their interactions with a social media platform based on the identified relevant privacy features. Recalling that we represent the privacy score of a user u based on the difference between u's profile and an ideal profile, a natural approach to compute such difference (and hence, a privacy score) consists in computing the *distance* between their vector representations. For simplicity and concreteness, we leverage the *Euclidean distance*, while noting that any distance (or, more generally, any similarity) notion could be applied.

Definition 3.5 (Privacy score). Given a social media platform P, the set $F_P = \{f_1, \ldots, f_n\}$ of features relevant to P, a user u with privacy profile $\Pi_{F_P}(u)$, a vector representation $\mathbf{v}_u$ of the privacy profile, and a vector representation $\mathbf{v}^*$ of the ideal privacy profile, the *privacy score* for u, denoted as $\text{score}_{F_P}(u)$, is defined as: $\text{score}_{F_P}(u) = \text{dist}(\mathbf{v}_u, \mathbf{v}^*)$, where $\text{dist}(\mathbf{v}_u, \mathbf{v}^*) = \sqrt{\sum_{i=1}^{|\mathbf{v}|}(\mathbf{v}_u[x_i] - \mathbf{v}^*[x_i])^2}$, and x_i is the i^{th} element in $\mathbf{v}$.

Clearly, the vector representation $\mathbf{v}_u$ of the privacy profile can be chosen between the privacy vector $\mathbf{pv}_u$ and the category-based privacy vector $\mathbf{cpv}_u$. The same holds for the choice of the vector representation for the ideal privacy profile (i.e., $\mathbf{v}^*$ either equal to $\mathbf{pv}^*$ or to $\mathbf{cpv}^*$).

4 INSTANTIATION OF THE MODEL ON THE TWITTER MICROBLOGGING PLATFORM

In this section we describe the instantiation of the proposed model, with respect to the scenario constituted by the Twitter microblogging platform. In particular, the data selection phase is described (Section 4.1), as well as feature building and feature classification into the four categories considered in this work (Section 4.2), and privacy score assessment (Section 4.3).

4.1 Twitter Data Selection

Twitter is the popular news and microblogging service provided by the Twitter, Inc. Company. In addition to representing a particularly fast and effective mode of communication, used by an increasing number of users, the Twitter platform proves particularly useful for data analysis purposes by providing public *Application Programming Interfaces* (APIs, more details of which will be provided in Section 5.1) for accessing and gathering data (either free with limited access, free for specific research purposes, or paid). This is another reason why we referred to such a platform to instantiate and test our model, in addition to the fact that a lot of personal data is spread on it. From an analysis of data that can be made freely accessible and downloadable by Twitter,[2] we selected the following *data items* to be used in our model, based on some consideration of their level of privacy risk.

- *Screen name*: this is the nickname that the user chooses during the registration phase. In many cases, users refer to their real name as the basis for their screen name, often in a shortened version. This can allow to infer personal information about the user, such as *gender* and *age*;

[2]https://developer.twitter.com/en/docs/twitter-api/data-dictionary/introduction

- *Location*: also this data item is provided during the registration phase; it is a non-mandatory field in which the user can write the name of a city, of a country, or an imaginary location. Providing this information can expose the geographical location of the user;
- *URL*: in most cases, the URL entered by the user refers to a personal Web page or to the profile of another social media platform. This data item permits to capture additional potentially sensitive and personal information from another source;
- *Protected*: this Boolean data item indicates whether the user has decided to keep the profile private. This means that the user's tweets will only be visible to the account's followers. A protected account does not mean that it does not contain sensitive information, but it certainly limits the viewing range;
- *Listed count*, *friends count*, and *followers count*: these data items represent the number of public lists of which the user is a member, the number of people the user follows, and the number of people who follow the user, respectively. All three data items provide an indication of the size of the user's network. In general, we can assume that the larger the size of the user's network, the greater the chance that potential personal data will be accessible to a wider audience, with consequent privacy issues;
- *Favorites count* and *statuses count*: the first data item represents the number of *likes* assigned by the user on the platform. It can be easily associated with the fact that the user has revealed a certain number of (personal) interests. The second data item represents the number of tweets written by the user. Also for this data item, the greater the number of tweets written, the higher the probability that more personal data have been released;
- *Default profile image*: this Boolean data item indicates if the user has changed or not the default profile image. Usually, profile pictures represent themselves or their family/friends, with obvious privacy impacts;
- *Description*: this data item is a text of up to 160 characters in which the user, while registering on the platform, often briefly describes interests, work activity, marital status, sex, age, etc. It is therefore clear that from this data item it it possible to infer a lot of personal and sensitive information.

