Challenges of protecting confidentiality in social media data and their ethical import

Arianna Rossi
SnT, University of Luxembourg
Luxembourg, Luxembourg
arianna.rossi@uni.lu

Mónica P. Arenas
SnT, University of Luxembourg
Luxembourg, Luxembourg
monica.arenas@uni.lu

Emre Kocyigit
SnT, University of Luxembourg
Luxembourg, Luxembourg
emre.kocyigit@uni.lu

Moad Hani*
University of Mons
Mons, Belgium
moad.hani@umons.ac.be

***Abstract*—This article discusses the challenges of pseudonymizing unstructured, noisy social media data for cybersecurity research purposes and presents an open-source package developed to pseudonymize personal and confidential information (i.e., personal names, companies, and locations) contained in such data. Its goal is to facilitate compliance with EU data protection obligations and the upholding of research ethics principles like the respect for the autonomy, privacy and dignity of research participants, the social responsibility of researchers, and scientific integrity. We discuss the limitations of the pseudonymizer package, their ethical import, and the additional security measures that should be adopted to protect the confidentiality of the data.**

***Index Terms*—Pseudonymization, security measures, GDPR compliance, research ethics, Named Entity Recognition**

1. Introduction

As big data analytics thrives, there is an urgent need for applications that help researchers and other data scientists to protect the personal and confidential data they collect and process while complying with the manifold obligations of applicable laws (e.g., the General Data Protection Regulation (GDPR)) and research ethics principles (e.g., respect for the autonomy, privacy, and dignity of research participants; scientific integrity; social responsibility), without investing excessive human and financial resources in daunting, time-consuming tasks that require manifold skills.

Very often, data gathered from online sources contain personal information, i.e., information that can identify the person to which it refers, and other kinds of information that should remain confidential. For instance, researchers from many disciplines recur to the scraping of social media [20] and analyze that data for various purposes, like sentiment analysis, disease tracking, and disinformation detection. Cybersecurity researchers collect data from online platforms to investigate cyberbullying, identity theft, identity impersonation, dissemination of pornography, fraud, and the like [29]. Such and other unstructured textual data (like e-mails, SMSs, etc.) are easily available "social gold mines" [8] that can be organized into usable, well-ordered data sets [20] with the goal of analyzing their informative content, training one's own machine learning models, and sharing knowledge with other researchers. Social media data sets are routinely employed in cybersecurity research, for instance, to build automated classifiers that evaluate the severity of software vulnerabilities [32] or to detect cyberattacks [18].

However, institutional review boards are often not adequately prepared on how to respond to ethical dilemmas in internet research (see e.g., the survey on US academic Institutional Review Board (IRB) [30]), including text and data mining. For example, we read in our IRB's authorization that social media data scraping is an "uncharted territory", even though this practice has existed for years. In order to overcome such challenges and contribute to the protection to the (possibly confidential) data collected at scale from social networks, we have developed an open-source Python package that pseudonymizes personal names, geopolitical locations, and companies mentioned in a text and can recover the original sentence at will, thereby maintaining data quality. The fact that this first version of the Pseudonymizer does not yet perform optimally provides the opportunity to discuss the technical and ethical import of such limitations and the additional security measures that should be adopted. It must be noted that pseudonymization alone is insufficient as the sole protection for publicly released data.

The main contributions of our work are: (i) analysis of the ethical and data protection issues that scientists need to consider when they scrape social media and other kinds of internet data; (ii) the development of an open-source Python package that allows the automated masking of companies and personal names and the generalizations of geolocations in unstructured text to help researchers protect the confidentiality of the data they collect and thereby contribute to the compliance with data protection provisions and the upholding of research ethics; it also

*This work was carried out as part of Moad Hani's student job at the SnT, University of Luxembourg.

DOI 10.1109/EuroS&PW55150.2022.00066

enables the recovery of the pseudonymized entities and the re-creation of the original text to protect scientific integrity, in particular data quality, and allow unwitting research participants to opt-out, thereby contributing to consent management; (iii) a presentation of the limitations of the pseudonymization of noisy textual data and their ethical import.

2. Legal and Ethical Challenges of Internet Data Research

On platforms that facilitate social interactions, both the data produced and published by users (e.g., comments, pictures, mentions of other users, URLs, etc.) and the relative metadata (e.g., location markers, timestamps, etc.), when linked together, can re-identify such users and should therefore be considered personal information. As a consequence, it is necessary to follow applicable laws on personal data protection to avoid the misuse of such information, including data breaches, and the associated financial and organizational risks (e.g., fines, reputation damage, loss of trust). Whenever an analysis of data containing personal information of individuals located in the EU is performed, the GDPR and its national implementations apply: no matter where situated, the members of the organization processing the data (e.g., academic researchers of a university, medical staff of a hospital, cybersecurity experts of a consultancy firm) must adopt organizational and technical measures (Art. 32.1 GDPR) to protect the confidentiality of such data.

