Extending off-the-shelf NER Systems to Personal Information Detection in Dialogues with a Virtual Agent: Findings from a Real-Life Use Case

Mario Mina **Carlos Rodriguez-Penagos** **Aitor Gonzalez-Agirre** **Marta Villegas**
Barcelona Supercomputing Center
{mario.magued|carlos.rodriguez1|aitor.gonzalez|marta.villegas}
@bsc.es

Abstract

We present the findings and results of our pseudonymisation system, which has been developed for a real-life use-case involving users and an informative chatbot in the context of the COVID-19 pandemic. Message exchanges between the two involve the former group providing information about themselves and their residential area, which could easily allow for their re-identification. We create a modular pipeline to detect PIIs and perform basic deidentification such that the data can be stored while mitigating any privacy concerns. The use-case presents several challenging aspects, the most difficult of which is the logistic challenge of not being able to directly view or access the data due to the very privacy issues we aim to resolve. Nevertheless, our system achieves a high recall of 0.99, correctly identifying almost all instances of personal data. However, this comes at the expense of precision, which only reaches 0.64. We describe the sensitive information identification in detail, explaining the design principles behind our decisions. We additionally highlight the particular challenges we've encountered.

1 Introduction and Context

With current advances in NLP relying on data-hungry machine learning systems and even more data-hungry language models, user-generated data is becoming increasingly important: data from conversations with chatbots, crawls of internet forums, posts on social media, etc can and are often used to train deep learning systems. At the same time, respecting user privacy is critical.

The General Data Protection Regulation (GDPR) came into effect as of the 25th of May 2018, affecting any data identifying or allowing the identification of a natural person. For instance, in the previous examples of user-generated data, identifiable data could take the form of a username, a full name, or an address, among others (Francopoulo and Schaub, 2020). At its core, the aim of the GDPR is to bring EU data protection legislation in line with the new ways that personal data is now being used by giving users more control over the ways their data is bring processed.

One of the implications of the GDPR is for there to be no way to trace data back to a specific individual or a group thereof. As a result, anononimized data is exempt of GDPR requirements. In turn, much effort has gone into perfecting anonymization and pseudonymisation techniques to allow NLP practitioners to work directly with user-generated data.

However, each domain presents its own unique challenges. In this paper we tackle anonymization in user-generated messages with a virtual chatbot. Text originating from this domain presents the same characteristics as other instances noisy user-generated text; we encounter different types of text, with some of the messages being characterised with non-standard spelling, use of slang, etc, while others are written in a formal register (Barbieri et al., 2020; Baldwin et al., 2015a). Furthermore, information is exchanged between the user and the virtual agent in a dialog fashion, such that it is possible for no individual message to allow the identification of the user, but the conversation, taken as a whole, could.

In this paper we describe findings from our particular real-life scenario of automatically identifying PIIs in user-generated data from conversations involving a virtual agent serving as an informative tool while not being able to directly access the data. Users adhering to the contemplated use-case could use the virtual assistant to make inquiries regarding COVID restrictions in their area of residence. Such exchanges are a perfect example of personal information that can be used to identify an individual based on their location.

As stated, different domains present different challenges for anonymization. With this in mind,

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 44–53

we design a flexible and modular pipeline[1] to anonymize GDPR protected text by allowing for different components that perform sensitive data identification and subsequent deidentification. We describe our experimental setup and methods used, and highlight particularly difficult aspects of working on real-life user-generated data in both Spanish and Catalan that we could not directly access.

2 Literature Review

The task of pseudonymisation is generally considered to be complex given that based on context, one can re-identify pseudonymised information and the. These difficulties can be in turn modulated by each domain's characteristics. Up until recently, most techniques were applied in either medical or legal domains, which were considered to be sensitive domains well before the GDPR (Sánchez-León, 2019; Langarizadeh et al., 2018; Yuwono et al., 2016). Methods typically vary between applications; generally speaking, pseudonymisation occurs in at least two steps: the first step is identifying personal information, where most of the our efforts in this paper are centered. Most methods that are applied to highly regular data rely on simple regular expressions, whereas less structured information requires more sophisticated Named Entity Recognition (NER) systems based on machine learning. Deidentification can vary more in terms of applicable methods, and is more dependent on the properties of the source text. That is to say, there are different methods that are more of less preferable depending on the use case (Belkadi et al.). Typically methods involve substituting sensitive information with a random sequence, a label, or a random entity of the same or similar type.