4.2 Feature Building and Classification

We built on top of the considered data items the following *privacy features*, classified into the four categories illustrated in Figure 1. For each feature, it is described the way in which its privacy risk value is computed.

4.2.1 Profile-based features. They are built on the *screen name*, *location*, *URL*, *protected*, and *default profile image* data items, constituting the user's profile information, being created during the registration phase to the platform.

- *Screen name feature*: to infer personal/sensitive data from the *screen name* data item, we relied on the *M3Inference* tool [35],[3] which extracts the probability of user sex and user age. Formally, we computed the privacy risk value associated with this feature as: $r_{sn} = 0.5 * p(a) + 0.5 * p(b)$, where $r_{sn} \in [0, 1]$, and $p(a), p(b)$ are probability values of user sex and user age generated by M3Inference. The higher the r_{sn} value, the higher the privacy risk;
- *Location feature*: *Named Entity Recognition* (NER) was used to identify if the data item *location* contains a place that actually exists. The privacy risk value associated with this feature, namely r_l, assumes the value 1 if an existing location is released, 0 if no location or a location not corresponding to an existing place is released;
- *URL feature*: the privacy risk value associated with this feature, namely r_{url}, assumes the value 1 if the *URL* data item is populated, 0 otherwise;

[3] https://github.com/euagendas/m3inference

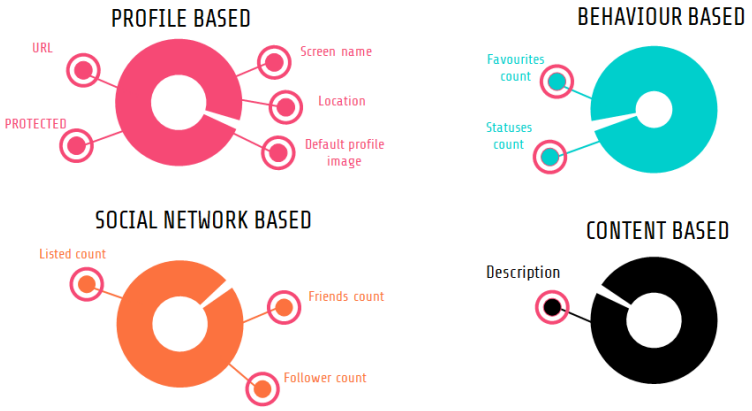

Fig. 1. Privacy features built on Twitter data items classified according to the four considered categories.

- *Protected feature* and *default profile image feature*: for the *protected feature*, the corresponding privacy risk value, namely r_p, assumes the value 1 if the user has the profile as public, i.e., visible to anyone and, hence, at risk (0 otherwise); regarding the *default profile image feature*, its corresponding privacy risk value, namely r_{dpi}, assumes the value 1 if the user has added a profile picture (0 otherwise).

4.2.2 Behavior-based features. The *favorites count feature* and the *statuses count feature* are built on top of the corresponding Twitter data items. In fact, both refer to a historical behaviors in user-platform interactions. Since these data items are already expressed in a numerical form in Twitter (e.g., 100 favorites assigned, 250 status updates), the privacy risk values of the features, namely r_{fc} and r_{sc}, are obtained by applying distinct normalization schemes (detailed in Section 5.3) in order to express them in the $[0, 1]$ interval. In this case, the closer the value is to 1, the greater the privacy risk, and vice versa, with respect to having provided more or less information about one's online behavior.

