Cloaked Classifiers: Pseudonymization Strategies on Sensitive Classification Tasks

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, Djamé Seddah

Cloaked Classifiers: Pseudonymization Strategies on Sensitive Classification Tasks

Arij Riabi **Menel Mahamdi** **Virginie Mouilleron** **Djamé Seddah**
Inria, Paris
{firstname,lastname}@inria.fr

Abstract

Protecting privacy is essential when sharing data, particularly in the case of an online radicalization dataset that may contain personal information. In this paper, we explore the balance between preserving data usefulness and ensuring robust privacy safeguards, since regulations like the European GDPR shape how personal information must be handled. We share our method for manually pseudonymizing a multilingual radicalization dataset, ensuring performance comparable to the original data. Furthermore, we highlight the importance of establishing comprehensive guidelines for processing sensitive NLP data by sharing our complete pseudonymization process, our guidelines, the challenges we encountered as well as the resulting dataset.

1 Introduction

Radicalization, fostered by online propaganda and offline indoctrination, has been the primary driver in most terror attacks and eruptions of public violence over the past decade (Farwell, 2014; Fernandez and Alani, 2021; Pellicani et al., 2023). It can be defined as a process by which an individual or group adopts increasingly radical viewpoints in opposition to a political, social, or religious system (Fink, 2014). These viewpoints cover, for example, far-right ideologies, religiously inspired extremism, and extreme conspirationism. Such content can spread rapidly, especially through social media, making radicalization challenging to detect (Nouh et al., 2019).

Natural Language Processing (NLP) methods have been used to detect and analyze radicalization mechanisms such as propaganda, recruitment, networking, data manipulation, and disinformation (Torregrosa et al., 2021; Aldera et al., 2021; Gaikwad et al., 2021). However, the effectiveness of such detection models depends on the availability and quality of training and evaluation datasets. Protecting user privacy, especially for sensitive tasks, is imperative when sharing such datasets. Finding the right balance between the obligation to build accurate anonymization methods and the need to maintain a decent level of performance is hard, as pertinent information may be contained through some identifiers (usernames, URLs, locations, etc.) and their associated socio-demographic or geographic markers. Hence, a *brutal* anonymization of a dataset can hinder its usability, especially in a domain where radicalization clues are often found through these indicators (Pellicani et al., 2023).

Ensuring the privacy of individuals is critical, especially in light of regulations such as the General Data Protection Regulation (GDPR)[1]. This is why we believe that despite implementing various laws to minimize harm and protect sensitive information, there is a need to explore how technological advancements intersect with data protection laws and impact the collection, storage, and use of confidential data (Nguyen and Vu, 2023; Lothritz et al., 2023).

In this work, we present our methodology for the manual pseudonymization of a radicalization dataset that (i) ensures performance to be comparable to the original data while maintaining its semantic properties and (ii) protects user privacy. We emphasize the importance of establishing a standard framework for privacy and usefulness when processing sensitive NLP data by sharing the complete pseudonymization process for our datasets and the challenges we faced (Vakili and Dalianis, 2022, 2023). It is a highly sensitive task that requires 100% accuracy; any oversight can render the dataset invalid.

Our dataset includes English, French, and Arabic content from various sources such as forums, Telegram and other social media platforms. The con-

[1]The GDPR is a comprehensive data protection law enacted by the European Union (EU). It aims to protect the privacy and personal data of individuals within the EU and the European Economic Area (EEA).

tent covers different radicalization domains (from white supremacy to jihadism) for each language. Our dataset will be available upon publication[2].

The manual annotation process we devised guarantees a high level of precision and enables us to better explore the interaction of our NLP tools and improve user safety. Furthermore, a critical component of our methodology involves identifying the exceptions for which anonymization does not need to be applied. For example, keeping well-known events and public figures enables us to leverage the knowledge embedded in the language model about specific entities and prevent pseudonymization from corrupting the relationships and alignment between named entities and other elements within the text, thereby enhancing the effectiveness of our system. Our evaluation results show that models trained on our pseudonymized data maintain similar levels of performance to their original counterparts.

To summarize, our contributions are as follows:

- We developed and share detailed guidelines[3] for our pseudonymization method.
- We release a pseudonymized multilingual radicalization detection dataset [4].
- We provide an analysis of performance, demonstrating that our method maintains the same level of effectiveness as the original data while protecting user privacy.

2 Related Work

2.1 Definitions

The GDPR provides a comprehensive definition of personal data, including any information related to an identified or identifiable natural person. According to Article 4 (1) of the GDPR, *"personal data means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person"*. Building on this definition, **anonymization** refers to the complete and irreversible removal of any data in a dataset that could potentially identify an individual, directly or indirectly. **De-identification** involves the removal of specific, predetermined direct identifiers from a dataset. **Pseudonymization** is replacing direct identifiers with pseudonyms or coded values while keeping the mapping between the pseudonyms and original identifiers stored separately. The definitions of these terms may vary across literature, and they are often used interchangeably (Lison et al., 2021; Lothritz et al., 2023).