Like thousands of cybersecurity experts, we upload information on the open-source threat intelligence and sharing platform MISP[1] with the goal of e.g., collaboratively analyzing trends of malware spread and phishing attacks. Several types of personal information can be disclosed on the platforms to investigate security incidents, including names, IP addresses, e-mail addresses, IBANs, Twitter IDs, and even the content of personal communications and images. Although the responsibilities for the legal and technical protection of personal data shared on MISP can take on various complex configurations, as a rule of thumb when an organization gathers and stores the data locally, processes them for its own purposes, and shares them with others like in our case, it becomes the data controller of such data [21]. Therefore, it inherits a set of obligations from the GDPR, that are meant to enhance the protection of such data and the accountability of those that process them.

Beyond legal obligations, when data are collected, processed and shared for scientific research purposes, additional safeguards should apply. Confidentiality and respect for the privacy and dignity of individuals constitute the central ethical tenet of research conducted on human beings [17], [22], which is intended to avoid any harm to research participants, like increased vulnerability, embarrassment, reputation damage, or prosecution [28]. Even when data are collected from the internet rather than directly from the individuals concerned, in the EU this is considered research with human subjects in its own right and is therefore governed by the same ethical principles [17]. However, most often the authors of online content are unaware that what they publish can be scraped and reused in other contexts for other purposes [9], [25]. No matter whether the information comes from available data sets or results from text and data mining research activities, participation in the research study cannot be considered informed, nor voluntary when the information is not collected directly from the individuals and these have not been provided with the opportunity to agree or disagree with the modalities and purposes of the use of their data [9]. Such unique conditions to internet research raise severe ethical issues. Moreover, although it may be believed that any information made available by people online is public and can thus be freely collected [17], [31], analyzed, and reported, ethical guidelines on internet research warn that it is not the case. Rather, contextual norms about confidentiality expectations (e.g., was the data shared on a password-protected group? How sensitive is the subject at hand?) and the possible consequences of data disclosure to a different audience outside of its original context should help researchers determine whether their actions are permissible [17]. This is why research with internet data is subject to several constraints: researchers should at least provide an easily accessible option to opt-out of the study and retract their data to those individuals whose information was harvested without their consent and should ask their permission before republishing such information [9], [31]. Thus, the possibility to manage consent and personal data should always be provided to research participants, especially when they are unaware of the data collection, but in the peculiar settings of social media data mining, standard solutions cannot be copy-pasted. Moreover, data integrity is key to the production of reliable findings, i.e., data should not be changed in an unreasonable manner [17], even out of privacy considerations. We refer the reader to our previous work [26], where we illustrate and discuss the often creative means we have developed to address the ethical and legal challenges of social media data scraping.

Cybersecurity researchers may also need to face possible legal actions from organizations that believe to have suffered reputation damage, like in the case of vulnerability disclosure. For example, our research gathers social media users' opinions about privacy designs they deem annoying, unethical, or manipulative (i.e., dark patterns) [12], and that could also be illegal. Researchers are subject to the principle of social responsibility [22] and should therefore be careful when companies are associated with negative judgments, as they may face defamation charges. Taking such a risk could be unnecessary if naming and shaming companies is not relevant for the analysis or when the data set is publicly released.

Compliance tasks are extremely burdensome and time-consuming, hence they deprive scientists and their organizations of useful financial and human resources that could be invested in more rewarding tasks [26]. Moreover, not all organizations, let alone academic ones, dispose of sufficient in-house resources to easily determine and implement appropriate ethics and data protection measures. Research integrity is not a "one-off" tick-box exercise, on the contrary: it means ongoing attention that is highly individual and context-dependent, [8] and inextricable from the method [10]. Whereas the Data Protection Officer, the IRB, and the legal team can only point to legal and

1. https://www.misp-project.org/

ethical issues and discuss possible solutions, it is the responsibility of researchers to implement them. When they do not have the expertise, the tools, or the motivation to do so, they may sidestep the rules and cause legal, financial, and reputation risks to the whole institution. For instance, none of the cybersecurity studies on social media data that we reviewed refers to any IRB's authorization nor research ethics or data confidentiality measure, confirming a tendency that is observable in other domains as well. However, as academic researchers and data controllers we are accountable for data management. Furthermore, a mishandling could have wider repercussions on the trust that the public places in science. When it comes to data privacy and security researchers, illegal or unethical personal data management becomes even less excusable.

3. Anonymization and Pseudonymization

Anonymization is an irreversible process that makes it impossible to re-identify the person to whom the data refer, at least in principle. Therefore, anonymized data are not considered personal data and are not subject to GDPR provisions. However evolving knowledge about the effectiveness of anonymization highlights that it is impossible to exclude with 100% certitude the risk of de-anonymization [14], especially in the case of graph data like social network data [16]. When a simple search online of a textual fragment or a picture can promptly re-identify its author, as in the case of indexed social posts [28], anonymization becomes simply impossible. Moreover, anonymizing data usually compromise a certain amount of data utility [23]: for instance, only when considered as a network certain events with attributes considered personal data (e.g., IP addresses) can be recognized as patterns (e.g., phishing attack) and analyzed as such, but not when they are individually analyzed.