4.2.3 Network-based features. As in the previous case, the *listed count feature*, the *friends count feature*, and the *followers count feature* are built on top of the corresponding Twitter data items. They have been regrouped into this category as they indicate social interactions in the platform. Again, the privacy risk values to be associated with them, namely r_{lc}, r_{frc}, and r_{foc}, are computed in the $[0, 1]$ interval, and obtained by applying appropriate normalization schemes (again, detailed in Section 5.3) to the corresponding data items. Also in this case, the closer the value is to 1, the greater the risk for privacy and vice versa, with respect to having engaged more or less in social interaction within the network.

4.2.4 Content-based features. The single *description feature* is built on top of the corresponding Twitter data item. As previously introduced, it represents textual content possibly containing a lot of personal and sensitive information about the user. Unlike tweets that are only available for users who have a public profile, the description is always accessible.[4] Specifically, within this data item, we identified information that can be defined as sensitive data according to the GDPR [33].[5] By analyzing the description data items belonging to the dataset employed in this work (full information about the dataset will be illustrated in Section 5.1), ten *personal data items* were recognized to be the most frequently recurring: (1) *profession*, (2) *marital status*, (3) *interests/passions*, (4) *place of birth/residence/work*, (5) *age*, (6) *sexual*

[4] Also for this reason, in this work, we decided not to focus on the textual element of users' tweets, but only on the textual content of their description.
[5] The EU *General Data Protection Regulation* (GDPR) is generally considered among the most advanced and protective tools for regulating the processing of personal data, and indicates special categories of data that should be treated with particular care.

orientation, (7) *health status*, (8) *religion*, (9) *political opinions*, (10) *ethnicity*. Following the special categories of data needing special care in the GDPR, data items (5) – (10) can be considered particularly sensitive. For this reason, in computing the privacy risk value to be associated with the *description feature*, we performed a *linear combination* of the distinct privacy risk values associated with each distinct personal data item (1) – (10). First, personal data items were interpreted and evaluated with respect to privacy risk as follows: a value equal to 0 is assigned to it if the user's personal data item is not present in the text, and 1 otherwise. Formally, this means dealing with 10 privacy risk values, namely r_i ($i \in [1, \ldots, 10]$), to be later linearly aggregated by considering distinct weights ω_i ($i \in [1, \ldots, 10]$, $\sum \omega_i = 1$), associated with them, as: $r_d = \sum r_i \omega_i$. The idea is to weight more those personal data items that appear in the GDPR as sensitive data (details on the specific weights employed in this work will be provided in Section 5.3). Through this formalization, if all the personal data items are present in the description, the final r_d privacy risk value associated with the *description feature* is equal to 1, i.e., indicating the maximum risk. In the absence of personal data items, the privacy risk value of the feature is equal to 0. To detect personal data items in the description, distinct dictionaries have been considered. For the *profession*, *health status*, *religion*, and *ethnicity*, publicly available online resources were employed [8, 16, 37, 38]. For the other personal data items, ad-hoc dictionaries were generated, by a word-frequency and pattern analysis of the considered dataset.

4.3 Privacy Score Assessment

At this point, having available the privacy risk values associated with each feature considered in the instantiation of the proposed model on Twitter, it is possible to calculate the *privacy score* for the target user. We refer in this instantiation to the model based on the category-based privacy vector representation (Definition 3.4) of the user privacy profile. Hence, it is first necessary to aggregate the privacy risk values for each considered category through a suitable *aggregation operator*. For simplicity, and noting that other more complex operators may be used without impacting on our general model, we relied on the *average* operator. We must then construct an m-dimensional vector where m represents the number of categories considered (i.e., $m = 4$). This means, formally, that the four feature categories c_{pb} (profile-based), c_{bb} (behavior-based), c_{nb} (network-based), and c_{cb} (content-based), are assigned a *category-based privacy risk value* R_{pb}, R_{bb}, R_{nb}, and R_{cb}, where each $R_i, i \in \{pb, bb, nb, cb\}$, is computed as: $R_i = \frac{\sum_{j=1,\ldots,|F(c_i)|} r_j}{|F(c_i)|}$, where r_j is the privacy risk value of the j^{th} feature in the i^{th} feature category c_i considered, and $F(c_i)$ is the set of features in category c_i. At this point, the *category-based privacy risk vector* for the user u takes the following form: $\mathbf{cpv}_u = [R_{pb}\ R_{bb}\ R_{nb}\ R_{cb}]$ and must be compared with the ideal category-based privacy risk vector $\mathbf{cpv}^* = [0\ 0\ 0\ 0]$.