Traditional manual methods for anonymizing text data may be inefficient, error-prone, and expensive, making it necessary to develop well-defined frameworks. Lison et al. (2021) point out a significant gap between NLP and privacy-preserving data publishing (PPDP) approaches, both of which have addressed aspects of anonymization independently without sufficient interaction (Papadopoulou et al., 2022). Given the complexity of text data, including indirect identifiers and nuanced semantic cues, there is a need for improved anonymization models that can effectively balance the trade-off between privacy protection and data utility.

The NLP-based approach usually turns text anonymization into a NER-like problem (Eder et al., 2022), where a set of categories set in advance are to be retrieved from the text. The PPDP approach uses "privacy models" (Sánchez and Batet, 2016, 2017; Brown et al., 2022), which are sets of requirements that are to be met by the anonymization system, often regarding identification by aggregation of data, degrees of anonymization and potential attacks.

Yermilov et al. (2023) compare three machine-learning-based pseudonymization techniques that consist of a *NER-based* classical approach, *seq2seq* (Lewis et al., 2020), which frames the task as a sequence-to-sequence transformation using an encoder-decoder model, and *LLM Pseudonymization*, which uses a two-step process with GPT-3 and ChatGPT: GPT-3 extracts named entities, and ChatGPT then pseudonymizes them.

Text pseudonymization usually requires three steps: (1) establishing relevant categories of personal data, (2) retrieving them, and (3) replacing them. We will briefly introduce the related works in the next subsections.

[2]Note that evaluating the radicalization detection task in itself is not the main point of the paper; here, we focus on our pseudonymization process.

[3]https://file.io/rmUwdPfvnmXq

[4]https://gitlab.inria.fr/ariabi/counter-dataset-public

2.2 Establishing categories

To our knowledge, there is no standardized set of categories, especially for non-medical, unstructured, online textual data that is processed in the European Union.

Since pseudonymization has mainly been used in the medical domain, most papers use the Personal Health Identifiers (PHI) enumerated in the American HIPAA regulations (HIPAA, 2004), either as a reference (Yang and Garibaldi, 2015; Dernoncourt et al., 2017) or as a starting point for further adaptation to the corpus (Velupillai et al., 2009; Dalianis and Velupillai, 2010; Megyesi et al., 2018; Eder et al., 2020). Some draw categories from data observation (Medlock, 2006; Adams et al., 2019; Çetinoğlu and Schweitzer, 2022). Adams et al. (2019) set 3 types of entities for their online chat corpus: Personal Identifying Information (PII), Corporate Identifying Information (CII), and Others, with only PII and CII being anonymized. Others create categories using the GDPR-based distinction between direct identifiers, indirect/quasi-identifiers, and sensitive data (Pilán et al., 2022; Volodina et al., 2020).

Still, making up an all-encompassing set of categories is not an easy task, and when it comes to non-clinical data, the line between what is to be anonymized and what is not becomes blurred for some entities. Çetinoğlu and Schweitzer (2022) resorted to heuristics and highlighted the subjective dimension of data pseudonymization. The datasets often display some special categories that have to be mentioned and taken into account in the annotation scheme:

- Indirect or quasi-identifiers: they are almost always anonymized (Adams et al., 2019; Volodina et al., 2020; Lison et al., 2021) following the GDPR. An argument cited by many is the study conducted by Sweeney (2000), which showed that 87% of the US population could be identified only by zip code, date of birth, and gender. Moreover, Identification by data aggregation and its prevention is a common theme in the literature.
- Sensitive information, such as ethnicity, political views or sexuality, are either anonymized or at least detected and annotated for further processing (Volodina et al., 2020).
- Public figures: briefly mentioned in Adams et al. (2019) and Çetinoğlu and Schweitzer (2022), they are not anonymized.
- Deceased people: there has been no mention of the case of deceased people. Although GDPR doesn't apply in this case, the French CNIL[5] has advised to apply data protection rules when it might impact families and close ones.

Finally, some have argued that one must not entirely rely on a closed, predefined set of categories: Pilán et al. (2022) suggest that all textual elements must be considered, as they can still be used for re-identification, either directly or indirectly through inference.

2.3 Data retrieval

Data retrieval can be done manually or with rule-based models (Neamatullah et al., 2008; Çetinoğlu and Schweitzer, 2022), but most of the related works employ machine learning and, more recently, focus primarily on deep learning approaches (Dernoncourt et al., 2017; Liu et al., 2017; Papadopoulou et al., 2022). Finally, anonymization pipelines and toolkits have also been proposed to coordinate human annotation and different anonymization techniques (Adams et al., 2019; Clos et al., 2022).

2.4 Substitution strategies

Textual data substitution usually falls into three categories. One can choose categorization (a term first used by Medlock (2006)), by which one exact string replaces all units from the same category. For example, the SOLID Twitter dataset (Rosenthal et al., 2021) replaces all usernames with the placeholder "@USER," and in Volodina et al. (2020), all bank accounts are replaced by the same standardized string "0000-00 000 00". Another method we call non-realistic pseudonymization consists of replacing each unit with a specific identifier that does not mimic natural language. Such is the case in the Dortmund Chat Corpus 2.1 (Lüngen et al., 2017), in which a person's name is replaced by an id, such as "`[_PERSONNAME-1_]`". A third method, which we call realistic pseudonymization, attempts to avoid loss of linguistic information by replacing the unit with a semantically similar identifier and that mimics natural language (Çetinoğlu and Schweitzer, 2022; Eder et al., 2022; Olstad et al., 2023). To preserve data quality, we chose this approach for our dataset.