Nevertheless, Adams et al. (2019) posit that the need for robust anonymization is being extended to other domains, due to the GDPR affecting other sources of data, which has made the task of automatic text pseudonymisation more relevant than ever. To that end they develop a machine learning-based toolkit to perform automatic pseudonymisation in human-computer dialogue while taking into account information that could potentially identify persons (PIIs) but also corporations (CIIs).

[1] https://github.com/langtech-bsc/AnonymizationPipeline

2.1 Regular expression-based sensitive information identification

Hassan et al. (2019) create ReCRF, a named entity extraction system that extracts features based on orthography, lexis and regular expressions from a specific token and its surrounding context to classify a token as containing PII or not in medical text. The interesting aspect to their feature crafting method is the use of a data-driven method to automatically generate regex-based rules. These features are then used as input to Conditional Random Field models.

Still involving the medical domain, Sánchez-León (2019) develop a pseudonymisation system for Spanish clinical text. They enrich a simple grammar formalism with regular expressions to take into account spelling variations and then apply each rule in order of reliability, with generally favourable results.

Yuwono et al. (2016) apply regular expressions similarly to detect PIIs in clinical discharge papers. On top of the regular expressions they construct hand-crafted heuristics involving minimum edit distance to account for spelling and formatting inconsistencies between documents. Their simple heuristics-based approach does not require any sort of fine-tuning, model training, or manual annotation, but they do make use of their own database when detecting patient information.

2.2 Machine learning-based detection of sensitive data

A variety of machine learning methods can be utilised in several ways when detecting sensitive information. Juez-Hernandez et al. (2023) perform a comprehensive assessment of PII detection methods using current state-of-the-art methods and propose a few of their own, with a focus on several languages. They perform several experiments to derive optimal solutions for PII identification in different types of Spanish text (clinical texts and law-enforcement reports). They pose different research questions regarding the performance of NER models of PII detection. Specifically, they contemplate the effects of using off-the-shelf models on performance in comparison to training a model for each specific domain, as well as if an ad hoc trained model can be used in a cross-domain fashion. Their findings suggest that while off-the-shelf models can be used for PII detection, training domain-specific models yields superior results, given the variabil-

ity across domains. Specifically, they note that a model trained on one domain can be used in another with acceptable performance, but performance will degrade when used on out-of-domain data.

In terms of methods, they test different NER architectures ranging from off-the-shelf Stanza (Qi et al., 2020) and Flair (Akbik et al., 2019) NER models to recurrent neural architectures with different combinations of embeddings, to pretrained transformers available on HuggingFace that were then fine-tuned on their task-specific data.

2.3 Challenges

Examining the requirements of the GDPR, and what constitutes genuinely anonymized data, a question that is continually asked is how do we manage the trade-off between privacy and utility? Francopoulo and Schaub (2020) determine that for an anonymization framework to be successful, it needs to: (1) avoid identifying the individuals in the text, (2) allow posterior analysis of the anonymized text, (3) allow for off-the-shelf NLP tools to be applied to the anonymized text, (4) produce a provable anonymization, (5) be usable in different European languages. They highlight that some of these are contradictory or at least that some requirements directly interfere with the effectiveness of the others, even if we assume a perfect detection of PIIs. The problem lies in that if resulting data from an anonymization or deidentification process is indistinguishable from a non-anonymized text, there would be no way to prove that it has actually been anonymized. For instance, anonymization by redaction (i.e. the elimination of PIIs by substituting them with a fixed character such as an *X*) leaves proof of the anonymization, but severely limits any posterior usability of the text. On the other hand, if a more sophisticated substitution is applied to the text, the result maximises posterior usefulness, but by definition should not leave a trace of the anonymization. As a solution they propose a relaxation of requirements based on the specific circumstance. They argue that requirement (3) is vital when using off-the-shelf tools within a secure environment, where requirement (4) can be relaxed, while requirement (4) is more important outside of a secure environment, where requirement (3) can be relaxed.

