Unveiling the Privacy Risk: A Trade-off between User Behavior and Information Propagation in Social Media

Giovanni Livraga[1], Artjoms Olzojevs[2], and Marco Viviani[2]*

[1] University of Milan, Computer Science Department
Via Celoria, 18 – 20133 Milan, Italy
giovanni.livraga@unimi.it
[2] University of Milano-Bicocca
Department of Informatics, Systems, and Communication
Edificio U14 (ABACUS), Viale Sarca, 336 – 20126 Milan, Italy
a.olzojevs@gmail.com, marco.viviani@unimib.it

Abstract. This study delves into the privacy risks associated with user interactions in complex networks such as those generated on social media platforms. In such networks, potentially sensitive information can be extracted and/or inferred from explicitly user-generated content and its (often uncontrolled) dissemination. Hence, this preliminary work first studies an unsupervised model generating a privacy risk score for a given user, which considers both sensitive information released directly by the user and content propagation in the complex network. In addition, a supervised model is studied, which identifies and incorporates features related to privacy risk. The results of both multi-class and binary privacy risk classification for both models are presented, using the Twitter platform as a scenario, and a publicly accessible purpose-built dataset.

Keywords: Complex networks, user privacy, user behavior, user-generated content, information propagation, social media

1 Introduction

Social media platforms have become a part of everyday life, enabling users to share various types of content and engage in diverse interactions with friends, acquaintances, and even strangers, in the complex networks that are generated on such platforms. The motivations driving this extensive content generation and sharing range from socio-psychological reasons, e.g., expanding social connections to feeling a sense of community [17] and boosting *social capital* [21], to "practical" and commercial purposes for using digital services and apps [24]. However, this widespread sharing exposes users to potential privacy risks as they leave behind a wealth of personal and sensitive information, such as birth dates, relationship status, political and religious beliefs, sexual preferences, health data,

* Corresponding author.

and family details. Additionally, the traces of their social interactions can further contribute to this information disclosure [5]. Unfortunately, many users remain unaware of how their data is precisely utilized, often due to negligence or difficulty understanding privacy disclaimers [22], and monitoring its diffusion on the network [23].

With the aim of increasing the user's awareness regarding the privacy risk connected to the extent of sensitive information they disclose, the proposed study tackles several challenges. These include the identification of sensitive information from user-generated content, the consideration of sensitive information propagation through social network connections, and the extraction and/or generation of suitable features that can be interpreted in terms of privacy risk. Taking these aspects into account, the study aims first to develop an *unsupervised model* to generate a *privacy risk score* related to the disclosure and dissemination of sensitive information on social media complex networks. In particular, this score is obtained by combining two distinct scores for each user. A first score considers the information released directly by the user and that involving the user released by other members of the social network, while a second score accounts for information propagated in the user's social circle. Additionally, a *supervised model* employing distinct *privacy-risk features* is proposed. These features are constructed to take into account the same privacy aspects as the unsupervised model. Finally, a comparison between the results obtained by the two models is conducted. Twitter is considered the target social media platform, analyzing users' tweets, the connections between their tweets, and the propagation level of users' tweets on the network. A dataset is built and made accessible to the wider research community, enabling further studies and advancements in the field of privacy risk assessment on social media platforms.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 presents the proposed unsupervised and supervised models, along with details on the construction of the labeled dataset employed to instantiate them. Section 4 illustrates the results of our experimental evaluation. Finally, Section 5 concludes the paper and highlights some potential further research.

2 Related Work

Our work is closely related to a research line investigating approaches evaluating and quantifying the potential privacy risk for users caused by their participation in online social media communities [11]. Among the first to address this problem, Liu and Terzi [19] assign a *privacy score* to users considering, in combination, the *sensitivity* of user data and their *visibility* on the social platform. In our work, we build on a similar idea (higher risks for users derive from more visibly releasing more sensitive information), but we explicitly consider unstructured textual content generated by the user (while [19] considers just user profile information such as name, email, and hometown), as well as the impact on privacy risk that content released by other users can have. The consideration of both sensitivity and visibility of released content is also pursued in [2], which, however, does not propose approaches for obtaining the specific data/content to be used for the privacy assessment, while we leverage NLP to identify and extract sensitive information. NLP is performed to extract sensitive information from tweets in [4], but the work only considers a supervised model (we also consider a non-supervised one) to assess privacy risks. Our work is also partly inspired, concerning the definition of sensitive information categories and the general problem of assessing privacy risks in social media, by another of our previous works [20]. In that work, however, we focused on user profile information and did not consider the textual content shared by users, nor the potential scope that such content may undergo, which are instead a key contribution of the present article.