[5] French data protection authority.

Some research purpose to extend pseudonymization efforts beyond the clinical domain (Lampoltshammer et al., 2019; Pilán et al., 2022; Yermilov et al., 2023). Nevertheless, these efforts are currently confined to a limited list of categories, such as names (Lothritz et al., 2023) or just names and addresses (Accorsi et al., 2012), in an artificial setting. We disclose the exhaustive list of entity categories and all the considerations taken into account during the anonymization for our task. Our position aligns with the recent research of Szawerna et al. (2024), who propose implementing a universal tagging system for categorizing personally identifiable information (PII) to improve pseudonymization processes. They emphasize that existing tagsets do not encompass all PII types found across various domains with the necessary level of detail for successful pseudonymization.

The pseudonymization of our dataset is important for sharing it for research purposes, as it minimizes information loss, which is a well-known undesirable side effect (Meystre et al., 2014; Sawhney et al., 2022; Lothritz et al., 2023). Additionally, Lampoltshammer et al. (2019) showed that even small changes in data anonymization can significantly impact sentiment analysis results even though Vakili et al. (2022) showed no significant change in performance after anonymization for clinical data. The results of our experiments that show almost no impact (Subsection 4.4) confirm their findings.

3 Methodology

We argue that the sensitive nature of certain tasks requires human annotators; therefore, a considerable amount of our pseudonymization process is done manually. Our guidelines are based on three primary sources: legal texts and recommendations from the French CNIL and the GDPR, existing research on data anonymization for NLP, and a thorough analysis of our corpus. As far as we know, no work has been published on the pseudonymization of radicalization data. We have also not found any official, standardized method for pseudonymizing textual data, neither from the GDPR/CNIL nor the literature.

3.1 Data types

We define three main types of data in our dataset: data related to individuals, data related to organizations, and data related to content sharing.

Data related to Individuals. We have systematically anonymized all direct identifiers (e.g. names, addresses, email addresses, phone numbers) associated with private individuals. For indirect identifiers (e.g., nationality, general location, age, gender), we decided to anonymize at least one in cases where multiple identifiers appear in the same text.

Following Adams et al. (2019); Çetinoğlu and Schweitzer (2022), public figures are not anonymized. We also include journalists, politicians, and authors in that category. Additionally, we introduced a category for **"Influencers,"** determined by criteria such as social media presence, follower count, and appearances in mainstream media. Although these profiles are not anonymized, specific sensitive direct and indirect identifiers (e.g., personal phone numbers and addresses) are anonymized to ensure their safety.

We balanced GDPR guidelines and CNIL advice for deceased individuals by not anonymizing deceased public figures while anonymizing private victims, in order to respect their memory and privacy. Regarding convicted individuals and terrorists, we excluded well-known and deceased terrorists from anonymization, considered age at the time of the crime, and anonymized those not found guilty or who underwent legal name changes, especially if they were minors.

Data related to organizations. We have chosen not to anonymize the names of organizations as a general practice. However, exceptions were made when the organization's name could serve as an indirect identifier of individuals, particularly those belonging to vulnerable groups or who might be targeted for their opinions. These cases include family/small businesses, companies providing specific religious services, student organizations based on ethnicity or religion, and workplaces of activists. Additionally, names of radical organizations displayed as usernames or group/channel names on social media were anonymized while preserving relevant semantic information. For instance, "@ProudBoys-Massachusetts-admin" (fictional) was transformed to "@Proud_Boys_MA_main".

Data related to content sharing. In the dataset, content is typically shared through URLs and titles of media. When the content is considered too radical or too private to share, it is anonymized or invalidated as appropriate. This includes URLs redi-

recting to fundraising campaigns, personal blogs or websites of private individuals (e.g., Tumblr, WordPress), social media channels of radical groups (e.g., Telegram, Gab) along with their usernames, and URLs and titles of videos, movies, and songs produced by members of radical groups.

3.2 Pseudonymization Pipeline

Retrieval. The first step was to use a fine-tuned model to generate NER pre-annotations automatically. This initial version of named entity annotations helped to extract aliases, individuals, and organizations. The model was fine-tuned on ANERcorp (Benajiba et al., 2007; Obeid et al., 2020) for Arabic, FTB NER (Ortiz Suárez et al., 2020) for French, and CONLL2003 (Tjong Kim Sang and De Meulder, 2003) for English. Moreover, regular expressions were used to extract data that followed stable patterns, such as links, hashtags, and emails (Figure 2 in Appendix A.1 for the distribution of the categories). Simultaneously, we fixed the silver NER annotations to add another layer of NER with a large tagset (See Table 7in Appendix A.1).

Manual anonymization. One annotator per language manually anonymized the entities and corrected pre-annotations. After each decision of anonymization was made, it was added to a token-level correspondence table for the languages to ensure that an entity has the same replacement across languages. To maintain the cultural and stylistic integrity of the content while avoiding the disclosure of sensitive information, we attempted to choose pseudonyms mimicking the original names or aliases. This involved picking pseudonyms that shared a phonetic resemblance, incorporated special characters or numbers, considered linguistic nuances, included wordplay, maintained similar token length, or even incorporated details about the author's origins, perceived ethnicity and cultural references (see Table 6 in Appendix A.1).

