bitext we help AI understand humans

Anonymization for GDPR compliance

1 Introduction

In this era of data, when fields such as Big Data or Data Science are standing out, the treatment our personal information, that we continuously transfer to companies in exchange of using their services, is a concern that grows higher every day among citizens worldwide.

The General Data Protection Regulation (GDPR) is the result of Europe trying to harmonize all the preexisting legislations on the matter. It became effective in 2016, but failing to fulfill it is starting to be punished from May, 25th on.

***Ensuring compliance with this rule, therefore, is key for companies** if they want to avoid considerable fines. Fortunately, they can trust the latest natural language processing (NLP) technologies to encrypt these sensible data and solve the problem.*

In this paper we will explain how anonymization is done and how it can help companies achieve GDPR compliance in a seamless process.

2 Personal data

The GDPR defines personal data as "any information relating to an individual, whether it relates to his or her private, professional or public life. It can be anything from a name, a home address, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer's IP address."

Detecting and tagging these kinds of information is essentially a named entities recognition (NER) problem, a classic task in NLP.

Bitext offers the anonymization service as a customized feature of our Entity Extraction service. This feature has a particular mission: **to replace personal data with special tokens**, which allow to know where these sensible pieces of information were placed, but not their original value.

Some examples in Spanish, Russian and Japanese follow (with a translation below):

Language	Original text	Edited text
Spanish	La profesionalidad de los comerciales, especialmente de Alfonso.	*La profesionalidad de los comerciales, especialmente de PROPERNAME.*
	Professionalism shown by sales representatives, especially Alfonso's.	*Professionalism shown by sales representatives, especially PROPERNAME's.*
Russian	менеджер Нилов был очень приветлив.	менеджер *PROPERNAME* был очень приветлив.
	Manager Nilov was very friendly.	*Manager PROPERNAME was very friendly.*
Japanese	社員藤川さんの対応が良いから	社員*PROPERNAME*の対応が良いから
	Employee Fujikawa's good response.	*Employee PROPERNAME's good response.*

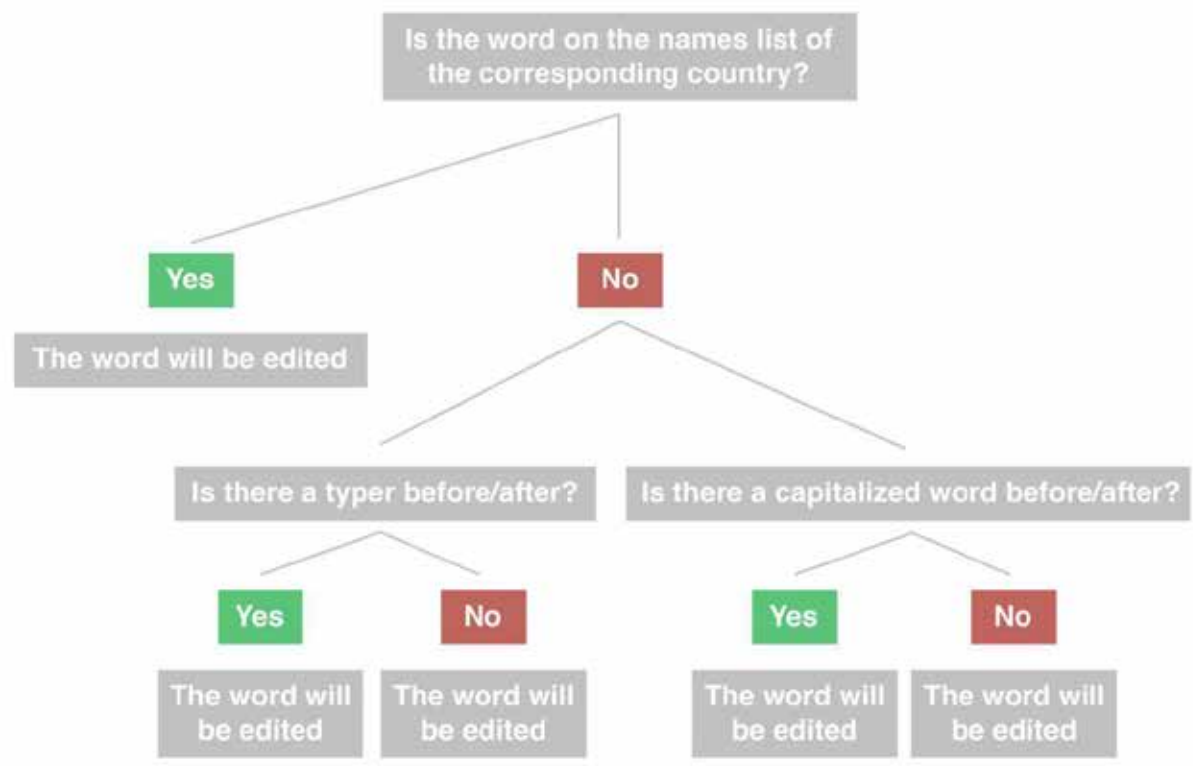
This way, the resulting text is still understandable by both humans or machines, but **whoever reads it won't see neither proper names nor surnames but that special token**, and the text can be processed safely.

3 How Bitext anonymization technology works

Our anonymization software is rule-based. This means its approach is not statistical, but rather uses linguistic data to make decisions. We will now describe the mechanics briefly in the case of proper names in English.

To find proper names, the software reads the given text and when it encounters a word that is listed in a set of proper names and surnames of the language in question, then it is edited, i.e., it is replaced by a special token, as explained above. We can guarantee the most comprehensive knowledge because we benefit from our **extensive experience in NLP in more than 20 languages**.

However, in case this isn't enough, there is also another way of detecting these kind of names: there are certain words that anticipate or follow proper names. We call them typers, and some examples are "sir", "madam", "Mr.", "Mrs.", occupations, etc. After discarding a word for edition because it is not a proper name, the software looks for it in the typers dictionary. If it is a typer and afterwards there's a word that begins with capital letter, then it gets edited.



Also, **for each language, we define all the possible element combinations that can integrate a complex proper name**. For example, in Japanese only a surname can appear before a typer, never a first name; in Italy, people usually write the surname before the first name too, etc.

Using regular expressions, we can also anonymize other kinds of personal data such as addresses, e-mail addresses, alphanumeric expressions such as phone numbers or identification numbers, bank accounts, etc.

A similar approach is used to encrypt offensive language as well. Taking advantage not only of our entity extraction technology, but also of our lemmatization software, **we can recognize any inflected form of the offensive word**. That is, either "fuck", "fucks", "fucked" or "fucking" will be detected as offensive, as they are forms of the same verb. This task can seem easy in English, but other languages such as Hungarian can have up to five thousand forms in its verbal conjugation, so lemmatizing is crucial.

4 Conclusion

Using Bitext anonymization service is the best way to ensure compliance with the GDPR in every language without any further data preprocessing. It is a seamless process that can be easily integrated with any system.