Other literature work has investigated related yet orthogonal issues, which for this reason are just listed in this section. They include studies on *privacy policies* (e.g., [25]), *privacy risks* entailed by establishing *new relationships* such as friendships (e.g., [3]), *data breaches* and *privacy violations* (e.g., [12, 18]), *privacy metrics* (e.g., [8, 9]).

3 Privacy Risk Assessment of Users

In this section, we present the two models proposed in this study for user privacy risk assessment. Firstly, we introduce the *unsupervised model*, which aims to identify patterns in the data without relying on pre-existing labels. Next, we delve into the *supervised model*, which utilizes the labels provided by human assessors to train three distinct classifiers. For this reason, we begin by introducing the *data* on which we instantiated both models, sourced from Twitter (prior to its rebranding to $\mathbb{X}$),[3] along with details about the *data labeling* process.

3.1 The Twitter Dataset and the Labeling Process

The Dataset. The dataset construction started with the identification of some *trending topics* from Twitter in December 2022 (the period in which this study was carried out). These trending topics encompassed various subjects, including the 2022 World Cup (`#fifa`, `#argentina`), technology (`#musk`, `iphone`), online communities (`#socialnetwork`), entertainment (`#netflix`, `#amazon`, `#disney`), public health (`#covid`), job opportunities (`#job`), political debates (`#politics`), conflicts (`#war`), religious themes (`#religion`), environmental sustainability (`#sustainability`), distinct aspects related to Sundays (`#sunday`), motivational content for Mondays (`#mondaymotivation`), the month itself (`#december`), festive season (`#christmas`), and general well-being (`#happiness`). Additionally, the approaching year was also considered among the trending topics (`#2023`). From those users discussing such trending topics, we randomly selected 100 real users (no spam profiles, no private profiles, no company profiles), 5 users for each topic. Subsequently, for each target user (referred to as user u for convenience),

[3] https://www.nytimes.com/2023/08/03/technology/twitter-x-tweets-elon-musk.html, accessed on September 1, 2023.

we downloaded up to 80 tweets directly from u, and 20 from other users who mentioned u using the `@` symbol (e.g., `hello @u!)` in their tweets. This to take into account the potential disclosure of user u's personal information by other users. In the end, a total of 9,210 tweets were collected for the 100 considered users (for some users it was not possible to download a number of tweets equal to 80). In addition to the textual content of the user's tweets, other related data and metadata, illustrated in Table 1, have been considered.

Table 1. Attributes and related data/metadata downloaded for each user.

Attribute	**Type**	**Type**
User	`string`	User u's username.
Tweet	`string`	The content of u's tweet. One user can post several tweets. Each tweet can contain up to 280 characters.
Biography	`string`	The designated section where u can provide a brief textual biography. This section is optional, and for certain users, it may remain empty, resulting in a `null` value.
Geolocation	`string`	The designated section where u can input their geolocation, such as a city, region, or State, representing their presumed place of birth or residence. This section is optional, and some users may leave it blank, resulting in a `null` value.
Followees	`integer`	The count of users followed by user u.
Followers	`integer`	The count of users who follow user u.
Likes	`integer`	The count of likes (or favorites) received by a given tweet.
Replies	`integer`	The total count of replies (or comments) on a given tweet.
Retweets	`integer`	The count of retweets on a given tweet.