Although it is often confused with anonymization, pseudonymization identifies a set of techniques that remove certain identifiers to contribute to data minimization and enhance data security (Art. 32 GDPR). Unlike anonymization, it can implement a recovery function that allows the association of the pseudonyms with the original identifiers [15] and thereby recover the original data. The automated masking of specific information (e.g., names, companies, locations, but also email addresses, Twitter IDs, etc.) with pseudonyms while retaining the keys to allow for recovery, can be useful in many cases, for instance to release a pseudonymized dataset with the keys only to trusted researchers and other stakeholders like cybersecurity experts, watchdogs, lawyers (see also the scenarios in [15]) like in our case. Similarly, researchers may want to identify the pseudonymized participants to ask for their authorization before republication of their data (e.g., tweet quotes on academic articles and data sharing platforms), allow them to retrospectively opt-out from the study, and re-establish the original data for analysis to maintain data integrity.

A risk-based approach should be adopted to determine which technique may be more appropriate for a certain context [15]. Identifying the intended protection level can derive from the type and sensitivity of data (e.g., biomedical data vs. publicly available social media data), the purpose of their analysis (e.g., derive aggregate statistics vs. compare specific people/companies/events), and their intended re-use (e.g., public disclosure of data set vs. private storage for internal analysis). Utility and scalability needs are also paramount [15], especially because there often is a trade-off between data utility (e.g., performing data classification over pseudonymized data) and data protection.

Using the keywords *pseudonymiz-e/ation* or *anonymize/ation*, we searched GitHub and PyPI for available open-source packages with the aim of pseudonymize a textual data set scraped from Twitter and Reddit on the topic of dark patterns. GoCept Pseudonymizer[2] is only composed of an encrypting function, while Pseudonymisation au Conseil d'État[3] [6] is explicitly developed for French judicial decisions. Data masking[4] and Python anonymizer[5] cannot mask names, locations and companies at once, nor they offer a recovery function. DeIdentify Twitter[6] also implements data masking, but only on structured Twitter data. Mainzelliste[7] [19] was designed to perform on medical joint research and biobank data stores. While Kodex[8] and FHIR pseudonymizer[9] have both pseudonymization and recovery functionalities, they were developed in `GO` and `C#`, and the latter only works on structured data expressed according to a biomedical standard. Therefore, we developed our own open-source package in Python. Since entity pseudonymization in noisy and informal social media sentences requires advanced Natural Language Processing (NLP) techniques like Named-Entity Recognition (NER), sophisticated and continuously updated open-source Python libraries are needed, such as Spacy and NLTK. Additionally, Python has the second-largest programming language community globally[10]. Therefore, an open-source tool developed with Python can be arguably improved by other developers.

4. The Pseudonymizer package

We developed a Pseudonymizer Python package[11] that works on English textual data and released it under a `GPL v2` license . This library works with structured and unstructured data, but in the case of unstructured data, and especially highly noisy data such as social media data, the challenge is greater and thus the performance is knowingly less accurate [5]. This software has three independent functionalities applied to different kinds of data: *Companies*, *Geolocations*, and *Personal Names*.

In the pseudonymization process, we act as the pseudonymization entity [15] that assigns the pseudonyms to each detected entity (via a mapping table) using a pseudonymization function. The Pseudonymizer package provides a list of *companies* and *personal names*, which are the inputs for generating the mapping tables of these

2. https://github.com/gocept/gocept.pseudonymize
3. https://github.com/etalab-ia/pseudo_conseil_etat
4. https://github.com/MWFK/Data-Masking/
5. https://github.com/nagyantal9312/python_anonymizer
6. https://github.com/qntfy/deidentify_twitter
7. https://github.com/cerbelding/medicalinformatics.mainzelliste
8. https://github.com/kiprotect/kodex
9. https://github.com/miracum/fhir-pseudonymizer
10. https://cutt.ly/yF84GLK
11. https://gitlab.uni.lu/irisc-open-data/2022-02-decepticon_pseudonymizer

two functionalities. Even if two or more organizations have the same list of companies and personal names, the mapping tables are unique as the package shuffles the list randomly using a permutation function. On the contrary, the mapping table for the *location* functionality is composed exclusively of the Geopolitical Entitys (GPEs) detected in the input data. The correspondence between pseudonyms and original pieces of information can be reestablished at will.