5 EXPERIMENTAL EVALUATION

In this section we detail the experimental evaluation carried out on the Twitter case study. For this reason, we first describe the data employed for the experimentation, explain some technical information about their gathering and pre-processing, and illustrate an initial exploratory analysis of the characteristics of these data (Section 5.1). Next, we turn to a description of the actual evaluation (Section 5.2) and a discussion of the results obtained (Section 5.3).

5.1 The Twitter Dataset

The purpose of the work is to assess the privacy of *target users* on social media. Therefore, the dataset that was generated for this purpose started with the identification of a set of users whose data was collected over a given period

of time. To proceed with *data gathering*, we established a connection with the Twitter Streaming API,[6] by means of the open-source *Tweepy*[7] Python library. User data was collected by considering some general-purpose hashtags in English, such as: *#today*, *#love #sport*, *#actuality*, *#fashion*, *#beautiful*, *#art*, *#photography*, *#happy*, *#summer*, *#friends*, *#life*, *#music*, *#motivation*, over a time-period of 3 weeks. It was chosen to use non-specific hashatags, such as hashtags concerning specific events or specific subjects, in order to collect heterogeneous users. Thanks to this method, we first collected the data of 3,500 users. Starting with this initial pool of users, we downloaded their follower data, coming up with a total of about 300,000 users. At this point, 1,500 users were randomly selected for evaluation purposes and human assessment (explained later in Section 5.2.1). Specifically, in selecting these 1,500 users, we avoided considering *public figure profiles* (e.g., celebrities) and outliers relative to the value of the data items we considered. In this work, it was chosen to eliminate 5% of the top extreme values. This represents a simple solution to eliminate in a rather coarse way potential profiles belonging to *bots* (more refined solutions dedicated to this purpose may be evaluated in the future).

A *pre-processing* phase was performed on these data, w.r.t. specific data items. In particular, a cleaning process has been performed on the *location* and *description* data items. Concerning *location*, being a non-mandatory field in which the user can add symbols, emoticons, grammatically incorrect places, or non-existent places, we have lowercase characters and we removed all special characters and punctuation symbols. In addition, *emojis*, which were very frequent, have been eliminated. The same cleaning process has been performed on *description*.

In the final dataset, *screen name*, *protected*, *followers count*, *friends count*, *listed count*, *statuses count*, *favorites count*, and *default profile image*, have 100% non-null values. Concerning the other data, *location* has 33% null values, *description* has 23% null values, *URL* has 79% null values. Concerning the distribution of values for Boolean data, *protected* accounts are the 3%, and the accounts with the *default profile image* are 60%.

5.2 Evaluation Scheme

In order to assess the effectiveness of the proposed model, we evaluated the goodness of fit of the automatically estimated privacy risk against the risk estimated by human assessors who analyzed the considered user data.

5.2.1 Human Assessment. Human assessors were asked to rate how much, in their opinion, users in the dataset release information that can put their privacy at risk on the platform. Each assessor was asked to evaluate 100 users. Before making the assessment, each assessor was briefly briefed on the contents of the GDPR regarding the definition of sensitive data and personal data. The assessor, in order to provide a privacy label concerning the privacy of users, had to analyze in detail the account of each Twitter user assigned. A number of 15 assessors were chosen, and they were selected so that there were people from different genders and age groups; in fact, it is possible to hypothesize that different age groups have different sensitivities to the concept of privacy. For example, it could be noted that digital natives are more likely to release sensitive information, and therefore consider the disclosure of some of this information less at risk. Each of the assessors was asked to rank each user against four privacy risk degrees, namely: *low risk*, *moderate risk*, *intense risk*, and *high risk*.