In some special cases where anonymization is not needed, such as for links and some specific usernames, we use invalidation by adding changing characters. Re-identification can still be possible in these cases, but direct access is not.

Finally, we choose anonymization out of caution when in doubt[6].

Accounting for re-identification We carefully considered re-identification concerns, basing our anonymization efforts on established insights. Recognizing re-identification as a significant concern in PPDP, we accounted for the "disclosure risk" by considering the "background knowledge" a potential attacker might have, as described by Sánchez and Batet (2016, 2017). This background knowledge includes all web pages accessible through search engines. Consequently, our anonymization process considered all data types that could be used with search engines to identify an individual.

4 Experiments

In this section, we analyze the variation of the performance of the model in different scenarios and compare the use of anonymized data to original data for radicalization detection task.

4.1 Tasks

Radicalization Detection Task Our dataset includes English, French, and Arabic examples from various sources (Figure 1 in Appendix A.1), each with distinct characteristics. The English dataset contains messages from platforms like Telegram and forums, where radical groups promote their movements. The French dataset consists mainly of comments from social media platforms such as Twitter and Instagram, while the Arabic dataset primarily comprises religious texts focused on jihadism from sources like Facebook and Twitter. Those texts included a lot of deceased persons that were not anonymized. We had a different annotator for each language.

For our experiments, we focus on the annotation of *Call for Action Classification* for English and French as their sizes are comparable, which entails categorizing content into one of five predefined levels based on the degree to which it motivates specific actions, ranging from "negative" to "very high" (See Appendix A.1 for more details).

4.2 Substitutions methods

In this section, we evaluate our pseudonymization technique by comparing it to four methods from the existing literature (Jegga et al., 2013; Berg et al., 2020). We use metadata from our annotations to

[6]We did not calculate the inter-annotator agreement for the anonymization process, but we frequently discussed difficult decisions to ensure consistency. For NER, we calculated inter-annotator agreement with 100 randomly selected sentences in both English and French. The English annotator annotated 100 French sentences, and vice versa. The Cohen's Kappa Score for French was 0.9124 and for English was 0.8266, indicating a high level of agreement between annotators, suggesting closely aligned decisions.

	Train	Dev	Test
		English	
# examples	1735	194	484
# anonymized entities	1143	146	326
		French	
# examples	1888	210	526
# anonymized entities	485	51	158
		Arabic	
# examples	-	-	1500
# anonymized entities	-	-	130

Table 1: Statistics for English, French and Arabic

generate three additional anonymized dataset versions. The strategies we considered are as follows:

- **Entity Deletion (S0)** This method involves deleting the entity to anonymize it. While this approach maximizes privacy, it sacrifices data utility and coherence.
- **Uniform Placeholder (S1)** This method replaces all entities in the dataset with the same placeholder. It retains some data utility while ensuring anonymity but lacks category-specific differentiation.
- **Category-Specific Placeholder (S2)** Each category of entities (e.g., names, organizations) is replaced with a unique placeholder specific to that category across the dataset. This strikes a balance between anonymization and preserving some context-specific information.
- **Unique Placeholder per Entity (S3)** A unique placeholder is assigned to each entity in each document, maintaining sentence coherence while ensuring anonymity.

Table 2 shows the differences between the different automatic methods and our methods.

4.3 Model training

We fine-tune XLM-T (Barbieri et al., 2022), an XLM-R (Conneau et al., 2020) model that has been fine-tuned on 200 million tweets (1 724 million tokens) scraped between May 2018 and March 2020, in more than 30 languages. This model has been shown to be more adapted for social media data (Montariol et al., 2022). To ensure the reliability of our findings, we fine-tuned the model using five different seeds and reported the average performance across these five runs.

4.4 Results

For each language, we trained six models: four models for the automatically anonymized versions, one on the original data, and one on our anonymized version.

Table 3 reports the average macro-F1 scores over 5 seeds for each fine-tuned model, evaluated on both the corresponding pseudonymized and original test sets. Our approach resulted in a macro-F1 score of 65.46 for the English language models on the corresponding test set, which closely aligns with the highest score of 65.55 achieved by S3. This demonstrates the effectiveness of our method in maintaining data usefulness while ensuring robust anonymization. When evaluated on the original test set, our method achieved a score of 64.80, outperforming all other methods and slightly outperforming the model trained on the original data (64.63). This indicates that our method introduces minimal noise, thereby preserving data quality and coherence.

The performance of our pseudonymization technique shows different tendencies in the English and French language models. While our method performed consistently well for the English models, this trend was not observed for the French models. Our method demonstrated a good balance between anonymization and data utility for the French dataset. However, it did not consistently outperform other methods across the corresponding pseudonymized and original test sets.