The pseudonymization function should not unreasonably alter the data to maintain data quality, thus the entity type of the input sentence must be detected. In order to fulfill this requirement, we implemented analytical and hybrid approaches to recognize and consequently pseudonymize the target entities. The analytical approach inspects the presence of the target entity in each sentence by iterating through the mapping table. The *company's* function is based entirely on this method. In contrast, the hybrid approach includes NER and analytical methods. The former inspects each phrase using NLP techniques and extracts the values of the target entity. Thus, if the entity is detected, the analytical method is initiated, which inspects through the mapping table.

4.1. Companies

This functionality uses data masking [23], a type of deterministic pseudonymization [15] that replaces the names of publicly listed companies with numbered pseudonyms, e.g., Ikea $(value) \rightarrow$ Company37 $(pseudonym)$. The list of companies was extracted from the Forbes Global 2000 database[12], which only contains a subset of the companies that can be found in a corpus, thus it can be enlarged with other sources e.g., other databases and/or user input. The mapping table is generated from the provided list of companies (n=4246). We also choose to pseudonymize some products as they can easily reveal the manufacturing company even when this is not explicitly mentioned. For instance, although *Chrome*, *YouTube*, *Maps* are part of the same company (*Alphabet*), each value was pseudonymized distinctively.

Pseudonymization function (P). The pseudonymization process is described in Alg. 1 and requires as input a dataset s, composed of a list of phrases or documents that may contain a company name. In this particular case, s is composed of 1562 phrases, scraped from Reddit.[13] Assuming that we use sentences without any pre-processing, the algorithm can output misleading results when dealing with strings that present spelling variations. Let us consider three phrases where the same entity is spelled differently (e.g., *IKEA*, *Ikea*, and *ikea*); only the phrase with the exact entity name would be pseudonymized. To avoid this kind of mismatching, the pre-processing of s is required to reduce the variations and improve the overall performance during the pseudonymization process. Thus, we lower-case the phrases, remove special characters, and ignore a bag of words that can be customized. After the pre-processing (s'), we inspect each sentence s'_i to search for any company listed in the mapping table. If so, the company is replaced by its pseudonym. Otherwise, the sentence remains unchanged. As output, we obtain a list of pseudonymized phrases ($\hat{s}$).

Algorithm 1: Pseudonymization function (P)

Input: List of phrases $s = (s_1, \ldots, s_n)$.
Output: List of pseudonymized phrases $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_n)$.
Data: MP $\{\text{values} = \{val_1, \ldots val_l\}, \text{nyms} = \{nym_1, \ldots, nym_l\}\}$

```
s' = (s'_1, ..., s'_n) ← Preprocess(s) Ignore bag-words, remove special characters
ŝ ← {}
for i ∈ {1, ..., n} do
    flag ← 0
    for j ∈ {1, ..., l} do
        if val_j ∈ s'_i then
            Replace each val_j in s'_i by nym_j
            flag ← 1
    if flag = 1 then
        Append s'_i to ŝ
return ŝ
```

4.2. Geolocations

This function uses a process of generalization [23] on geolocations, i.e., it recognizes specific cities or regions mentioned in English and replaces them with the corresponding country name thanks to a dedicated API. The Geonamescache library[14] was imported to create the mapping table which contains cities and their corresponding country.

Pseudonymization function (P). We recur to a hybrid approach that includes NER to determine the entity type, after applying normalization techniques such as removing non-alphanumeric characters or lower-case words. If the entity type is a GPE, its value is checked exclusively against the list of cities generated by the Geonamescache library. Additionally, it is replaced by the city's country if found on geonames.org. If the detected entity is a country, the entity value is preserved. Furthermore, a mapping table, including an ID, is created for the process of recovery. The implementation of NER aims to prevent the detection of false positives. For example, Paris is a city in France but it can also be the name of a person, thus the entity type should be detected before pseudonymization.

4.3. Personal Names

This functionality uses data masking with random pseudonyms. A database s,[15] composed of phrases that may contain first names and/or last names, was exploited to check whether the detected entities are personal names. Moreover, different open-source NLP python libraries, such as `spaCy` and `NLTK`, were used to tokenize the sentences and find the entity types.

12. https://data.world/aroissues/forbes-global-2000-2008-2019
13. https://www.reddit.com/r/antiassholedesign/
14. https://github.com/yaph/geonamescache
15. https://libraries.io/pypi/names-dataset

Pseudonymization function (P). We implement traditional normalization techniques in our datasets, such as removing non-alphanumeric values and lower-casing. Although removing stop-words is also one of the common ways of cleaning up the text data, we observed that removing them decreased the NER performance since it broke the sentence structure. Thus, we first implement the NER to each phrase, split all words, label the entity types (if matched), and then perform the data cleaning process. If the entity type of the word is a "PERSON", the text is checked against the available database for possible matching. If so, the text is replaced with a random fake name. Significant challenges for the pseudonymization of personal names are the vocabulary size and semantic ambiguity. For instance, "Antonio" or "Costa" are names, but if used as "San Antonio" or "Costa Rica", they designate a city and a country, respectively. Some names, especially last names, can also correspond to verbs or objects in the sentence like "Jobs". Therefore, we implemented a hybrid approach for pseudonymization due to the nature of the data and the targeted entity type.