The Labeling Process. Twelve human assessors were tasked with assessing the privacy risk associated with the considered users in the dataset thus constructed. Each assessor was well-informed about the potential risks arising from sharing sensitive information on social media platforms and was familiar with the types of information considered sensitive (more details about this information are provided in Section 3.2). The assessors were carefully chosen to represent various professional fields, ensuring a balanced representation of gender (seven men and five women), and encompassed a wide age range from 20 to 70 years. Each assessor was assigned randomly selected Twitter user profiles to analyze. Assessors were required to gauge the risk for each user based on reading at least the user's 50 most recent tweets and considering the interactions with those tweets and possibly the other attributes related to the user. For each of the 100 Twitter users considered, the goal was to obtain five distinct privacy risk assessments. The privacy risk assessment was initially conducted using a multi-graded scale, i.e., $1-3$, where 1 denotes "Not at Risk", 2 "Partially at Risk", and 3 "At Risk". In cases where there was no *majority agreement* among the five assessments, extra evaluations were required from assessors who had not participated in the initial assessment for that user. Subsequently, the same assessors for each user were required to perform a binary privacy risk assessment (i.e., on a $1-2$ binary scale) to assign a final score again based on the majority of assessments. In this case, 1 to denote a "Not at Risk" user while 2 an "At Risk" user.

3.2 Unsupervised Privacy Risk Assessment

The *unsupervised privacy risk assessment model* is designed to create *two privacy risk scores* that consider two essential factors: (*i*) sensitive information release and (*ii*) its dissemination scope. The first score, namely *Sensitive Information Release Risk Score* (SIRRS), is derived through the assessment of the release of sensitive information in the user-related content, while the second score, namely *Potential Scope Risk Score* (PSRS), involves the number of interactions (detailed in the following) for each user across all their generated content. The two scores are then *aggregated* to yield the final *Global Privacy Risk Score* (GPRS). This score plays a critical role in determining a potential risk class for each user, based on the selection of a given *privacy threshold*.

Sensitive Information Release Risk Score. This score aims to assess the tendency of user u and other users who have mentioned u to release sensitive information within the textual content. It is constituted by four distinct sub-scores: (*i*) *utw*, which considers the release of sensitive information in u's tweets, (*ii*) *otw*, which considers the release of sensitive information in tweets mentioning u, (*iii*) *ub*, which considers the release of sensitive information in the biography of u, and (*iv*) *ul*, which considers the release of geolocation information in the profile of u. Concerning (*i*) – (*iii*), the presence of sensitive information was detected using lists of *sensitive terms* associated – as proposed in [20], by taking inspiration from the definition of sensitive data in the EU GDPR – with ten *sensitive information categories* that include: (*i*) *health status*, (*ii*) *ethnicity*, (*iii*) *religion*, (*iv*) *political affiliation*, (*v*) *sexual orientation*, (*vi*) *geolocation*, (*vii*) *profession*, (*viii*) *marital status*, (*ix*) *interests/passions*, and (*x*) *age*. We note that the first five categories represent *special category personal data* according to Art. 9 of the EU GDPR, and are therefore deemed *highly sensitive* and in need of specific protection, unlike the remaining categories that we denoted as *less sensitive*. Lists of sensitive terms for each category have been taken from [6] (health status), [26] (ethnicity), [28] (religion), [16] (profession), [27] (political affiliation), [1] (sexual orientation), [13] (geolocation), [10] (marital status), [20] (interests/passions).

From the point of view of calculating the SIRRS, we first specify how its sub-constituents are computed. Concerning *utw*, for each u's tweet t, for each highly sensitive information i present at least once in it, a score α_{ti} is assigned. Similarly, the presence of each less sensitive information j, yields another score β_{tj}. The maximum overall score attainable, obtained by summing up all 10 scores for (both highly and less) sensitive information, equals 1 per tweet. The overall *utw* value for u is given by the average of these values over the total number N

of tweets. The same holds for *otw*, but in this case, the considered tweets are those mentioning u. Formally:

$$utw = otw = \frac{\sum_{t=1}^{N}\left(\sum_{i=1}^{5}\alpha_{ti} + \sum_{j=1}^{5}\beta_{tj}\right)}{N} \tag{1}$$

Concerning α_{ti} and β_{tj} values, it is possible to assign them in different ways. However, for simplicity, the ones already assigned and tested in [20] were used, i.e., $\alpha_{ti} = 0.15$ and $\beta_{tj} = 0.05$ in the presence of the release of sensitive information with respect to the categories considered. As *utw* and *otw* are defined, their values may vary in the range $[0 - 1]$.