The differences in trends observed between the French and English datasets can be attributed to the unique content and characteristics of the data for each language. The English dataset primarily consists of messages from platforms like Telegram and forums such as 4chan, where radical groups actively promote their movements and share propaganda. The figures (Figure 1 in Appendix A.1) further illustrate these differences, showing the diverse range of platforms for the English dataset and a higher proportion of radical content compared to the French dataset. As a result, it contains a significantly higher number of usernames and links that need to be anonymized. In contrast, the French dataset mainly includes posts from social media platforms like Twitter and Instagram. While personal data is less frequently encountered in the French dataset, it requires equal vigilance due to the presence of sensitive information, such as personal addresses and family business details. Table

Original	Hit me up @marie.delattre1, @handsomephilantropist on Insta. Shoutout to Moshe Chaya! At Rue Alphonse Metayer.
S0	Hit me up, on Insta. Shoutout to ! At.!
S1	Hit me up placeholder, placeholder on Insta. Shoutout to placeholder! At placeholder.
S2	Hit me up username, username on Insta. Shoutout to name! At location.
S3	Hit me up username11, usernsme22 on Insta. Shoutout to name44! At location55.
Ours	Hit me up @jane.doe1, @attractivehumanitariant on Insta. Shoutout to Raj Avrom! At Rue Hubert Couturier.

Table 2: Examples (Fictional) of different substitutions methods

Training data	Lang	Corresponding Test	Original Test
Original	en	-	64.63 (±2.0)
S0		62.11(±3.5)	60.81(±3.3)
S1		64.99(±1.5)	63.81(±1.1)
S2		62.34(±2.6)	59.91(±2.8)
S3		65.55(±1.6)	63.50(±1.4)
Ours		65.46(±1.0)	64.80(±2.2)
Original	fr	-	65.65(±1.8)
S0		64.13(±6.1)	66.78(±7.8)
S1		65.89(±4.1)	66.41(±5.4)
S2		63.52(±5.0)	62.31(±4.9)
S3		64.87(±4.2)	66.10(±4.5)
Ours		64.72(±4.8)	63.97(±4.3)

Table 3: Results for each fine-tuned model on the original training and the different anonymized training sets when **tested on the original test set (right)** and **the corresponding anonymized test sets (left)**. (Average Macro-F1 Scores over 5 Seeds)

Testing data	Lang	Macro-f1
Original	en	64.63(±2.0)
S0		62.93(±2.0)
S1		62.56(±2.1)
S2		63.41(±2.6)
S3		63.14(±1.9)
Ours		65.24(±2.7)
Original	fr	65.65(±1.8)
S0		65.57(±3.5)
S1		65.46(±3.8)
S2		65.69(±3.6)
S3		65.86(±3.5)
Ours		67.88(±2.3)

Table 4: Results for the model **trained on original data** and **tested on the test sets corresponding to different substitution methods** (Average Macro-F1 Scores over 5 Seeds)

1 shows the distribution of the categories for both languages and total entities for the test sets.

What to use for training? A commonly asked question after pseudonymization is, should we use the pseudonymized version for training? Does the added noise make the training more robust? Recent model attacks have demonstrated that it is possible to extract training data from a publicly shared model (Song et al., 2017; Carlini et al., 2021). To investigate this question, we report in Table 4 the results of models trained on the original training data and tested on each version of the pseudonymized test set similarly to Lothritz et al. (2023). We do not observe the same tendencies for both languages. For English, training on the anonymized train set (Table 3, corresponding test set column) gave better results than the counterpart model trained on the original data for almost half the models. While the results were inconsistent for English, we noticed that the original model performed consistently better in almost all cases when tested on the anonymized test sets for French. This suggests that the model learns more easily on the original data and generalizes well on the pseudonymized test sets.

Despite those trends, Brown et al. (2022) argue that language models should be trained on data that can be publicly published to guarantee privacy.

Even though it is not the main topic of this paper, we present in Table 8 in Appendix A.2 the results for the NER task on the original data and our anonymized data. We opted not to conduct experiments on the automatic substitution strategies because adding the category of the entity provides the named entity in the text, and removing it alters the token count, making the results non-comparable. We observe similar performance trends to the classification task with very close scores between the model trained on the original data and the model trained on our pseudonymized data.

5 Challenges

Public figures and influencers The lines between public figures, "influencers", and "private figures" are often blurred, making it challenging to determine if a journalist for a small news website should be considered a public figure. Similarly, categorizing scholars and less renowned authors

also poses difficulties.

Links redirecting towards radicalized content and far-right media websites It was often tough to decide what was to be anonymized for two reasons: the definition of "mainstream" can become entirely subjective, especially when a medium can be considered renowned in its circle but not enough for global recognition. Moreover, even when a medium is categorized as mainstream, leaving it as such still poses an ethical dilemma, as it can contribute to sharing propaganda.

Data related to terrorists and attackers In the English and mainly Arabic datasets, there were a lot of names of deceased terrorists, mainly from the Far-Right or from ISIS. While it is common for ISIS terrorists to have acquired names that do not always correspond to their birth names, and thus the risk of identification is lower, it is still a dilemma as to what should be left in the dataset.