5.2.2 Evaluation Tasks and Metrics. Having the labels provided by the human assessors, it was decided to evaluate the effectiveness of the model with respect to two classification tasks, i.e., (*i*) *multi-class classification* and (*ii*) *binary classification*, and (*iii*) with respect to a task of *ranking* the results against the automatically generated privacy scores. To evaluate the effectiveness of the model with respect to the two classification tasks, standard evaluation metrics

[6] https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time
[7] https://www.tweepy.org/

Table 1. Evaluation results with respect to multi-class and binary classification, in terms of accuracy and f-score, with respect to the three normalization schemes used.

Model configuration	Accuracy	F-score
MC(min-max)	0.41	0.40
MC(vector)	0.42	0.42
MC(interval)	0.45	0.44
BC(min-max)	0.66	0.65
BC(vector)	**0.75**	**0.73**
BC(interval)	0.70	0.69

belonging to the set-based family, namely *accuracy* and *f-measure*, were used. These measures represent a standard for assessing the effectiveness of classification tasks in the literature [31]. To evaluate the model w.r.t. the ranking task, the *Normalized Discounted Cumulative Gain* (NDCG) measure [14], belonging to the family of rank-based metrics, was used. This measure is commonly employed to evaluate the effectiveness of Information Retrieval Systems in producing a ranking list of relevant documents w.r.t. a query. The goodness of the produced ranking is evaluated with respect to an ideal ranking of the documents, in which the most relevant ones should be in the highest positions and gradually the less relevant ones in the lower ones. In our case, a ranking is produced based on the privacy scores associated with the users. This ranking is evaluated with respect to the ideal ranking represented by the users ranked on the basis of the labels assigned to them by the human assessors. The smaller the difference between the ranking produced from the model and that ideal one, the greater the effectiveness of the model and vice versa. The three evaluation measures were calculated in this work through the use of the `sklearn.metrics` Python library.[8]

5.3 Evaluation Results

The results of several *evaluation configurations* conducted against the tasks and measures detailed above are provided below. Three evaluation configurations were first assessed against the multi-class classification, each configuration against a given *data normalization scheme*. As anticipated in Section 3, to carry out the evaluations, we considered distinct types of normalization, three in particular (i.e., the *min-max* normalization, *vector* normalization, and *interval* normalization). For this reason, the first three evaluation configurations assessed are denoted as **MC(min-max)**, **MC(vector)**, and **MC(interval)**, and, in the computation of the distance between the category-based privacy vector and the ideal vector, use equal weights assigned to each category. Next, three other evaluation configurations were assessed against the binary classification, again one configuration against each normalization scheme. These configurations are denoted as **BC(min-max)**, **BC(vector)**, and **BC(interval)**, and again use equal weights assigned to each category. The results of these six evaluation configurations are illustrated in Table 1. Note that the results in this (and the following) table were obtained by assigning higher weights to the personal data items (5) – (10) illustrated in Section 4.2.4 in the computation of the privacy risk value for the *content-based feature*.

Finally, different evaluation configurations were considered with respect to the ranking task. The NDCG results with respect to the rankings obtained using the different normalization schemes are illustrated in Table 2. In particular, they are denoted as **R(min-max)**, **R(vector)**, and **R(interval)**.

[8] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

Table 2. Evaluation results, with respect to the ranking task, in terms of NDCG.