As for *ub*, this value is calculated in the same way as the two previous scores, but limited to the biography of user u. From a formal point of view:

$$ub = \sum_{i=1}^{5}\alpha_{bi} + \sum_{j=1}^{5}\beta_{bj} \tag{2}$$

where α_{bi} and β_{bj} are the scores obtained for each sensitive information released in u's biography. Also in this case, the value of *ub* may vary in the range $[0 - 1]$.

Finally, as regards the calculation of *ul*, this score takes on a value of 1 if there is a matching between a geolocation value released by u in the user profile and a geolocation value from among those in [13]. Otherwise, it takes the 0 value.

The overall SIRRS value for user u is obtained as a *linear combination* of the previous values, which allows us to weigh some components more heavily at the expense of others (as we shall see in the experimental evaluations). Formally:

$$\text{SIRRS} = \omega_{utw} \cdot utw + \omega_{otw} \cdot otw + \omega_{ub} \cdot ub + \omega_{ul} \cdot ul \tag{3}$$

where, $\forall x \in \{utw, otw, ub, ul\} : \omega_x \geq 0$, $\sum \omega_x = 1$. Hence, the final value of SIRRS may vary in the range $[0 - 1]$.

Potential Scope Risk Score In this preliminary study, a pretty simple strategy was used to consider the potential privacy risk associated with the propagation of information in one's social network. One must first take into consideration that each user has different perceptions and purposes with respect to the dissemination of their information online. Some believe they have control over the level of its dissemination; others are not affected by this concern. Between these two extremes, there are many users who do not have a clear idea of the actual audience to which their content may be exposed. Our goal in this case was to identify tweets from selected users that achieve a high level of interaction compared to the average number of interactions of tweets from those same users. Interactions include the *number of likes*, *retweets*, and *comments* a tweet receives. In practice, the *Potential Scope Risk Score* (PSRS) allows the identification of tweets that have a high potential to be viewed by a large audience, exceeding the normal reach of the tweets of the users who posted them.[4]

[4] This can happen, for example, when a tweet is retweeted or mentioned by users with a large following, thus amplifying its reach.

Specifically, the PSRS score for u is derived by first calculating the *average interaction degree* of u. If the interactions for a given tweet t surpass this average, a value of $\gamma_t = 1$ is returned; otherwise, a value of $\gamma_t = 0$ is returned. The overall PSRS value is again given by the average of the values obtained from each tweet related to u. Formally:

$$\text{PSRS} = \frac{\sum_{t=1}^{N} \gamma_t}{N} \tag{4}$$

where N is the total number of tweets considered for the user u. To compute the average interaction degree, it was necessary to remove *outliers*. They were identified by considering *Interquantile Range* (IQR) [30], with $k = 1.5 * (Q3 - Q1)$, where $Q3$ and $Q1$ represent the third and first quartiles. Values greater than $Q3 + k$ or less than $Q1 - k$ are considered outliers.

Global Privacy Risk Score This overall score aims to identify the privacy risk of each u user by associating the riskiness of the published content, captured by the SIRRS, and its propagation, captured by the PSRS. The *Global Privacy Risk Score* (GPRS) is hence obtained by *linearly aggregating*, through different combinations of *importance weights*, the SIRRS and the PSRS. Formally:

$$\text{GPRS} = \omega_s \cdot \text{SIRRS} + \omega_p \cdot \text{PSRS} \tag{5}$$

where ω_s and ω_p represent the importance weights, and $\omega_s + \omega_p = 1$. In this work, different values for these weights were tested and illustrated in the experimental evaluations. As per definition, the GPRS assumes values in the $[0 - 1]$ range.

3.3 Supvervised Privacy Risk Assessment

This section involves the *supervised privacy risk assessment model*, which is contrasted with the previously discussed unsupervised model. The foundation of this supervised model lies in the utilization of labeled data, as elaborated in Section 3.1. This data was coupled with standard Machine Learning models and the extraction of pertinent features that pertain to the disclosure of sensitive information. The supervised models employed encompass *Logistic Regression*, *K-Nearest Neighbors*, and *Random Forests*.