6 Conclusion

In this paper, we presented our approach to pseudonymization specifically tailored for a radicalization dataset. Our method aimed to fill the gap in research on pseudonymization in sensitive domains, such as online radicalization. Our technique balances the need for privacy protection while maintaining the usefulness of the data for research and analysis. We highlighted the challenges encountered during the pseudonymization process, particularly the nuances of handling different types of personal data. These challenges underscore the importance of a detailed and cautious approach. Our multilingual radicalization dataset will be released upon publication. We advocate for developing a standardized framework for pseudonymizing sensitive NLP data. Overall, our work contributes to the growing body of research advocating for enhanced privacy measures in the processing and sharing of sensitive data, aligning with recent efforts to establish universal standards for categorizing and anonymizing personally identifiable information (Szawerna et al., 2024).

Limitations

Legal implications of pseudonymization Social media data processing and publishing cannot be exempt from anonymization techniques. Article 4 of GDPR defines pseudonymization as *"the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, [...]"*, which *" is kept separately and is subject to technical and organizational measures [...]."*. This "additional information" is often shaped through correspondence tables between the original data and its pseudonymized counterpart. Pseudonymization is recommended by GDPR (art.89) as an example of "appropriate safeguard[s]" to process personal data. Pseudonymization is not a completely fireproof method. According to the CNIL (2022) and GDPR, personal data can still be recovered by accessing the correspondence tables or tertiary data. Thus, since private information can theoretically be recovered, pseudonymized data still falls under GDPR.

Ethics Statement

This paper aims to outline the challenges encountered during the pseudonymization of this dataset. We share the resultant dataset as a scientific artifact in line with the principles of open science. We cannot stress enough This dataset cannot be used to train any radicalization model used in real ground conditions. Having been annotated by domain experts from different countries, it may contain biases that can harm different communities.

We recognize the sensitive nature of this work and stress the importance of striking a balance between privacy and effectiveness. We understand that the task of detecting radicalization is inherently subjective. Although we chose not to anonymize information about public figures, we took special care to anonymize contact and address information to prevent doxxing. For example, in one case from the English dataset, an individual with a somewhat public status in academia had their personal information -such as professional email addresses and phone numbers- revealed by the author of the post to incite harassment due to the individual's political beliefs. Despite the public status of the individual, we determined that it was too dangerous to keep this information in the dataset.

Note that the whole annotation process was particularly challenging for our annotators due to the violent, if not borderline traumatizing in some cases, nature of the data, which had an impact on their psychological well-being.

A mental health professional service and support from human resources services were made available to the team. A process dedicated to evaluating

the psychological impact induced by annotating this content was put in place. Its results (through extensive surveys—similar in depth to PTSD evaluation forms—and debriefing interviews) are currently under evaluation at our institution.

Acknowledgements

This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021607. The authors warmly thank the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Pierre Accorsi, Namrata Patel, Cédric Lopez, Rachel Panckhurst, and Mathieu Roche. 2012. Seek and hide: Anonymising a french sms corpus using natural language processing techniques. *Lingvisticæ Investigationes*, 35(2):163–180.

Allison Adams, Eric Aili, Daniel Aioanei, Rebecca Jonsson, Lina Mickelsson, Dagmar Mikmekova, Fred Roberts, Javier Fernandez Valencia, and Roger Wechsler. 2019. AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland. Linköping Electronic Press.

S. Aldera, Ahmad Emam, Muhammad Al-Qurishi, Majed Alrubaian, and Abdulrahman Alothaim. 2021. Online extremism detection in textual content: A systematic literature review. *IEEE Access*, 9:42384–42396.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*, pages 143–153, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy?

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models.

Özlem Çetinoğlu and Antje Schweitzer. 2022. Anonymising the SAGT speech corpus and treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5557–5564, Marseille, France. European Language Resources Association.

Jeremie Clos, Emma McClaughlin, Pepita Barnard, Elena Nichele, Dawn Knight, Derek McAuley, and Svenja Adolphs. 2022. PriPA: A tool for privacy-preserving analytics of linguistic data. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 73–78, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hercules Dalianis and Sumithra Velupillai. 2010. How certain are clinical assessments? annotating Swedish clinical text for (un)certainties, speculations and negations. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CodE alltag 2.0 — a pseudonymized German-language email corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4466–4477, Marseille, France. European Language Resources Association.

Elisabeth Eder, Michael Wiegand, Ulrike Krieg-Holz, and Udo Hahn. 2022. "beste grüße, maria meyer" — pseudonymization of privacy-sensitive information in emails. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 741–752, Marseille, France. European Language Resources Association.

James P. Farwell. 2014. The media strategy of isis. *Survival*, 56(6):49–55.

Miriam Fernandez and Harith Alani. 2021. Artificial intelligence and online extremism: Challenges and opportunities. -.

Louis Fink. 2014. Understanding radicalisation and dynamics of terrorist networks through political-psychology. *International Institute for Counter-terrorism*.

M. Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and K. Kotecha. 2021. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404.

HIPAA. 2004. *The Health Insurance Portability and Accountability Act*. U.S. Dept. of Labor, Employee Benefits Security Administration.

Anil Jegga, Imre Solti, Katalin Molnar, Keith Marsolo, Laura Stoutenborough, Louise Deleger, Megan Kaiser, Qi Li, Todd Lingren, Guergana Savova, and Fei Xia. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, 20(1):84–94.