These concerns are echoed in Mozes and Kleinberg (2021). They argue that current methods do not correctly quantify anonymization performance, given that if a text contains several instances of PIIs, it is enough for one of them to go undetected to identify the person in question. Many metrics would still assign a high performance to the anonymization system, as evaluation is typically applied on a sentence or instance level, despite the anonymization essentially failing.

They propose specific evaluation criteria to measure the effectiveness of the anonymization. The criteria presented in TILD take into account an anonymization system's **t**echnical performance, the **i**nformation **l**oss resulting from the anonymization, and the human ability to **d**eanonymize the redacted documents. They highlight the importance of information loss and robustness against de-anonymization; to guarantee posterior utility, the authors argue that the anonymization process must introduce as minimal changes as possible to the original document such that utility loss (difference in performance when using anonymized data in comparison to the original data) and construct loss (difference according to a higher order construct) are minimised. However, while ensuring minimal differences between anonymized and original texts, the anonymization process must be irreversible, such that a human intruder with the ability to use external resources would not be able to identify the original PIIs.

We can draw parallelisms between Francopoulo and Schaub (2020) and Mozes and Kleinberg (2021). Both papers highlight the importance of the actual detection component (requirement (1) of Francopoulo and Schaub (2020) and criterion T of TILD), and both are concerned with the posterior utility of the data in terms of the analyses that can still be carried out (requirements (2) and (3) in Francopoulo and Schaub (2020) and criterion IL of TILD). However, we observe that Francopoulo and Schaub (2020) suggest modulating the importance of that requirement based on intended use and level of exposure of the anonymized data, while Mozes and Kleinberg (2021) make no such statement. After that point, both criteria diverge. Francopoulo and Schaub (2020) highlight the importance of having a provable anonymization, while Mozes and Kleinberg (2021) place more emphasis on the anonymization being non-reversible while maintaining the properties of the original data.

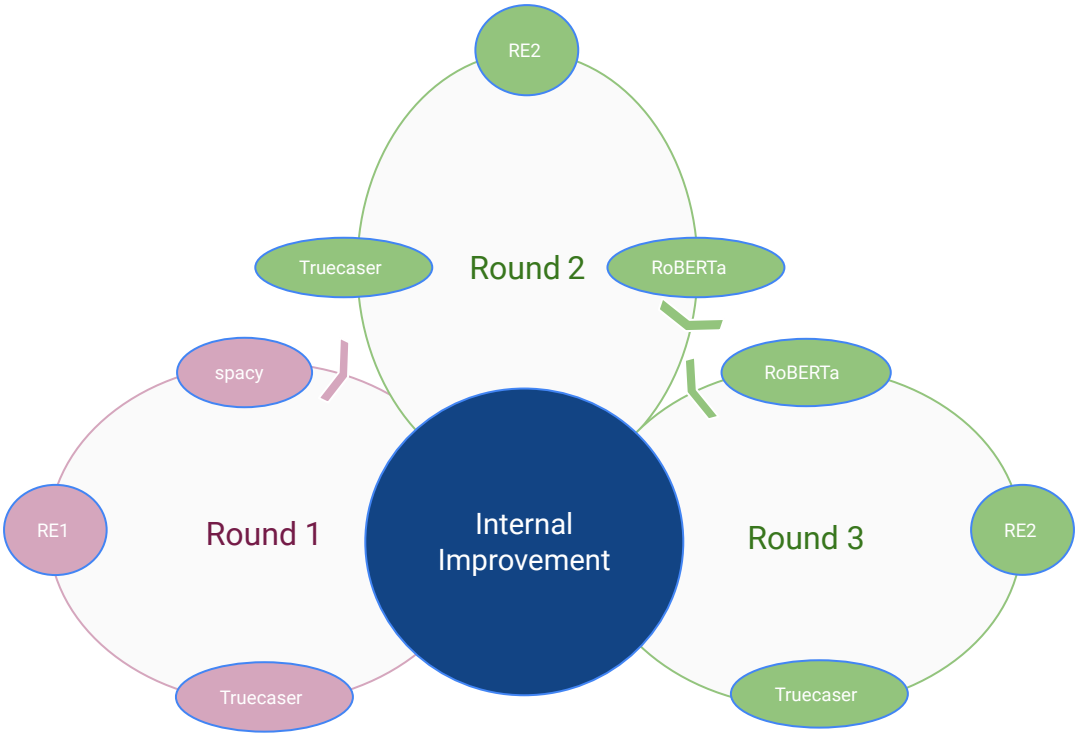

Figure 1: Diagram of our experimental design. round 1 was performed on the randomly sampled dataset (left, in pink), while rounds 2 and 3 were performed on the entire conversation dataset (top and right, in green).

