Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP

Gil Francopoulo, Léon-Paul Schaub

Proceedings of the workshop on Legal and Ethical Issues (Legal2020), pages 9–14
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP

Gil Francopoulo, Léon-Paul Schaub
Akio + Tagmatica, Akio + LIMSI-CNRS
43 rue de Dunkerque, 75010 Paris, France
gil.francopoulo@wanadoo.fr, lpschaub@akio.com

Abstract

The General Data Protection Regulation (GDPR) is the regulation in the European Economic Area (EEA) law on data protection and privacy for all citizens. There is a dilemma between sharing data and their subjects' confidentiality to respect GDPR in the commercial, legal and administrative sectors of activity. Moreover, the case of text data poses an additional difficulty: suppressing the personal information without deteriorating the semantic argumentation expressed in the text in order to apply a subsequent process like a thematic detection, an opinion mining or a chatbot. We listed five functional requirements for an anonymization process but we faced some difficulties to implement a solution that fully meets these requirements. Finally, and following an engineering approach, we propose a practical compromise which currently satisfies our users and could also be applied to other sectors like the medical or financial ones.

Keywords: anonymization, pseudonymization, GDPR, NLP

1. Introduction

The General Data Protection Regulation (GDPR)[1] is the regulation in the European Economic Area[2] (EEA) law on data protection and privacy for all citizens. The aim is to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying the regulation within the EEA. The main evolutions brought by the law are:

- New concepts are created or instated: profiling, right to be forgotten, privacy by design (Spiekermann, 2012),
- Transparency becomes an obligation (Goddard, 2017),
- Responsibilities are re-balanced (Lindqvist, 2017).

The regulation contains provisions and requirements of personal data of individuals and applies to any enterprise established in the EEA countries. This regulation changes the way we manage our data (Kamocki et al., 2018)(de Mazancourt et al., 2015). Business processes that handle personal data must be designed with consideration of the principles and provide safeguards to protect data, for example using anonymization, so that the data sets are not publicly available without explicit and informed consent. De-identification like data anonymization is the process of removing personally identifiable information from data sets, so that people whom the data describe remain anonymous (Ji et al., 2017).

Fully anonymized data that meet the legal bar set by European data protection law is no longer 'personal data' and is therefore not subject to the obligations of the GDPR at all. It should be added that a process akin to anonymization is pseudonymization in which personable identifiable information are replaced by one or more artificial surrogates. Pseudonymization[3] is defined in the GDPR Article 4(5) 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, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

A single pseudonym for each replaced field makes the data record unidentifiable while remaining suitable for data analysis. The main point to consider is the validity coverage of the pseudonym. There are two sorts of pseudonymization: one could be called as 'local' in which the pseudonym is only valid within a single text and the second could be called as 'global' in which the pseudonym is constant from a text to another within a corpus. The main difference between these three options is that anonymous and local pseudonymization data can't be re-identified while global pseudonymization still allows for some re-identification because various clues may be picked and linked together.

Two modes of treatment are concerned:

- During the development phase, in batch mode, large collections of texts need to be collected in order to feed various machine learning processes and statistical computations,
- During exploitation, in real-time mode, a constant flow of information needs to be inserted into real time data analysis or chatbots.

[1] Council Regulation 2016/679, 2016 O.J. (L 119) (EU) 1.

[2] Let's recall that the EEA is the European Union plus Iceland, Liechtenstein and Norway.

[3] Pseudonymization becomes now an active field of research to such an extent that a workshop has just been devoted to it (Ahrenberg and Megyesi, 2019)

The focus of this article can be summed up as: **how to adjust the cursor, in order to respect the personal privacy of the citizens while allowing in depth semantic data analysis at the level of a large group of people and texts?**

2. Industrial context

The context is the design and use of an anonymization tool within CRM which means usually Customer Relationship Management for private companies (Garcia-Crespo et al., 2010) but when applied to administration can be formulated as Citizen Relationship Management. The content is either email messages, social media flows or chatbot dialogues.