Thomas J. Lampoltshammer, L″orinc Thurnay, and Gregor Eibl. 2019. Impact of anonymization on sentiment analysis of twitter postings. In *Data Science – Analytics and Applications*, pages 41–48, Wiesbaden. Springer Fachmedien Wiesbaden.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Harald Lüngen, Michael Beißwenger, Laura Herzberg, and Cathrin Pichler. 2017. Anonymisation of the dortmund chat corpus 2.1. In *Proceedings of the 5th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora17)*. cmc-corpora conference series.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1051–1056, Genoa, Italy. European Language Resources Association (ELRA).

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

Stéphane Meystre, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014. Can physicians recognize their own patients in de-identified notes? *Studies in Health Technology and Informatics*, 205:778—782.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, 8(1):32.

T. Nguyen and X. Vu. 2023. Privacy and trust in iot ecosystems with big data: A survey of perspectives and challenges. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 215–222, Los Alamitos, CA, USA. IEEE Computer Society.

Mariam Nouh, Jason R. C. Nurse, and M. Goldsmith. 2019. Understanding the radical mind: Identifying signals to detect extremist content on twitter. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 98–103.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for

Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Annika Willoch Olstad, Anthi Papadopoulou, and Pierre Lison. 2023. Generation of replacement options in text sanitization. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 292–300, Tórshavn, Faroe Islands. University of Tartu Library.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France. European Language Resources Association.

Anthi Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 217–229, Online only. Association for Computational Linguistics.

Antonio Pellicani, Gianvito Pio, Domenico Redavid, and Michelangelo Ceci. 2023. Sairus: Spatially-aware identification of risky users in social networks. *Information Fusion*, 92:435–449.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.

Ramit Sawhney, Atula Tejaswi Neerkaje, Ivan Habernal, and Lucie Flek. 2022. How much user context do we need? privacy by design in mental health nlp application.

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 587–601, New York, NY, USA. Association for Computing Machinery.

Latanya Sweeney. 2000. Uniqueness of simple demographics in the US population. LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh.

Maria Irena Szawerna, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Xuan-Son Vu, and Elena Volodina. 2024. Pseudonymization categories across domain boundaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13303–13314, Torino, Italy. ELRA and ICCL.

David Sánchez and Montserrat Batet. 2016. C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology*, 67(1):148–163.

David Sánchez and Montserrat Batet. 2017. Toward sensitive document release with privacy guarantees. *Engineering Applications of Artificial Intelligence*, 59:23–34.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.

Javier Torregrosa, Gema Bello Orgaz, Eugenio Martínez Cámara, Javier Del Ser, and David Camacho. 2021. A survey on extremism analysis using natural language processing. *CoRR*, abs/2104.04069.

Thomas Vakili and Hercules Dalianis. 2022. Utility preservation of clinical text after de-identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388, Dublin, Ireland. Association for Computational Linguistics.

Thomas Vakili and Hercules Dalianis. 2023. Using membership inference attacks to evaluate privacy-preserving language modeling fails for pseudonymizing data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 318–323, Tórshavn, Faroe Islands. University of Tartu Library.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Sumithra Velupillai, Hercules Dalianis, Martin Hassel, and Gunnar H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and f-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, 78(12):19 – 26.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hui Yang and Jonathan M. Garibaldi. 2015. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:30 – 38.

Oleksandr Yermilov, Vipul Raheja, and Artem Chernodub. 2023. Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Datasets

Each document of the original dataset is annotated with different information. We describe here the **Call for Action levels** that indicates whether a specific content should be flagged:

- **Negative (No Call for Action)**: Content that exhibits no indications of radicalization or encouragement of extremist activities.
- **Low Call for Action**: Content that expresses radical views or ideologies without explicitly advocating for violence or extremist actions. This may include mere approval of extremist actions or actors.
- **Moderate Call for Action**: Typically involves content that subtly suggests participation in extremist activities or ideologies but stops short of direct advocacy.
- **High Call for Action**: Content that demonstrates clear support or admiration for extremist groups or indicates involvement in such groups' activities, likely inciting further radical actions.
- **Very High Call for Action**: Represents the most extreme level, where content explicitly calls for violent action against individuals or groups.

Figure 2, Figure 1, Table 5, Table 6 and Table 7 represent statistics on our dataset and details about the annotations layers.

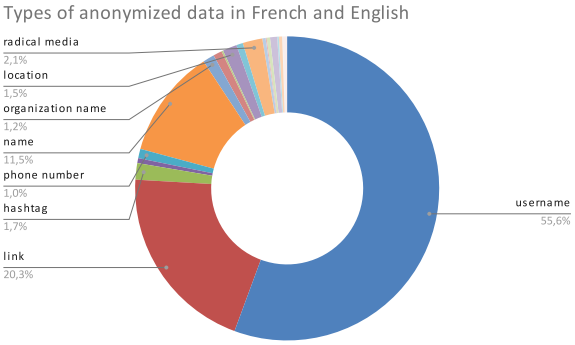


Figure 2: Types of anonymized data in French and English

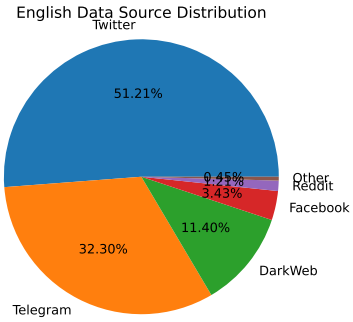
	English	French	Arabic
PER	2234	1802	4100
LOC	1783	1496	1656
ORG	1963	681	637
OTH	613	783	180
COMP	58	122	6

Table 5: Named entity repartition in the datasets.