Recovery function (R). The recovery process is described in Alg. 2 which requires a list of pseudonymized phrases $\hat{s}$ as input (this function is the same for the three functionalities). During this process, the algorithm searches for the presence of pseudonyms in each phrase $\hat{s}_i$. If one or more pseudonyms are matched, they are replaced by the names of the corresponding entity through the same mapping table used during the pseudonymization process. Otherwise, the phrase remains unaltered.

Algorithm 2: Recovery function (R)

Input: List of pseudonymized phrases $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_n)$
Output: List of phrases $s = \{s_1, \ldots, s_n\}$
Data: MP $\{\text{values} = \{val_1, \ldots val_l\}, \text{nyms} = \{nym_1, \ldots, nym_l\}\}$

```
s ← {}
for i ∈ {1, ..., n} do
    flag ← 0
    for j ∈ {1, ..., l} do
        if nym_j ∈ ŝ_i then
            Replace each nym_j in ŝ_i by val_j
            flag ← 1
    if flag = 1 then
        Append ŝ_i to s
return s
```

5. Evaluation

We used a confusion matrix to measure the overall performance of the three functionalities. We extracted the entities of each function according to the approaches described in section 4. The analytical approach checks each word in the text and, even when its usage is different from the target entity type, it pseudonymizes it when it is available in the mapping table. Although it is expected that NER models can be more successful for the target detection, we noticed that the used NER libraries did not perform any better due to the informal and non-standard language of social media content, thus we implemented a hybrid approach. According to a previously established codebook, the first three authors manually labeled the entities indicating *companies*, *locations*, and *personal names* in the different dataset s. Thus, we can extract insights from the results by estimating the True-Positives (TPs), False-Negatives (FNs), False-Positives (FPs), and the True-Negatives (TNs) and observe where our classification approach presents issues during the prediction. In our study, target entities are classified as positive values.

The **accuracy** metric, which is the ratio between the correctly predicted observations (TPs and TNs) and the total observations, was higher than 96% for all target entities, as observed in Table 1. We can attribute this performance to the fact that TNs examples constituted the majority of the entities in the data set. The **precision** values, which describe how good a model is at predicting the target entity over the total number of retrieved positives, show that *personal names* presented the lowest score while the *location* function was quite successful. Even though the same hybrid approach was implemented in both functions, the target entity affected the detection of FPs and therefore the precision values.

The **recall** describes how well the model predicts the positive class when the actual outcome is positive. We observe in Table 1 that the *location* functionality underperformed if compared to the *personal names* function. In other words, personal names sensitivity was better, due to the low number of FNs. The **miss rate**, the metric that measures the FNs performance of the model, performed better in the analytical approach than in the hybrid approach. The **false discovery rate**, which is the expected proportion of false positives and indicates when the model incorrectly predicts the positive class, was quite low for the *location*'s functionality. This result was directly related to the low amount of the predicted positive detection, and the entity-type checked by NER preventing high FPs. The pseudonymization of the personal names had the worst scores overall.

We observed that the *company's* function presented a high **accuracy**. However, its **precision** was low meaning that some non-companies were identified as such and pseudonymized, thereby contributing to the number of FPs; e.g., *"make sure she never sees the **light** of day"*, the function P pseudonymized the *light* value. Likewise, the **recall** was relatively low, meaning that many companies were not identified and therefore were not pseudonymized, e.g., the company *marks and spencer* was not detected in the sentence *"I really respect this move from **marks and spencer**"*, increasing the number of FNs. This metric can be improved by increasing the company's list. A crucial result is that the words having overlapping entity types affect the performance of the personal name and location pseudonymization functionalities in terms of precision and recall like in the following phrases e.g., *"Paris in Wisconsin bought jumpsuit 13 minutes ago"*. *Paris* was detected as a location and this increased FPs. Another factor that raised FPs and decreased **precision** in personal name functionality, is when there are mistakes like the NER classifying *Java* as personal name in the sentence *"Java recommends you delete it ..."*.

TABLE 1. RESULTS OF THE DIFFERENT METRICS.

	Accuracy	Precision	Recall	Miss Rate	False Discovery
Companies	0.964	0.602	0.609	0.391	0.398
Locations	0.994	0.965	0.451	0.549	0.035
Personal Names	0.984	0.348	0.859	0.141	0.652

We additionally measured the **execution time** of the different approaches. The *geolocations* and *personal names* functionalities performed quite similarly and on average each sentence was pseudonymized in 0.01 seconds. However, the *company's* function took on average 0.12 seconds per phrase for the pseudonymization. We also observed that the execution time for the recovery process is quite efficient because the targeted phrase does not require any NER inspection, but it only requires iterations through the mapping table (0.0001 seconds per phrase on average).