We operate in both the private and the public domains. In a private context, we work in the domain of e-commerce and retail where NLP techniques are used to compute customer satisfaction (or dissatisfaction) features under GDPR (Sun et al., 2017). In a public context, the communication department of the Prime Minister of an important EEA country receives hundreds of personal complaints and questions per day within a secure perimeter implemented with on-premise servers and firewall protection. The problem arises when these data should be given to another administrative department or to external sub-contractors for data analysis purposes in order to understand what are the concerns of the population directly from the verbatim corpus using NLP techniques. Up until now, this externalization was not possible under GDPR.

3. Related works

The problem of customer data anonymization is older than GDPR. (Zhong et al., 2005) show the efficiency of k-anonymization for customer privacy during automatic process. K-anonymization is defined as: 'Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful' (Samarati and Sweeney, 1998). Although (Nergiz and Clifton, 2007) outperformed k-anonymization with clustering-based algorithms. However these techniques were not effective to anonymize unstructured data as shown by (Angiuli and Waldo, 2016) and the GDPR introduced several changes in the definition of an anonymized text (Hintze and El Emam, 2018).

(Di Cerbo and Trabelsi, 2018) introduce an overview of supervised techniques for anonymization. In the medical domain, NLP tasks are grandly concerned by GDPR (Fraser et al., 2019; Kirinde Gamaarachchige and Inkpen, 2019; Berg and Dalianis, 2019; Dalianis, 2019). (Chevrier et al., 2019) propose a survey on specific techniques and issues of anonymization for medical datasets. (Goddard, 2017) propose a clustering approach for medical reports anonymization in order to limit the information loss and the data utility.

In the didactic field, (Megyesi et al., 2018) build a GDPR-compliant corpus for foreign language learner: their method can be partially reused in many domain because of the complete named-entities anonymization they realize.

Several open-source tools recently appear to anonymize texts according to GDPR (Adams et al., 2019; Kleinberg et al., 2017). Nevertheless, as far as we know, there is no formal approach of text anonymization for opinion mining-based tasks in the customer-relationship management field. Some works processing customer data just could not anonymize their corpora because of the task complexity : (Bechet et al., 2012) developped a corpus of call-center human-human spoken conversation from the Parisian public transport network (RATP) but were not able to distribute it because of the absence of anonymization. Moreover, the GDPR was not voted yet, we guess that anonymization of such a corpus would be even harder today.

4. Requirements

The main functional requirements are as follows:

- REQ#1 Avoid identifying the individuals mentioned in the text,
- REQ#2 Allow in-house semantic data analysis which could eventually be adapted to a certain kind of input,
- REQ#3 Allow off-the-shelf NLP tools,
- REQ#4 Prove that an anonymization has been done in case of a complaint from someone mentioned in a specific text or in case of lawsuit or journalists' investigation,
- REQ#5 Usable in different European languages.

These requirements are somehow contradictory. For instance, from the original text:

> My name is Paul Smith, and I moved from Leeds to Paris.

an anonymization will black out all identifiable information and will produce:

> My name is X and I moved from X to X.

In this case, REQ#1 and REQ#4 are fulfilled but the semantic processing of REQ#2 and REQ#3 will be deeply disrupted. Another option could be to replace a name with a random name from a dictionary while respecting the type of the name as:

> My name is John Wilson and I moved from Berlin to Madrid.

In this case, the realistic surrogates give the impression that the text is original but REQ#4 is not fulfilled. We cannot afford global pseudonymization because it is not really a secure anonymization (as mentioned in the introduction) but local pseudonymization seems a good compromise fulfilling four out of five requirements giving a sentence like:

> My name is _People1 and I moved from _City1 to _City2.