Twenty distinct *metadata privacy-risk features* were considered, some derived from the unsupervised model as well as additional new features: (i) *number of characters* in the user's biography, (ii) presence of *geolocation* information, (iii) number of *followees*, (iv) number of *followers*, (v) average number of *likes*, (vi) average number of *comments*, (vii) average number of *retweets*, ($viii$) average *character count* of all tweets associated with u, (ix) the *utw* score, (x) the *ub* score, and $(xi) - (xx)$ the *average score* of the sensitive information released by the user for each of the ten sensitive information categories. This means calculating the average of the α_{ti} and β_{tj} values of each category with respect to the number of u's tweets. In addition, *textual privacy-risk features* extracted from user biography text, user tweets, and tweets mentioning target users were also considered. Specifically, they are *unigram features*, *bi-gram features*, and *tri-gram features* constituted by single terms, pairs of terms, and triples of terms with their associated *Term Frequency - Inverse Document Frequency* (TF-IDF) values, and, for each n-gram category, the top 500 terms related to their TF-IDF values.

4 Experimental Evaluation

This section discusses the experimental results obtained with respect to the unsupervised and supervised models presented in this work. Before detailing them, some technical details about the development of these models and the evaluation metrics used are presented.

4.1 Technical Details and Evaluation Metrics

Technical Details. The proposed models, both unsupervised and supervised, were implemented using the *Python* language. In particular, with regard to the classifiers used in the supervised model, the implementations provided within the `scikit-learn` library were used, with default parameters.[5] Also for the evaluation of the results, the implementations of the evaluation metrics (illustrated in detail below) provided in the `scikit-learn` library were used.[6] The `snscrap` library was used to crawl the tweets of the selected users and their additional data and metadata.[7] To address the problem of *class imbalance* in both multi-class and binary classification, the *Synthetic Minority Oversampling Technique* (SMOTE) [7] technique was used. SMOTE aims to increase, via K-Nearest Neighbours, the number of observations of a class that has fewer observations than the one with the most observations within a dataset. This way, the dataset for multi-class classification grew from 100 to 141 total observations, with 47 observations for each class, while the dataset for binary classification grew from 100 to 128 observations, with 64 observations for each class. Finally, for the supervised model, *k-fold cross-validation* [29] with $k = 5$ was used, by employing the again the `scikit-learn` library.[8]

Evaluation Metrics. To evaluate both the unsupervised and supervised models with respect to (multi-class and binary) *classification effectiveness* versus privacy risk, standard metrics, such as *accuracy* (Acc.) and F1-*score* (F1) [14], were used. For the unsupervised model, being able to obtain real privacy risk values associated with each user (i.e., the Global Privacy Risk Score), it was also possible to assess *ranking effectiveness* with respect to privacy risk, using the *normalized Discounted Cumulative Gain* (nDCG) [15].

[5] https://scikit-learn.org/stable/supervised_learning.html

[6] https://scikit-learn.org/stable/modules/model_evaluation.html

[7] https://github.com/JustAnotherArchivist/snscrape

[8] https://scikit-learn.org/stable/modules/cross_validation.html

4.2 Results: Unsupervised Privacy Risk Assessment

In the evaluation of the unsupervised model, the contribution of SIRRS and PSRS to the improvement of evaluation results was addressed. Specifically, we first considered several ways in which the construction of the SIRRS can contribute to increasing both classification and ranking effectiveness. As shown in Section 3.2, the SIRRS consists of a linear combination of distinct subcomponents (see Equation 3). Hence, we tested different weight combinations associated with them. Specifically, the three combinations are: (i) $\omega_{utw} = 0.4$ and $\omega_{otw} = \omega_{ub} = \omega_{ul} = 0.2$; (ii) $\omega_{utw} = 0.5$, $\omega_{otw} = \omega_{ub} = 0.2$, and $\omega_{ul} = 0.1$; and (iii) $\omega_{utw} = 0.6$, $\omega_{otw} = 0.2$, and $\omega_{ub} = \omega_{ul} = 0.1$. In addition, we evaluated effectiveness with respect to the contribution that both SIRRS and PSRS have by also combining them linearly with respect to different combinations of weights. The results of these evaluations with respect to both (multi-class and binary) classification and ranking (performed for both multi-class and binary labels) are shown in Table 2, in which the GPRS threshold was chosen by means of a *greed search* strategy maximizing the evaluation results.