6. Discussion

Performance. This study shows that the pseudonymization process and its performance are closely correlated with the target entity type and the data set attributes. Proposing a generalized solution for pseudonymization is not realistic. Training solutions on specific types of information (e.g., see the training on the corpus of judicial decisions in [6]) will supposedly yield to more accurate results. It will be crucial to train our model on more extensive and balanced domain data sets (e.g., with more TPs).

We observed high miss rate scores for each functionality because the noisy structure of social media data affected the NER's performance. Future work intends to analyze if data sets based on formal, standard language (e.g., news articles) can provide better results and investigate how to implement other approaches to optimize the performance on social media data. In terms of false discovery rate, the wide personal names vocabulary size contains many common words, e.g., "Rose" can refer to an object or a person. Therefore, identifying the data set domain and using topic modeling can be crucial to contextualize the entity type for the detection algorithm. Similarly, we should implement mechanisms that consider contextual cues to disambiguate between entities like "apple" as an object or a company.

The training process should contain a feedback mechanism to improve the rules or checklists, as the FPs words can be detected and added to the ignore words list. For instance, we decided to ignore "Don" as it was classified as the first name for many instances e.g., "*Don buy this!*" or "*Don consider the price*". Additionally, the precision of the company's functionality is associated with the company database size and completeness, which we fed with a NER-based feedback mechanism. Thus, researchers who analyze social media data need to implement customized or domain-specific techniques and data processing phases to achieve better performance, even though this would subtract time from other research tasks which casts doubt on the actual feasibility of such approaches.

Concerning the execution time difference between the three functionalities, it depends on the fact that, for the company function, we used regular expressions whose time complexity is $O(n)$. Still, it may require up to $O(2^m)$ construction time/space, where m is the regular expression size. This execution time can be improved by implementing other techniques or parallelizing the process. Finally, recovery is a vital functionality of pseudonymization to provide availability of the original data and allow the research participants to opt-out of the study or to be recontacted in case of republication of their data. Even though the mapping table is securely stored, the recovery flow enables access to the personal data, which needs to be protected with additional measures (see Sec. 7).

Ethical implications. The results show that many questions still need to be addressed before the package can be promoted and help other researchers to automatically and reliably pseudonymize their data sets. Even assuming that the results will enhance thanks to the application of additional approaches (see Sec. 8) and reach state-of-the-art on noisy social media data NER [5], it is crucial that we address an ethical question: what would be our responsibility in the case of inaccuracies when others use the package? False negatives entail that not all entities that should have been pseudonymized have been detected, thereby exposing data that should remain confidential. On the other hand, false positives imply that certain pieces of information have been mistakenly altered, thus impacting data quality, i.e., their utility, transparency, and reproducibility [22].

That said, even if a human check should always be performed, it is undeniable that an automated approach offers benefits for researchers, who would otherwise need to manually pseudonymize the data, buy a dedicated software (provided that such software exists for social media text), or invent and develop their own automated process. We intend to disclose training data sets together with the package whenever possible, as well as be transparent about the performance of the package on such data, and update the results when we implement new approaches and functionalities. However, this would mean that those using the package should have at least some level of data literacy to interpret the caveats that we describe. Many academic research teams do not dispose of such skills, though.

The question about developers' liability raises another aspect that we have encountered when we have examined the license of the components we used to determine the most appropriate license for the Pseudonymizer. We noticed that one of the packages we heavily rely on to form the list of names in the analytical approach is described as deriving from the massive Facebook data leak, where the personal data of 533 million users in 2021 [24] (i.e., phone numbers, Facebook IDs, full names, locations, birth dates, bios, email addresses) were illegally scraped and published online by exploiting a feature offered by the platform [4]. Thus we wonder if it is ethically acceptable to use such a data set and whether the research benefits outweigh the risks and harms. We will create a synthetic data set (e.g., by randomizing the name order) to protect the identity of the people present in the leaked data set.

Similarly, the company's data set is available online under an unspecified "public domain" license. We asked its author the reasons for such license and the origins of the data set but have not received any answer at

the moment of writing. As mentioned in section 2, just because data is available online does not mean that it can be scraped or reused by others at will. It can not be safely assumed that developers have sufficient knowledge of applicable laws (i.e., data protection and intellectual property) and research integrity when they create data sets and develop packages that are published and adopted by millions of other developers worldwide. Moreover, developers are not always aware of how to correctly use licenses for the piece of software they develop, which has consequences on the reuse opportunities and the licenses that can can applied. This is why there are approaches [13] intended to enhance accountability in data set creation, publication, and maintenance. However, we should further reflect on our responsibility to use such data sets would be when alternatives do not exist.