Req.	substitution by X	global pseudo.	local pseudo.	random substitution
REQ#1	yes	no	yes	yes
REQ#2	no	yes	yes	yes
REQ#3	no	no	no	yes
REQ#4	yes	yes	yes	no
REQ#5	yes	yes	yes	yes

Table 1: Requirements vs solutions

The only drawback of the approach is that the text cannot be given to an NLP process which is not prepared for this sort of mangling like an automatic translation, and therefore the REQ#3 target is missed. In fact, **REQ#3 and REQ#4 are contradictory**. Thinking again about this problem, we realized that certain requirements need not to be satisfied in all circumstances. REQ#4 is important when producing the data out of a secure perimeter while REQ#3 is important when using off-the-shelf tools internally within a secure perimeter. The dilemma can be resolved by implementing a Boolean parameter when running the anonymization associated to REQ#3 or REQ#4 fulfillment. Thus, the anonymization is able to produce:

> My name is _People1 and I moved from _City1 to _City2.

when there is a need to externalize, as well as:

> My name is John Wilson and I moved from Berlin to Madrid.

in case of internal processing, depending on the option. The requirement fulfillment is summed up in table-1.

5. Implementation

The idea is to chain three processes: 1) a named entity recognition, 2) an entity linker, and 3) a substitution. These processes should run within a secure environment and should not produce any traces of execution which could break the anonymization. That is to say that only the result of the substitution is authorized to be publish outside the running environment.

Named entity recognition (NER) is processed by Akio's named entity detector which takes the output of a syntactic parser whose name is Tagparser (Francopoulo, 2008). The parser combines statistical induction and robust syntactic rules. The NER is implemented by a cascade of patter matching rules to detect names of human beings, locations, companies, marks, email addresses and all sorts of numeral forms like dates, amounts of money, flight numbers, IBANs, phone numbers, passport numbers and social security numbers[4]. For proper names, the NER makes use of language-based local clues combined with a list of 1.2M proper names which have been automatically extracted from Wikidata. This is an industrial detector used to process currently an average of 1M texts every day in six languages (English, French, German, Italian, Spanish, Portuguese). There is a specific parser for each language whereas most named entity detections are language-neutral, that is there are the same in all our six covered languages. In fact, only a small set of cultural differences like vehicle identifications are different[5]. The program includes a specific spelling checker to process ill-formatted inputs based on a 10 years' experience of badly formatted input collection.

The aim of the entity linker is to gather named entities appearing in different places of the text possibly with some encyclopedic or orthographic variations. For instance, in 'Nicolas Sarkozy said... Sarko replied...' where 'Sarko' being a nickname for 'Nicolas Sarkozy', the two names should be linked. Another example is 'N Sarkozy' vs 'Nicolas Sarkozy' where 'N' is not ambiguous and should be considered as a given name. The objective is to link these utterances in a common structure.

The objective of the substitution is to replace a selection of entity types which are:

- **city** for the names of cities and agglomerations, like 'Paris' (a city) or 'Cergy-Pontoise' which is not formally a city but is an agglomeration.
- **contractNumber** for the combination of digit and letters which seems to be something else than a word or a number. This category includes some specific personal categories like IBANs (International Bank Account Number) and BICs (Bank Identifier Code).
- **emailAddress** for email addresses.
- **personName** for the names of individuals which are human beings.
- **identificationNumber** for the identifier of an individual like a social security number or a passport number.
- **IPAddress** for Internet Protocol addresses.
- **phoneNumber** for the various forms of a phone number.
- **vehicleIdentification** for the vehicle registration plates.
- **zipCode** for postal codes.

It should be noted that the NER detects other entity types like for instance, countries, regions, organizations, amounts of money or flight numbers. Obviously, it is technically easy to substitute these entities but the question is: what is the rationale to do so? These entities are less personal and without any personal clues there is no danger in keeping

[4]The reader can reproduce our work by using another NER provided that all the precise and personal forms like social security identifiers are correctly detected. Obviously, the quality of the whole process is highly dependent on that of the NER.

[5]The French system is not able to recognize German number plates, for instance, but the situations where it is necessary are extremely rare.

the original string, provided that the more the text is transformed, the more difficult the semantic parsing is. Due to the fact that city is replaced, the exact localisation cannot be determined, so there is no need to substitute the address in full, in addition to the fact that the recognition of the section indicating the street is very difficult because of the many possible forms.