3 Methods

3.1 Anonymization Data

As stated in section 1, our evaluation data originates from conversations between users and a virtual agent. For context, the conversations took place during the COVID-19 pandemic. We examine two subsets of the data. Initially, we randomly sample 23.000 messages for simplicity. However, after our initial assessment, discussion with annotators, and following Mozes and Kleinberg (2021), we decide to include full conversations, given that in some occasions, referents can be identified using contextual cues that do not individually constitute PIIs.

We decide to include messages from full conversations such that the individual messages sum to 23.000; annotators were instructed to compile a second dataset such that all messages sent by the users from a conversation were included. This resulted in the generation of a second evaluation dataset consisting of 23.000 messages from 953 unique conversations. We highlight that due to privacy restrictions, we only use the data for evaluating our system, as the data cannot be used to train or fine-tune a base model.

Regarding the annotation process, the data was selected and processed by two annotators, and then revised by a third such that the third annotator could essentially act as a tie-breaker. Cases where no consensus was reached were excluded from the experiments (this was the case for fewer than 15 messages in total, taking into account both datasets)

In terms of structure, only messages from the users are included. Messages are assigned two identifiers: a unique identifier and an identifier specifying to which conversation it belongs. Within each conversation, messages are ordered chronologically. Furthermore, the messages are unlabeled. We do not explicitly work with a gold standard. Instead, we rely on the annotators to examine the data on our behalf. They additionally analyse the performance of our system by checking what information is correctly pseudonymised and which information is incorrectly pseudonymised.

Our setup is as shown in Figure 1. The data is kept by the third party such that we cannot directly access or manipulate it. Given this data access constraint, the annotators were hired to perform three evaluation rounds. In each round, we submit a version of the pipeline which is run on the third party's systems. They in turn evaluate the PII identification

component performance.

3.2 The Anonymization Pipeline

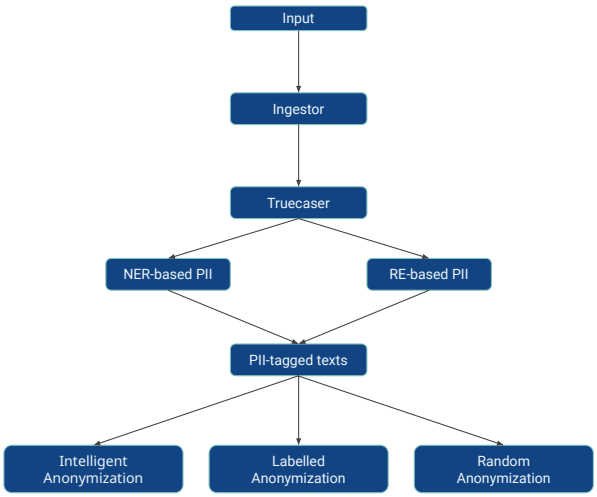

Figure 2: Diagram of the presented pipeline.

Despite the complexity of the task of pseudonymisation, the pipeline we present in this paper is relatively straightforward. In this subsection we proceed to describe our pipeline as shown in Figure 2.

Input data is provided textually. Currently, the pipeline supports different forms of textual input through different ingestors that, when fed a specific format, would output the data in a normalized format that the rest of pipeline can manipulate. The pipeline currently has ingestors for .csv, .json and .txt formats.

As explained in section 2, user-generated data is known to be noisy; while we do not explicitly add a preprocessing module, we have empirically determined during initial testing that in our specific case, performance is hindered by poor textual formatting. To mitigate this issue, we apply an implementation of the NLTK truecaser module[2] (Bird et al., 2009). This resolves simple cases where names of people, locations, organizations, etc are incorrectly cased.