7. Additional data protection measures

The different files obtained through the pseudonymization process must be stored securely, kept secret, and known only to the data controller. Otherwise, internal and external attacks could compromise pseudonymization by retrieving the entirety of the information and reversing the function. Given that some pseudonymization techniques can more securely protect such data, we recommend combining our approach with other measures, especially when the data at hand are sensitive (e.g., biomedical data). For instance, the mapping table can be kept encrypted to avoid any potential leakage to external adversaries. For each query, the search can be done over the encrypted data. There are different ways to achieve this: deterministic public-key encryption [1], searchable symmetric encryption [3], or public-key encryption with keyword search [2]. The mapping table can also be exposed to other threats, such as its deletion by mistake or by a malicious adversary; thus, it can be outsourced to a database server to prevent data loss and ensure continuous backup. The database server stores the data in the encrypted domain and does not have access to the cryptographic keys. Format preserving encryption and hashing functions can be imported from other open-source available packages mentioned in Sec. 3.

Even the communication channels that serve to share the data sets, for example on Git repositories, should be encrypted and be furthermore subject to a strong authentication procedure via SSH public key authentication or password authentication [26]. Each of them presents advantages and disadvantages: e.g., more or less protection, risk of losing encryption keys, and the difficulty of password management across different devices. A role-based access control [27], that is often implemented on Git repositories, distinguishes between the access rights of various individuals (e.g., students, senior researchers, principal investigators, etc.) and ensures further protection from unauthorized access and data misuse. Git repositories that are locally managed avoid resorting to external cloud services that could expose to the risk of data access from other jurisdictions without an adequate level of legal protection. All other data protection measures (e.g., purpose and storage limitation, transparency, etc.) and ethics research principles (e.g., exclusion of minors' data, data quality, etc.) must also be applied to the scraping of social media data, described in our previous work [26].

If one wishes to publicly release a data set containing personal information, it should be noted that with pseudonymization alone it cannot be claimed that such information is sufficiently protected. Re-identification can easily happen not only via the mapping table, but also thanks to contextual information and indexed content online, as well as other methods. Other technical and organizational measures should therefore be adopted, as well as a thorough ad hoc reflection about research ethics issues.

8. Limitations and Future work

This first work addresses only a few of the challenges that we have identified. We can expect that the results will improve if the model is trained on data generated by specific communities (e.g., gamers) or focused on specific topics (e.g., video games). Moreover, it has been recently argued that NER tasks need a more nuanced evaluation approach than classical F1 measures [11], thus, future work will also be devoted to this. The pseudonymization process should be customized with entity-type specific rules to minimize miss rate and false discovery rate, customized rule-based models can be included in the package. Moreover, it will be crucial to evaluate the package according to additional security measures, like confidentiality and integrity. That said, even assuming perfect performances, pseudonymized entities can be easily deduced by other contextual elements, like the mention of a famous soccer player who acts in the advertisement of a renamed company: just by recognizing the person, the company can be identified as well.

Additional functionalities could also be implemented. First, it would be useful to extend the pseudonymization to languages other than English, including resource-poor languages that are notably more in need of tools for automated language processing. Different pseudonymization approaches should also be devised. For example, hashing functions could strengthen the data protection measures together with additional processes that ensure the unlinkability of individual values like tweets [15], while differential privacy applied to authorship [7] can remove unique stylistic cues and thus enhance the de-identification of authors.

Furthermore, we aim to build a user-friendly application where users do not require any programming experience to apply our pseudonymization process. Such an application could also incorporate a feedback mechanism for the analytical approach to enlarge the database of companies and names. This could also be achieved by using existing APIs[16] and names other than the English ones. However, the accuracy of such approaches may suffer.

9. Conclusions

In this article, we have addressed the pseudonymization process of social media data, which consist of noisy

16. Like OpenCorporates, available at: https://api.opencorporates.com/documentation/API-Reference

textual data, and presented its usefulness to comply with EU data protection provisions and research ethics principles, as well as its limitations that still need to be overcome. While discussing the ethical and legal aspects of this process, we have also presented the first version of an open-source Pseudonymizer package that can help scientists to mask companies and personal names and generalize locations at scale. The package can recover the pseudonymized entities and thus reconstitute the original text, thereby ensuring data integrity and allowing the authors of such content to opt-out and be recontacted when authorization for republication is necessary. Although the results show that improvement is needed before we can safely promote the package among other researchers, we have established a clear pathway to do so. We also call upon fellow scientists to contribute to such a goal.

Acknowledgment

This work has been partially supported by the Luxembourg National Research Fund (FNR): "Deceptive Patterns Online (Decepticon)" IS/14717072 and No more Fakes "NOFAKES" PoC20 / 15299666 / NOFAKES-PoC.

References

[1] M. Bellare, A. Boldyreva, and A. O'Neill. Deterministic and Efficiently Searchable Encryption. In A. Menezes, editor, *Advances in Cryptology - CRYPTO 2007*, pages 535–552, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[2] D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano. Public Key Encryption with Keyword Search. In C. Cachin and J. L. Camenisch, editors, *Advances in Cryptology - EUROCRYPT 2004*, pages 506–522, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[3] Z. Brakerski and G. Segev. Better Security for Deterministic Public-Key Encryption: The Auxiliary-Input Setting. *Journal of Cryptology*, 27(2):210–247, apr 2014.