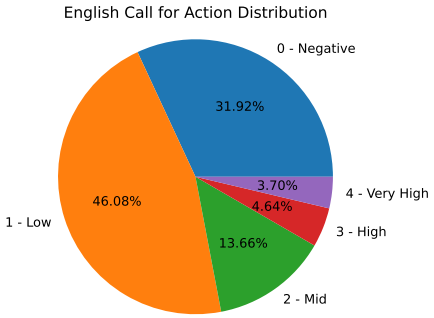
Original	Replacement
Myriam Zegman	Rachel Kaufman
Virginia	Mary
Muhammed	Ahmed
@MaryJohanson1987	@LaraWilson1989
https://wa.me/+93722758	https://wa.me/+93824556

Table 6: Examples (fictional) of replacements

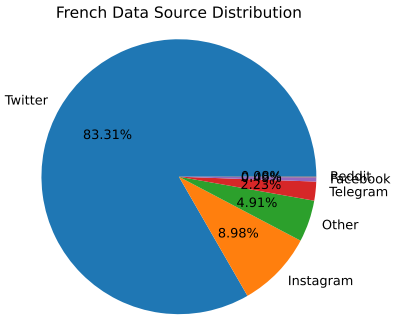
A.2 Additional Results



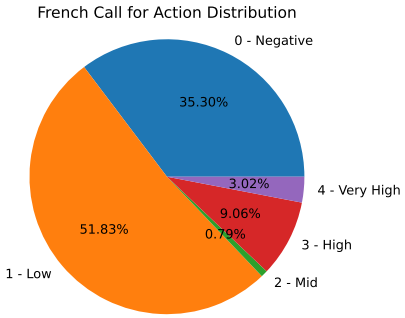
(a) English Data Source Distribution



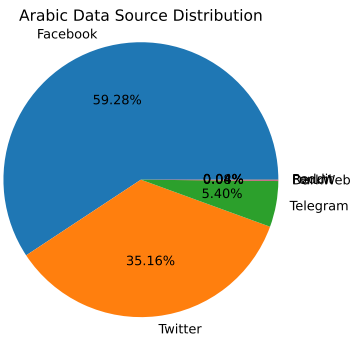
(b) English Call for Action Distribution



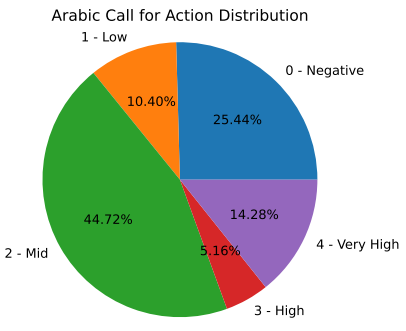
(c) French Data Source Distribution



(d) French Call for Action Distribution



(e) Arabic Data Source Distribution



(f) Arabic Call for Action Distribution (

Figure 1: Data source and call for action distributions for English, French, and Arabic

Label	Description
PER	mentions of names, aliases, and hashtags when they refer to a single person or user
PER:IMG	Fictional characters from manga, movies, books, and common culture.
PER:REL	References to individuals existing in a religious representation of the world.
COMP	Mentions of commercial enterprises and companies.
LOC	Mentions of locations, including neighborhoods, cities, and countries.
LOC:IMG	Fictional places.
LOC:REL	Religious locations.
ORG	Political, educational, or association-like organizations.
ORG:MEDIA	Media organizations, including radio or TV shows, podcasts, and newspapers.
OTH:BOOK	Books, mostly religious texts such as the Quran and the Bible.
OTH:GAME	References to games with mentions like "Minecraft."
OTH:MOVIE	Movies and series.
OTH:MUSIC	Musical entities, with mentions like "La isla Bonita."
OTH:DIS	Diseases.
OTH:SYMB	This category encompasses symbolic entities, including representations like the "Swastika" and religious symbols like the "Étoile de David."
OTH:EVENT	Reserved for recurring events, historical events, and religious events
OTH:CONSPI	This category is dedicated to concepts related to conspiracy theories.

Table 7: List of Named Entities used for the NER annotation layer.

Training data	Lang	Corresponding Test	Original Test
Original	en	-	87.04(±0.6)
Ours		87.01(±0.5)	86.83(±0.5)
Original	fr	-	78.96(±1.9)
Ours		78.96(±1)	78.01(±1.1)

Table 8: NER results for each fine-tuned model on the original training and our anonymized training sets when **tested on the original test set (right)** and **our anonymized test set (left)**. (Average Macro-F1 Scores over 5 Seeds)

Testing data	Lang	Macro-f1
Original	en	87.04(±0.6)
Ours		86.01(±0.8)
Original	fr	78.96(±1.9)
Ours		77.87(±1.5)

Table 9: NER results for the model **trained on original data** and **tested on our anonymized test set** (Average Macro-F1 Scores over 5 Seeds)