Following ingesting the data, normalizing the input representation, and applying mild preprocessing, we proceed to the sensitive information identification task by combining regular expressions and machine learning models as described in subsection 3.3. While we take into account the labels provided by our regular expressions and one NER module, we highlight that the pipeline supports the use of multiple NER modules. In the case of a mismatch between any of the components, we establish a ranking such that the labelling of one can be determined to be more "trustworthy" and therefore take scope over the other in case of a discrepancy.

Once the sensitive information has been identified within the text, the pipeline can perform simple deidentification. Three methods are included: *random*, *labelled*, and *intelligent*. The *random* method substitutes the sensitive span with a series of random characters of varying length, the *labelled* method substitutes the span with the category of sensitive data detected (e.g. PERS, ID, LOC, etc). The *intelligent* method performs a limited substitution that attempts to substitute the marked span with a different entity of the same category. Currently, it is entirely possible for the *intelligent* anonymization system to substitute a street name with a city for instance, as we do not have a more fine-grained method of PII detection available. We conduct all of our experiments with the *labelled* setting to ease the annotation task.

3.3 PII identification

Given that pseudonymisation is a complex task and PIIs can occur in varying contexts, our pipeline is designed with flexibility and modularity in mind, such that components can be substituted based on the requirements and difficulty of the task. We differentiate between structured and non-structured PIIs and detect them using different methods; instances of structured data include phone numbers, zip codes, emails, etc. On the other hand, non-structured PIIs could include person and location names.

We take into account the domain properties of our domain of intended use. User-generated text is notorious for being noisy (Baldwin et al., 2015b; Jose and Raj, 2014). This can harm the robustness of PII detection module of our pipeline, increasing the number of false positives and negatives. We mitigate these problems in our pipeline differently for structured and unstructured data.

According to the intended use-case of the virtual agent, users are expected to provide information regarding their location, identity and contact in the form of structured PIIs of zip codes, ID number, email, phone number, and even land registry

[2] https://huggingface.co/HURIDOCS/spanish-truecasing

identifiers. To detect this sort of information we hand-craft regular expressions to match such information, allowing for some variation by users (e.g. a missing digit in a phone number, lower-case letters rather than upper-case letters in a license plate number).

For non-structured data such as location names and full names whose formats can vary, we use machine learning and deep learning methods for detection. We experiment with a large spacy model[3] and a RoBERTa NER[4] model that have been fine-tuned on Catalan and Spanish NER data. We additionally experiment with a truecasing module to ease the detection of named entities.

During each of the three evaluation rounds, we ask the annotators to classify each message using criteria that we provide. We establish the following typology to categorise PII detection performance:

A) Information that should not be anonymized (false positives)

B) PIIs that should have been detected but were not (false negatives)

C) PIIs that have been detected but assigned an incorrect type (true positives)

D) Correctly identified PIIs (true positives)

E) Potential PII but not in this context (true positives)

F) Not PIIs (true negative)

Many PIIs are contextually modulated, in the sense that the same span of text may allow the identification of the individual depending on the information in the surrounding context. For instance, a first name on its own might not identify an individual, but a full name probably would, and both would be detected by most off-the-shelf NER models. Similarly with locations, a user stating their city of residence may not be providing sensitive information. However, the likelihood of being able to identify the user increases the fewer the inhabitants that live in the area denoted by the message. Given our limited access to the data, we cannot use any of the messages to fine-tune a model and tailor it to our specific domain, and must rely on models trained on other datasets. This limits our system's ability to take this contextual modulation into account.

That being said, we still instruct the annotators to take into the account all messages sent by the user during the exchange in line with the points raised in Mozes and Kleinberg (2021) and the TILD evaluation framework; in one of our evaluation paradigms, if our system fails to detect critical PIIs that allow the identification of the individual, the entire exchange is labelled as B). We additionally instruct the annotators to highlight instances where users specify entities that would typically be detected by NER systems, but do not constitute PIIs, thereby creating category E). As stated, the models we use in this case are not specialised in anonymization, and therefore they are unable to pick up on explicit contextual cues that allow distinguishing PIIs from named entities (NEs). In light of this and that that sensitivity of specific entities is contextually modulated, for our evaluation we still consider them to be true positives. Similarly, we also consider correctly detected PIIs that are not correctly categorised (e.g. a location that is classified as a name) to be true positives, given that our main focus is PII identification.