6. Example

From this (invented) original text:

> Dear Sir/Madam,
> I am writing today to complain of the problem I have with www.ameli.fr. I'd like to create an account but my social security identifier 200 11 99 109794 on my carte vitale is not the same as the one of my mutual insurance 201 11 99 109794. How could I do?
> Best regards,
> Paul Watson,
> tel 01 23 34 34 56 pwatson@aol.fr

Note that the Carte Vitale is the health insurance card of the national health care system in France. The anonymization produces the following text, provided that the pseudonymization option is selected:

> Dear Sir/Madam,
> I am writing today to complain of the problem I have with www.ameli.fr. I'd like to create an account but my social security identifier _SSid1 on my carte vitale is not the same as the one of my mutual insurance _SSid2 How could I do?
> Best regards,
> _People1,
> tel _Phone1 _Email1

Due to the fact that the pseudonyms are renumbered starting at one in each text, it is not possible to induce any personal data from this text or to make any correlation with another text, so GDPR is respected. However, provided that the digital analytics program is specially adapted to orthographically handle pseudonyms and to interpret the pseudonym as a semantic named entity value, it is still possible to compute that the author has:

- A complaint concerning a given web site,
- A complaint of mismatch concerning different social security identifiers,
- A question.

This is fully satisfactory. It is typically the kind of results which are produced by our in-house product Akio Analytics but such a result could also be computed by another product implementing ABSA (Aspect Based Sentiment Analysis) as we do (Pontiki et al., 2014).

7. Method used for validation

The manual verification of a large corpus iteratively with alternations of correction / verification is a very heavy burden. Our test corpus is a collection of 18138 French verbatim from the legal and administrative sector of activity and we cannot verify the whole corpus after every improvement of the detector. We started by excluding randomly 300 verbatim as a test corpus, to be used afterwards.
The main focus is not to avoid noise but mainly to avoid silence, that is, we consider that it is not very important when a character string is over-substituted. On the contrary, missing a person name substitution is a serious mistake. We use the fact that the text is transformed after pseudonymization and if some proper names of the nine types are remaining, there is a good chance that there is an error. We tested the system for French following a three-fold approach iteratively on the development corpus containing 18138-300=17838 texts:

- Step#1, the corpus is anonymized with the local pseudonymization option,
- Step#2 the named entity is applied again and the result is filtered to retain the named entities of the nine types which do not begin with the character underscore, this character identifying a pseudonym. When there is a result, there is a good chance that this is an error.
- Step#3 the NER errors are fixed and the process is applied again at Step#1. We stopped when we have not found any error.

The different phases of the validation are presented in table-2.

rounds	nb of processed texts	nb of errors
phase-1	17838	284
phase-2	284	53
phase-3	53	0

Table 2: Results of validation

Evaluation of the test corpus is presented in table-3:

Nb of texts	Recall	Precision	FMesure
300	100	99.5	99.7

Table 3: Quality evaluation

The total distribution over the whole corpus (development and test) by type of entity is shown in Table 4.

8. Future work

The NER is currently used everyday in order to compute e-reputation and commercial data analysis in six languages for several big companies, but so far, we did not had time to work on anonymization in all these languages. In the near future, we plan to test the anonymization in languages other than French.
We also plan to extend the substitution to another entity which does not directly identify an individual but which by its context can do so, what is usually called a context-sensitive entity. We plan to substitute all organizational

type	nb of occ.	distrib.
city	6408	19%
contractNumber	17	0%
emailAddress	2141	6%
personName	20835	63%
identificationNumber	146	0%
IPAddress	89	0%
phoneNumber	1721	5%
vehicleIdentification	97	0%
zipCode	1687	5%
total	33141	100%

Table 4: Entity types distribution

names for (relatively rare) cases like: "as president of Danone".

9. Conclusion

After a presentation of the context of use which is rather broad, namely citizen and customer relationship management, we listed five precise requirements and discussed the various options to provide an effective implementation. Our requirements are not specific to our context and could be applied to another context like a medical or financial application.