Model configuration	NDCG
R(min-max)	0.939
R(vector)	0.942
R(interval)	0.945

5.4 Discussion

From the obtained results, several considerations can be made. Not surprisingly, the results of the binary classification are always better than those obtained with respect to the multi-class classification. This is due to the fact that, in the process of assigning a user to a class, it is evaluated in the same way to have mistakenly assigned a user labeled as *low risk* to the class *moderate risk* or to the class *high risk*, while the first case should be evaluated in a less severe way. In the binary classification, acting on only two classes, where *low risk* and *moderate risk* were considered equivalent classes, as well as *intense risk* and *high risk*, this problem is blunted. However, given the use of a four-risk-value scale, in which belonging to one of the four values very much depends on the perception of the human assessors who performed the labeling, we believe it is more correct to estimate the effectiveness of the model by referring to the task of ranking the users according to the risk score obtained and comparing it with the ideal ranking constituted by the labels provided by the human assessors. In this case, by means of the rank-based NDCG measure, we can observe how, in effect, the proposed method achieves excellent results in identifying users estimated by assessors w.r.t. their privacy risk. In general, we can also say that the choice of a particular normalization scheme does not appreciably affect the results obtained.

6 CONCLUSIONS AND FURTHER RESEARCH

In this article, we have proposed a model for assessing the privacy of users on social media. Our model considers several characteristics that may impact a privacy score of the users of a social media platform, ranging from the (personal) data that are disclosed by them on the platform, to their behavior in terms of engagement with the platform, to the network of users that may access their body of information. To this end, we have identified different categories of features that can be built on top of the data related to users that is accessible from a social media platform. We defined a metric for assigning scores to users based on their evaluated privacy risk level. To test the effectiveness and applicability of our approach, we have instantiated our model on the Twitter microblogging platform, and we have reported the results of a series of experiments demonstrating that our model achieves good effectiveness.

The approach proposed in this article is in some ways to be considered preliminary and can be enriched with the consideration of other aspects. Among these, we plan to study at a higher level the composition of individual feature classes and their impact in generating the final privacy score, the use of a less compensatory approach in aggregating the values obtained for each feature class, and the evaluation of the proposed approach on other social platforms and datasets.

ACKNOWLEDGMENTS

This work was supported in part by the EC within the H2020 Program under project MARSAL, and by the Italian Ministry of Research within the PRIN program under project HOPE.

REFERENCES

[1] E Aghasian, S Garg, L Gao, S Yu, and J Montgomery. 2017. Scoring users' privacy disclosure across multiple online social networks. *IEEE Access* 5 (2017), 13118–13130.

[2] C Akcora, B Carminati, and E Ferrari. 2012. Privacy in social networks: How risky is your social graph?. In *Proc. of ICDE 2012*. 9–19.

[3] J Bak, C-Y Lin, and A Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proc. of EMNLP 2014*. 1986–1996.

[4] A Barak and O Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior* 10, 3 (2007), 407–417.

[5] A Caliskan Islam, J Walsh, and R Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proc. of WPES 2014*. 35–46.

[6] T Calvo, G Mayor, and R Mesiar. 2002. *Aggregation operators: new trends and applications*. Vol. 97. Springer Science & Business Media.

[7] J Casas-Roma, J Herrera-Joancomartí, and V Torra. 2017. A survey of graph-modification techniques for privacy-preserving on networks. *Artificial Intelligence Review* 47, 3 (2017), 341–366.

[8] cdc.gov. 2022. List of all diseases. https://www.cdc.gov/diseasesconditions/az/a.html. [Online; accessed 2-March-2022].

[9] S De Capitani di Vimercati, S Foresti, G Livraga, and P Samarati. 2012. Data Privacy: Definitions and Techniques. *IJUFKBS* 20, 6 (2012), 793–817.

[10] C Dwork. 2006. Differential Privacy. In *Proc. of ICALP 2006*. 1–12.

[11] S E Embretson and S P Reise. 2013. *Item response theory*. Psychology Press.

[12] L Fang and K LeFevre. 2010. Privacy wizards for social networking sites. In *Proc. of WWW 2010*. 351–360.

[13] Z Foreman, T Bekman, T Augustine, and H Jafarian. 2019. PAVSS: Privacy Assessment Vulnerability Scoring System. In *Proc. of CSCI 2019*. 160–165.

[14] D Harman. 2011. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3, 2 (2011), 1–119.

[15] D J Houghton and A N Joinson. 2012. Linguistic markers of secrets and sensitive self-disclosure in Twitter. In *Proc. of HICSS 2012*. 3480–3489.