3.4 Evaluation Rounds

Cleaning and aggregation On one hand, we believe that whatever PII detection system that is applied or deployed in a given environment should show robust performance despite noisy input. But on the other, in our specific case, without being able to adapt a model to this type of task and domain, the noise in the input negatively skews our results, both in terms of performance and evaluation. With the aid of the annotators, we identify two main issues:

1. Much of the input is noisy. Many users will misspell several words (e.g. *weno chao* instead of *bueno ciao* to end a conversation) in their messages or simply send nonsense (e.g. button mashing or sending the same random characters multiple times) to the virtual agent, which is detected by the model

2. Some users send several instances of the same message. If PII detection of that specific message is incorrect, it is then overrepresented in the data

Essentially, the first problem causes the model to detect several false positives through errors of

[3] https://huggingface.co/PlanTL-GOB-ES/es_anonimization_core_lg

[4] https://huggingface.co/BSC-LT/roberta_model_for_anonimization

type A). By sending several copies of some messages, the first problem is essentially exacerbated, such that false positives are overrepresented in our evaluation.

To mitigate this problem, we first detect poorly formatted or spelled messages similarly to Kudugunta et al. (2023); we apply the fasttext language identifier (Joulin et al., 2016) to each message. The language identifier outputs a probability distribution over languages. Poorly formed messages will have a lower probability associated with the expected languages. We discard any message with a probability lower than 0.8 of being either in Spanish or Catalan. Furthermore, in addition to evaluating performance by considering each message individually, we also examine performance by considering entire conversations. That is to say, rather than assign labels to individual messages, we assign them to the entire conversation. We do this by establishing a hierarchy of error types, such that graver errors take higher scope. The hierarchy is as follows: B > A > C > D > E > F. For instance, if in a conversation, one message is correctly anonymized (i.e. type D), but a critical PII is missed in another message belonging to the same conversation(i.e. type B, or a false negative), then that whole conversation is marked as B.

As stated in subsection 3.1, we do not have direct access to the data and instead provide the task to a third party. We iteratively make improvements to our pipeline based on their feedback. In Figure 1 we illustrate how we proceed through each evaluation round. We perform three sequential rounds of evaluation. During each round, we update the pipeline and make the new version available to the third party. The pipeline is then downloaded and run on their systems where the data is kept. For readability, model performance is based on the label-based pseudonymisation. Model performance is manually examined by comparing the original data with the anonymized data to examine if PIIs were correctly detected and replaced with the correct labels. The results of the examination are then forwarded back to us in terms of the error typology presented in the beginning of subsection 3.1. This feedback is then taken into account for the following round of evaluation.

Round 1 For the first round of evaluation, we experiment with lightweight approaches. We use a large spacy NER pipeline (which includes POS tagger, dependency parser, attribute matcher, and lemmatizer) (Honnibal et al., 2020). Initial experiments in-house additionally showed a benefit in performance by adding a truecaser as a preprocessing step. We use our initial set of regular expressions (RE1).

Round 2 For the second evaluation, we take into account the results and feedback from the second round and include a larger and more robust RoBERTa NER model to increase the quality of the PII detection. We additionally perform in-house experiments to determine if the truecaser adds any benefit and decide to still include it.

Round 3 After the second round, we observe that our system manages to detect the majority of the PII instances in the evaluation set. However, discussion with the annotators revealed that some instances were not detected due to user error (e.g. a phone number missing a digit, a misspelled email). We refine the regular expressions such that they are more flexible to account for user error (RE2). We additionally observe that the truecaser introduced whitespaces in specific contexts which interfered with the RoBERTa model tokenization, negatively impacting precision. We resolved this issue for the third and final round with the aim of reducing the number of false positives.

We show the results for each round in Table 1, presenting precision, recall and F_β ($\beta = 2$)for each round. We additionally present results for the datasets after filtering our the noisy text and assigning a label to each conversation, rather than each individual message.

4 Results

We present our results in Table 1. Within each round, we evaluate the effects of data cleaning and applying our evaluation metrics in different ways; we explore the effects of aggregating the data differently (as shown in the *Aggregation* column), and the effects of removing poorly formatted messages from consideration (expressed by the *-c* (for clean) suffix). Each round is separated by a horizontal line in the table. Cleaned and non-cleaned versions of the data are separated by a dashed line within each evaluation round.