[4] M. Clark. The facts on news reports about facebook data. https://about.fb.com/news/2021/04/facts-on-news-reports-about-facebook-data/, Apr 2021.

[5] L. Derczynski, E. Nichols, M. van Erp, and N. Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, page 140–147. Association for Computational Linguistics, 2017.

[6] EtaLab IA. Guide à la pseudonymization decisions ce. https://github.com/etalab-ia/pseudonymisation_decisions_ce, Jan 2020.

[7] N. Fernandes, M. Dras, and A. McIver. *Generalised Differential Privacy for Text Document Processing*, volume 11426 of *Lecture Notes in Computer Science*, page 123–148. Springer International Publishing, 2019.

[8] C. Fiesler, N. Beard, and B. C. Keegan. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:187–196, May 2020.

[9] C. Fiesler and N. Proferes. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366, Jan 2018.

[10] A. S. Franzke, A. Bechmann, M. Zimmer, and C. M. Ess. Internet research: Ethical guidelines 3.0: Association of internet researchers, 2019.

[11] J. Fu, P. Liu, and G. Neubig. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, 2020.

[12] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, page 1–14, Montreal QC, Canada, 2018. ACM Press.

[13] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 560–575. ACM, Mar 2021.

[14] Ireland Data Protection Commission. Guidance on anonymisation and pseudonymisation, Jun 2019.

[15] M. Jensen, C. Lauradoux, and K. Limniotis. *Pseudonymisation techniques and best practices. Recommendations on shaping technology according to data protection and privacy provisions*. European Union Agency for Cybersecurity (ENISA), November 2019. DOI 10.2824/247711.

[16] S. Ji, P. Mittal, and R. Beyah. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials*, 19(2):1305–1326, 2016.

[17] L. K. Kaye, C. Hewson, T. Buchanan, N. Coulsoun, Branley-Bell, C. Fullwodd, and L. Devlin. *Ethics Guidelines for Internet-mediated Research*. The British Psychological Society, 2021.

[18] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan. Crowdsourcing cybersecurity: Cyber attack detection using social media. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, page 1049–1057, Singapore Singapore, Nov 2017. ACM.

[19] M. Lablans, A. Borg, and F. Ückert. A restful interface to pseudonymization services in modern web applications. *BMC Medical Informatics and Decision Making*, 15(1):2, Feb 2015.

[20] N. Marres and E. Weltevrede. Scraping the social? issues in live social research. *Journal of Cultural Economy*, 6(3):313–335, Aug 2013.

[21] MISP. Information sharing and cooperation enabled by gdpr. https://www.misp-project.org/compliance/GDPR/, Jan 2018.

[22] J. Oates, D. Carpenter, M. Fisher, S. Goodson, B. Hannah, R. Kwiatkowski, K. Prutton, D. Reeves, and T. Wainwright. *BPS Code of Human Research Ethics*. The British Psychological Society, Apr 2021. ISBN 978-1-85433-792-4.

[23] A. . W. Party. Opinion 05/2014 on anonymisation techniques, 2014.

[24] J. Peters. Personal data of 533 million facebook users leaks online. https://www.theverge.com/2021/4/4/22366822/facebook-personal-data-533-million-leaks-online-email-phone-numbers, Apr 2021.

[25] N. Proferes. Information flow solipsism in an exploratory study of beliefs about twitter. *Social Media + Society*, 3(1):2056305117698493, Jan 2017.

[26] A. Rossi, A. Kumari, and G. Lenzini. *Unwinding a Legal and Ethical Ariadne's Thread out of the Twitter's Scraping Maze*. Springer Nature, Venice, sebastien ziegler, adrian quesada rodriguez and stefan schiffner edition, In press.

[27] W. Stallings. *Operating system security (Chapter 24)*, pages 24.1–24.21. Wiley, 6 edition, 2014.

[28] L. Townsend and C. Wallace. *Chapter 8: The Ethics of Using Social Media Data in Research: A New Framework*, volume 2, page 189–207. Emerald Publishing Limited, Dec 2017.

[29] E. van der Walt, J. H. P. Eloff, and J. Grobler. Cyber-security: Identity deception detection on social media platforms. *Computers & Security*, 78:76–89, Sep 2018.

[30] J. Vitak, N. Proferes, K. Shilton, and Z. Ashktorab. Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics*, 12(5):372–382, Dec 2017.

[31] M. L. Williams, P. Burnap, L. Sloan, C. Jessop, and H. Lepps. *Users' Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm*, volume 2, page 27–52. Emerald Publishing Limited, Dec 2017.

[32] S. Zong, A. Ritter, G. Mueller, and E. Wright. Analyzing the perceived severity of cybersecurity threats reported on social media. In *Proceedings of NAACL-HLT*, page 1380–1390, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.