Table 2. Results of the unsupervised model w.r.t. multi-class (MC) and binary (BC) classification (or labels for ranking) taking into consideration the components of SIRRS via combinations (i)–(iii) and the contribution of SIRRS (S) and PSRS (P) to GPRS.

GPRS	**SIRRS** (i)			**SIRRS** (ii)			**SIRRS** (iii)		
Computation	**Acc.**	**F1**	**nDCG**	**Acc.**	**F1**	**nDCG**	**Acc.**	**F1**	**nDCG**
MC: $S * 0.7 + P * 0.3$	0.48	0.48	0.96	0.50	0.50	0.96	0.53	0.53	0.95
MC: $S * 0.6 + P * 0.4$	0.49	0.49	0.96	0.54	0.54	0.96	0.54	0.55	0.96
MC: $S * 0.5 + P * 0.5$	0.52	0.52	0.95	0.57	0.58	0.96	0.54	0.55	0.96
MC: $S * 0.4 + P * 0.6$	0.53	0.54	0.95	0.52	0.53	0.95	0.52	0.53	0.95
MC: $S * 0.3 + P * 0.7$	0.51	0.52	0.94	0.49	0.50	0.94	0.49	0.50	0.94
BC: $S * 0.7 + P * 0.3$	0.78	0.78	0.95	0.78	0.78	0.95	0.78	0.78	0.95
BC: $S * 0.6 + P * 0.4$	0.78	0.78	0.95	0.78	0.78	0.95	0.76	0.76	0.95
BC: $S * 0.5 + P * 0.5$	0.78	0.78	0.95	0.78	0.78	0.95	0.76	0.76	0.95
BC: $S * 0.4 + P * 0.6$	0.74	0.75	0.95	0.76	0.76	0.95	0.72	0.72	0.94
BC: $S * 0.3 + P * 0.7$	0.68	0.69	0.94	0.70	0.71	0.94	0.68	0.69	0.94

4.3 Results: Supervised Privacy Risk Assessment

The supervised model was evaluated with respect to (multi-class and binary) classification effectiveness versus the privacy risk of users. Here, the results are illustrated with respect to the three classifiers considered, i.e., *Logistic Regression* (LR), *K-Nearest Neighbors* (K-NNs), and *Random Forests* (RFs), also taking into account different feature configurations, among those illustrated in Section 3.3. Specifically, they are referred to as:

i. TF-IDF: it includes, in addition to the 20 basic features, the TF-IDF values of the individual terms in the corpus. In this case, 22,721 unigrams and their TF-IDF values are considered;

ii. TF-IDF BEST-500: in this case, the 20 basic features and the 500 highest TF-IDF values of the unigrams extracted from the texts are taken into account;

iii. BI-GRAM: as the case (*i.*), but considering bi-grams instead of unigrams. In this case, we have 95,254 textual features;

iv. BI-GRAM BEST-500: as the case (*ii.*), but considering bi-grams;

v. TRI-GRAM: as the case (*i.*), but considering tri-grams. In this case, we have 106,384 textual features;

vi. TRI-GRAM BEST-500: as the case (*ii.*), but considering tri-grams;

vii. BI-TRI-GRAM BEST-500: the 500 bi-gram or tri-grams with the highest TF-IDF values.

The results of the multi-class and binary classifications with respect to the three classifiers and the different feature configurations are shown in Table 3, in terms of classification accuracy.

Table 3. Results of supervised multi-class and binary classification with each of the seven proposed feature configurations.