In spite of not being able to directly access the data, Table 1 shows the clear benefits of our iterative evaluation paradigm. We can observe non-trivial improvement from one round to the next; the first round, using the Spacy model, yields moderate

Round	NER component	RE set	Aggregation	Precision	Recall	F_β
R1	Spacy	RE_1	Total	0.43	0.74	0.65
	-	-	-	-	-	-
R2	RoBERTa	RE_1	Total	0.06	0.95	0.23
			CONV	0.27	0.93	0.62
R2-c	RoBERTa	RE_1	Total	0.1	0.95	0.35
			CONV	0.29	0.93	0.64
R3	RoBERTa	RE_2	Total	0.40	0.99	0.77
			CONV	0.63	0.99	0.89
R3-c	RoBERTa	RE_2	Total	0.54	**0.99**	0.85
			CONV	**0.64**	0.99	**0.90**

Table 1: Results as classified by annotators for each evaluation round. Best performance in bold. β = 2. RE_n indicates the set of regular expressions used, whereas the *-c* suffix indicates that noisy messages have been removed from the dataset.

precision but relatively low recall. For the second round, we incorporate a more robust RoBERTa model into the pipeline, which drastically raises recall at the cost of precision. For the third and final round, we modify the system tokenization scheme and augment the set of regular expressions, further improving both recall and precision, ultimately yielding the highest F_β-score of 0.90.

Furthermore, we can see the clear impact of the data quality on pipeline performance. For each evaluated dataset, we create a *clean* counterpart after filtering out messages we believe have significant orthography or formatting issues. We observe superior model performance on cleaner datasets, especially in terms of precision. We observe this effect in evaluation rounds 2 and 3.

However, we observe a much stronger difference in precision based on the way we choose to aggregate the data; by aggregating the data by conversation we observe major improvements, which are more representative of actual PII detection performance. That is to say, by assigning a single label to each conversation based on whether a correct detection of PIIs was carried out, we attain much better precision. We observe these effects in both rounds where we collected several messages from the same conversation (rounds 2 and 3).

5 Discussion

PII detection performance The results shown in Table 1 in section 4 show clear improvement between consecutive iterations. In terms of trade-off between precision and recall, we note that performance is most balanced using the Spacy model. However, it does also present the lowest recall, which we consider to be the to be the most relevant metric given the sensitive nature of the data. In light of this, we find recall to be prohibitively low using the Spacy model. Comparing performance between the Spacy and RoBERTa models in similar conditions (i.e. considering individual messages and unclean data), it is clear that the RoBERTa models show lower precision. That said, their higher recall makes them more desirable in this context.

Our findings are in line with those of Juez-Hernandez et al. (2023). While we are able to achieve high recall with models trained on out-of-domain data, we do observe that performance is not optimal, given the relatively low precision of the RoBERTa models. That said, we have been able to determine that the low performance is largely due to the overrepresentation of noisy input in the data, which essentially interacts with the imperfect robustness of our model in this context, contributing to the deflation of the aforementioned metric. We have more or less mitigated this issue so that the results more accurately reflect model performance, but we also highlight that if the NER models were more robust to the noise in the input, the number of false positives would be drastically lower.

Francopoulo and Schaub (2020) and TILD Given the high recall of our system, we consider that we fulfill the first item of the criteria presented in TILD (**T**echnical performance) and Francopoulo and Schaub (2020), which is to ensure that the data does not contain PIIs. However, we do note the low precision of our system may negatively impact **I**nformation **l**oss, as obscuring more data than necessary may render the data less useful. On the

other hand, tagging more entities than necessary as PIIs and their subsequent anonymization (via redaction or substitution) is more likely to make deanonymization much more difficult. In light of this, we argue that depending on use, the system we present in this paper could be more than adequate.

6 Conclusion and Future Work

In conclusion, we present the results and findings from a real life use-case where we have had to develop a PII detection system to pseudonymise exchanges between users and a virtual agent. We demonstrate the effectiveness of our system and the issues that can arise when extending a NER system beyond its original domain. We highlight the specific problems we have encountered with user-generated data. We additionally note that given the differences between the domains of training and deployment, our system performs well, achieving very high recall. We argue that the inter-domain differences may be detrimental to performance in general, PII detection can be achieved with robust off-the-shelf NER models, given that our system managed to detect almost all instances of PII.