Our process anonymizes critical information through a step-wise named entity recognition implementation and entity linking. It identifies contextual information and replaces them with a semantic-preserving category label which allow semantic data analytics except that the character string of certain proper names and numeric expressions are hidden but remain manageable. As an option, the program allows the replacement with a random value simulating an original character string for off-the-shelf NLP tools.

10. Acknowledgements

This work was co-financed by ANRT and Akio under the CIFRE contract 2017/1543.

11. Bibliographical References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., and Wechsler, R. (2019). AnonyMate: A toolkit for anonymizing unstructured chat data. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 1–7, Turku, Finland, 30 September. Linköping Electronic Press.

Lars Ahrenberg et al., editors. (2019). *Proceedings of the Workshop on NLP and Pseudonymisation*, Turku, Finland, 30 September. Linköping Electronic Press.

Angiuli, O. and Waldo, J. (2016). Statistical tradeoffs between generalization and suppression in the deidentification of large-scale data sets. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 589–593, June.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., ElbÃ¨ze, M., Mori, R. D., and Arbillot, E. (2012). Decoda: a call-center human-human spoken conversation corpus. In *In International Conference on Language Resources and Evaluation (LREC*.

Berg, H. and Dalianis, H. (2019). Augmenting a deidentification system for Swedish clinical text using open resources and deep learning. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 8–15, Turku, Finland, 30 September. Linköping Electronic Press.

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., and Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *J Med Internet Res*, 21(5):e13484, May.

Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pages 16–23, Turku, Finland, 30 September. Linköping Electronic Press.

de Mazancourt, H., Couillault, A., Adda, G., and Recourcé, G. (2015). Faire du TAL sur des données personnelles : un oxymore ? In *22eme Conférence sur le Traitement Automatique des Langues Naturelles*, Caen, France, June.

Di Cerbo, F. and Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In András Benczúr, et al., editors, *New Trends in Databases and Information Systems*, pages 118–126, Cham. Springer International Publishing.

Francopoulo, G. (2008). Tagparser: well on the way to ISO-TC37 conformance. In *ICGL (International Conference on Global Interoperabilty for Language Resources)*, Hong Kong, January.

Fraser, K. C., Linz, N., Lindsay, H., and König, A. (2019). The importance of sharing patient-generated clinical speech and language data. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 55–61, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Garcia-Crespo, A., Colomo-Palacios, R., Gomez-Berbis, J. M., and Ruiz-Mezcua, B. (2010). Semo: A framework for customer social networks analysis based on semantics. *Journal of Information Technology*, 25(2):178–188.

Goddard, M. (2017). The EU general data protection regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6):703–705.

Hintze, M. and El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. *Journal of Data Protection & Privacy*, 2(2):145–158.

Ji, S., Mittal, P., and Beyah, R. (2017). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys Tutorials*, 19(2):1305–1326, Secondquarter.

Kamocki, P., Mapelli, V., and Choukri, K. (2018). Data management plan (DMP) for language data under the new general da-ta protection regulation (GDPR). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kirinde Gamaarachchige, P. and Inkpen, D. (2019). Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64, Hong Kong, November. Association for Computational Linguistics.

Kleinberg, B., Mozes, M., Toolen, Y., and Verschuere, B. (2017). Netanos - named entity-based text anonymization for open science. June.

Lindqvist, J. (2017). New challenges to personal data processing agreements: is the GDPR fit to deal with contract, accountability and liability in a world of the Internet of Things? *International Journal of Law and Information Technology*, 26(1):45–63, December.

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C.-J., Sundberg, G., Wirén, M., and Volodina, E. (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, November. LiU Electronic Press.

Nergiz, M. E. and Clifton, C. (2007). Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3):622 – 645. 25th International Conference on Conceptual Modeling (ER 2006).

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report.

Spiekermann, S. (2012). The challenges of privacy by design. *Commun. ACM*, 55(7):38–40, July.

Sun, S., Luo, C., and Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10 – 25.

Zhong, S., Yang, Z., and Wright, R. N. (2005). Privacy-enhancing k-anonymization of customer data. In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 139–147, New York, NY, USA. ACM.