Classification	**Multi-class**			**Binary**		
Features *Classifier*	**LR**	**K-NNs**	**RFs**	**LR**	**K-NNs**	**RFs**
TF-IDF	0.42	0.53	0.64	0.59	0.75	0.82
TF-IDF BEST-500	0.41	0.53	0.64	0.59	0.75	0.83
BI-GRAM	0.43	0.54	0.72	0.59	0.76	0.78
BI-GRAM BEST-500	0.41	0.53	0.74	0.59	0.75	0.84
TRI-GRAM	0.45	0.54	0.66	0.59	0.76	0.78
TRI-GRAM BEST-500	0.41	0.53	0.68	0.59	0.75	0.85
BI-TRI-GRAM BEST-500	0.44	0.53	0.75	0.59	0.75	0.87

4.4 Results: Discussion

Unsupervised Model. From the results illustrated in Section 4.2, we can observe without great surprise that, from a macro point of view, the effectiveness of binary classification is far more satisfactory than that of multi-class classification. This can be due to the fact that it is difficult to correctly classify a "Partially at Risk" user. This semantic can be easily affected by the subjectivity of the evaluation of the human assessors. We can in fact observe that in the case of the nDCG measure, which is based on the evaluation of a ranking and not a classification, the values are more than satisfactory in both cases.

If we delve deeper into the factors that emerge as the basis of user privacy risk, there are interesting observations that apply to the classification tasks. First of all, we can observe how the best results are given, in binary classification, by the composition of the GPRS in which greater importance is given to SIRRS, or in any case until the two scores have the same importance. In multi-class classification, results depend more on the SIRRS composition; the best ones are obtained in relation to the *(ii)* SIRRS configuration, and when SIRRS and PSRS are equally important to GPRS. As regards the binary classification, the composition of the SIRRS does not seem to have any major impact on the final results, even if in this case they are slightly better with the *(ii)* configuration.

Supervised Model. Even in the case of the supervised model results illustrated in Section 4.3, we can observe that the multi-class classification performs less satisfactorily than the binary one, but nevertheless better than the multi-class classification from the unsupervised model. This observation is also not surprising. In this case the classifiers, in particular the one based on Random Forests, are able to produce a model that makes the most of the privacy-risk features identified in this work. Globally, RF results improve with the selection of the 500 best textual features, particularly for the model that uses the BI-TRI-GRAM BEST-500 features. This could suggest a significant impact on the privacy risk of the information released in the texts, some of which could be sensitive (as in the case of the unsupervised model); this, however, would need further in-depth analysis of the impact of individual features through explainable AI methods.

5 Conclusions and Further Research

In this study, we delved into the landscape of privacy risks that can affect social media users. In particular, we performed a preliminary investigation focused on analyzing the risk related to the release of sensitive information in user-generated content and its diffusion within the social network, by developing both unsupervised and supervised models. The unsupervised model, capable of generating privacy risk scores, took into account not only the direct release of sensitive information by users but also the cascading effects of content propagation. Simultaneously, the supervised model harnessed distinct privacy-risk features to pinpoint and incorporate potential vulnerabilities. Our evaluation encompassed multi-class and binary classification scenarios, using data extracted from the Twitter platform. The insights gleaned from our preliminary study, especially from the unsupervised model, suggest a positive interplay between individual sensitive information release and its far-reaching influence across social circles.

The proposed models and the obtained results would benefit from further refinement and analysis. In fact, the interplay mentioned earlier should be quantified in greater detail, both in the unsupervised and the supervised models. Concerning the unsupervised model, it will be necessary to carry out further analyses on the impact of the importance of the SIRRS components. Furthermore, concerning the supervised model, there would be a need to introduce an element of explainability concerning the effectiveness of individual features in relation to the classification process. Based on this further research, the results of the two models may also be related.

Acknowledgements

This work was supported in part by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, by project KURAMi (20225WTRFN) under the PRIN 2022 MUR program, and by the EC under grants MARSAL (101017171) and GLACIATION (101070141).

Data Availability

The labeled dataset generated and used in this work is available on request from the corresponding author.