[16] joblist.com. 2022. List of all jobs. https://www.joblist.com/b/all-jobs. [Online; accessed 2-March-2022].

[17] K Kircaburun, S Alhabash, Ş B Tosuntaş, and M D Griffiths. 2020. Uses and gratifications of problematic social media use among university students: A simultaneous examination of the Big Five of personality traits, social media platforms, and social media use motives. *International Journal of Mental Health and Addiction* 18, 3 (2020), 525–547.

[18] N Kökciyan and P Yolum. 2016. PriGuard: A semantic approach to detect privacy violations in online social networks. *IEEE TKDE* 28, 10 (2016), 2724–2737.

[19] M Kosinski, D Stillwell, and T Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *PNAS* 110, 15 (2013), 5802–5805.

[20] H Liang, F Shen, and K-w Fu. 2017. Privacy protection and self-disclosure across societies: A study of global Twitter users. *New Media & Society* 19, 9 (2017), 1476–1497.

[21] K Liu and E Terzi. 2010. A framework for computing the privacy scores of users in online social networks. *ACM TKDD* 5, 1 (2010), 1–30.

[22] G Livraga and M Viviani. 2019. Data Confidentiality and Information Credibility in Online Ecosystems. In *Proc. of MEDES 2019*. 191–198.

[23] P Matthews. 2016. Social media, community development and social capital. *Community Development Journal* 51, 3 (2016), 419–435.

[24] A Mazzia, K LeFevre, and E Adar. 2012. The PViz comprehension tool for social network privacy settings. In *Proc. of SOUPS 2012*. 1–12.

[25] A M McDonald and L F Cranor. 2008. The cost of reading privacy policies. *ISJLP* 4 (2008), 543.

[26] M McPherson, L Smith-Lovin, and J M Cook. 2001. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* 27, 1 (2001), 415–444.

[27] G Salton, A Wong, and C-S Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18, 11 (1975), 613–620.

[28] P Samarati. 2001. Protecting Respondents' Identities in Microdata Release. *IEEE TKDE* 13, 6 (2001), 1010–1027.

[29] A Squicciarini, S Karumanchi, D Lin, and N DeSisto. 2014. Identifying hidden social circles for advanced privacy configuration. *COSE* 41 (2014), 40–51.

[30] A Squicciarini, S Sundareswaran, D Lin, and J Wede. 2011. A3P: Adaptive policy prediction for shared images over popular content sharing sites. In *Proc. of HT 2011*. 261–270.

[31] A Tharwat. 2020. Classification assessment methods. *Applied Computing and Informatics* (2020).

[32] P Umar, C Akiti, A Squicciarini, and S Rajtmajer. 2021. Self-disclosure on Twitter During the COVID-19 Pandemic: A Network Perspective. In *Proc. of ECML–PKDD 2021*. 271–286.

[33] P Voigt and A Von dem Bussche. 2017. *The EU general data protection regulation (GDPR)*. Springer.

[34] I Wagner and D Eckhoff. 2018. Technical privacy metrics: A systematic survey. *ACM CSUR* 51, 3 (2018), 1–38.

[35] Z Wang, S Hale, D I Adelani, P Grabowicz, T Hartman, F Flöck, and D Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *Proc. of WWW 2019*. 2056–2067.

[36] J Watson, H R Lipford, and A Besmer. 2015. Mapping user preference to privacy default settings. *ACM TOCHI* 22, 6 (2015), 1–20.

[37] Wikipedia contributors. 2022. List of contemporary ethnic groups — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_contemporary_ethnic_groups&oldid=1074262633. [Online; accessed 2-March-2022].

[38] Wikipedia contributors. 2022. List of religions and spiritual traditions — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_religions_and_spiritual_traditions&oldid=1074913144. [Online; accessed 2-March-2022].

[39] M Yang, Y Yu, A K Bandara, and B Nuseibeh. 2014. Adaptive sharing for online social networks: A trade-off between privacy risk and social benefit. In *Proc. of TrustCom 2014*. 45–52.