While the performance of our system is more than adequate given the circumstances, anonymisation and pseudonymisation are tasks that are gaining more and more urgency and importance. In light of this we consider it of critical importance to develop more resources for domain-specific and domain-general pseudonymisation.

The focus of this paper has been examining the effectiveness of adapting off-the-shelf NER systems to the task of PII detection. Our future work should aim to address explore robust ways of deidentifying the data in accordance with the established literature (Francopoulo and Schaub, 2020; Mozes and Kleinberg, 2021).

7 Limitations

Given the relatively novel nature of this task, one of the major limitations of the work presented is only taking into account the benefits of examining entire conversations over individual messages from the second round of evaluation onwards. This negatively impacts the comparability of our results; we cannot compare the performance of the Spacy model with the RoBERTa model when considering entire conversations.

Additionally, we mention in section 3 that our system contains a rudimentary deidentification component that can substitute detected PIIs with either a sequence of random characters, a label, or a similar entity which was randomly sampled. For the purposes of our experiments, we have only considered the label-based deidentification (which is similar to redaction in the literature), as it made the anonymized text much more readable, and subsequently simplified the annotation task. We leave evaluating this component for future work.

8 Ethical Statement

The development of anonymisation or pseudonymisation systems is central to people's right to privacy. We view the work presented in this paper as a positive contribution, given that we provide the tools (pipeline, models, etc) to detect and deidentify sensitive data in Spanish and Catalan. Furthermore, we highlight the weaknesses we have observed both in our system and in early iterations of our improvement cycle with the aim of helping researches avoid similar pitfalls. However, while we do not foresee the methods described here to be used for unethical purposes, discussing any potential system weaknesses may facilitate system attacks down the line.

9 Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina project. We would like to extend our thanks to the anonymous reviewers for their kind and insightful comments that have helped improve this paper. We additionally want to thank Belén Alemán and 1millionbot for their invaluable help in analysing user generated data used in this study.

References

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.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Timothy Baldwin, Marie-Catherine De Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015a. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, page 126–135, Beijing, China. Association for Computational Linguistics.

Timothy Baldwin, Marie-Catherine De Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015b. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the workshop on noisy user-generated text*, pages 126–135.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Lydia Belkadi, Martine De Cock, Natasha Fernandes, Katherine Lee, Christina Lohr, Andreas Nautsch, Laurens Sion, Natalia Tomashenko, Marc Tommasi, Peggy Valcke, et al. 4.2 metrics for anonymization of unstructured datasets. *Privacy in Speech and Language Technology*, page 73.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Gil Francopoulo and Léon-Paul Schaub. 2020. Anonymization for the gdpr in the context of citizen and customer relationship management and nlp. In *workshop on Legal and Ethical Issues (Legal2020)*, pages 9–14. ELRA.

Fadi Hassan, Mohammed Jabreel, Najlaa Maaroof, David Sanchez, Josep Domingo-Ferrer, and Antonio Moreno. 2019. Recrf: Spanish medical document anonymization using automatically-crafted rules and crf. In *IberLEF@SEPLN*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Greety Jose and Nisha S Raj. 2014. Noisy sms text normalization model. In *International Conference for Convergence for Technology-2014*, pages 1–6. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Rodrigo Juez-Hernandez, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. Agora: An intelligent system for the anonymization, information extraction and automatic mapping of sensitive documents. *Applied Soft Computing*, 145:110540.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Mostafa Langarizadeh, Azam Orooji, Abbas Sheikhtaheri, and D Hayn. 2018. Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review. *eHealth*, 248:80–87.

Maximilian Mozes and Bennett Kleinberg. 2021. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization. *arXiv preprint arXiv:2103.09263*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Fernando Sánchez-León. 2019. Resource-based anonymization for spanish clinical cases. *IberLef@SEPLN*, pages 704–711.

Steven Kester Yuwono, Hwee Tou Ng, and Kee Yuan Ngiam. 2016. Automated anonymization as spelling variant detection. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 99–103, Osaka, Japan. The COLING 2016 Organizing Committee.