References

1. M. Abrams, H. Weiss, S. Giusti, and J. Litner. 47 terms that describe sexual attraction, behavior, and orientation. https://www.healthline.com/health/different-types-of-sexuality, 2023. [Online; accessed 1-April-2023].
2. E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery. Scoring users' privacy disclosure across multiple online social networks. *IEEE Access*, 5:13118–13130, 2017.
3. C. Akcora, B. Carminati, and E. Ferrari. Privacy in social networks: How risky is your social graph? In *Proceedings of ICDE 2012*, pages 9–19, 2012.
4. A. Caliskan Islam, J. Walsh, and R. Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of WPES 2014*, pages 35–46, 2014.
5. B. Carminati, E. Ferrari, and M. Viviani. Online social networks and security issues. In *Security and Trust in Online Social Networks*, pages 1–18. Springer, 2014.
6. Centers for Disease Control and Prevention. List of all diseases. https://www.cdc.gov/health-topics.html, 2022. [Online; accessed 2-March-2022].
7. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *JAIR*, 16:321–357, 2002.
8. S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. Data privacy: Definitions and techniques. *IJUFKBS*, 20(6):793–817, 2012.
9. S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati. k-Anonymity: From theory to applications. *Transactions on Data Privacy*, 16(1):25–49, 2023.
10. E. S. Explained. Glossary: Marital status. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Marital_status, 2019. [Online; accessed 2-March-2022].
11. E. Ferrari and M. Viviani. Privacy in social collaboration. In *Handbook of Human Computation*, pages 857–878. Springer, 2013.
12. Z. Foreman, T. Bekman, T. Augustine, and H. Jafarian. PAVSS: Privacy Assessment Vulnerability Scoring System. In *Proceedings of CSCI 2019*, pages 160–165, 2019.
13. Geonames. All cities with a population > 1000. https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000, 2023. [Online; accessed 1-April-2023].

14. M. Hossin and M. N. Sulaiman. A review on evaluation metrics for data classification evaluations. *IJDKP*, 5(2):1, 2015.
15. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*, 20(4):422–446, 2002.
16. joblist.com. List of all jobs. https://https://www.joblist.com/b/all-jobs, 2022. [Online; accessed 2-March-2022].
17. K. Kircaburun, S. Alhabash, Ş. Tosuntaş, and M. D. Griffiths. 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. *IJMHA*, 18:525–547, 2020.
18. N. Kökciyan and P. Yolum. PriGuard: A semantic approach to detect privacy violations in online social networks. *IEEE TKDE*, 28(10):2724–2737, 2016.
19. K. Liu and E. Terzi. A framework for computing the privacy scores of users in online social networks. *ACM TKDD*, 5(1):1–30, 2010.
20. G. Livraga, A. Motta, and M. Viviani. Assessing user privacy on social media: The Twitter case study. In *Proceedings of OASIS 2022*, Barcelona, Spain, June 2022.
21. P. Matthews. Social media, community development and social capital. *CDJ*, 51(3):419–435, 2016.
22. A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 4:543, 2008.
23. S. Pei, L. Muchnik, S. Tang, Z. Zheng, and H. A. Makse. Exploring the complex pattern of information spreading in online blog communities. *PloS one*, 10(5):e0126894, 2015.
24. J. Shibchurn and X. Yan. Information disclosure on social networking sites: An intrinsic–extrinsic motivation perspective. *CHB*, 44:103–117, 2015.
25. J. Watson, H. R. Lipford, and A. Besmer. Mapping user preference to privacy default settings. *ACM TOCHI*, 22(6):1–20, 2015.
26. Wikipedia contributors. List of contemporary ethnic groups — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/List_of_contemporary_ethnic_groups, 2022. [Online; accessed 2-March-2022].
27. Wikipedia contributors. List of generic names of political parties — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/List_of_generic_names_of_political_parties, 2022. [Online; accessed 2-March-2022].
28. Wikipedia contributors. List of religions and spiritual traditions — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/List_of_religions_and_spiritual_traditions, 2022. [Online; accessed 2-March-2022].
29. T.-T. Wong and P.-Y. Yeh. Reliable accuracy estimates from k-fold cross validation. *IEEE TKDE*, 32(8):1586–1594, 2019.
30. J. Yang, S. Rahardja, and P. Fränti. Outlier detection: how to threshold outlier scores? In *Proceedings of AIIPCC 2019*, pages 1–6, 2019.