Paper TT07

Automated anonymization of protected personal data in clinical reports

Azad Dehghan, DeepCognito Ltd., Manchester, UK
Cathal Gallagher, d-Wise Technologies Inc., Raleigh, USA

ABSTRACT

The application of **N**atural **L**anguage **P**rocessing (NLP) methods to textual data such as **C**linical **S**tudy **D**ocuments (CSDs) has shown that automated identification and classification of **P**rotected **P**ersonal **D**ata (PPD) is feasible. Our investigation indicates that anonymization of CSDs can be automated to a large extent, and therefore significantly reduce resource costs whilst maintaining human benchmarks. Our approach includes a set of knowledge- and data-driven methods that complement each other to address the challenges and peculiarities of the problem space and inherit data characteristics. The baseline methods developed are designed to be part of a learning system that adapts and improves over time.

BACKGROUND

The **E**uropean **M**edicines **A**gency (EMA) Policy 0070 have set an expectation – that companies must now make detailed clinical trial findings available to the public. Transparency must now be the default rather than an afterthought. And there are plenty of good reasons for the industry to accommodate the new requirements. Although the EMA has been working towards transparency for some time, the reality of accessing requested documents has been far from ideal – often taking researchers several months to secure even basic information about previous trials. Policy 0070 has changed that with documents now being made available publicly. Progress has been slow so far but it has been steady. Redaction is no longer the default as anonymization is what the EMA is insisting upon.

BLUR

Blur is a scalable solution for de-identification and anonymization of clinical trial data and reports. This application-based de-identification solution enables emerging rules for de-identification and integration of the execution of these rules into a comprehensive workflow-driven process that provides automated documentation, traceability and audit trails. Blur is constantly evolving as de-identification rules evolve and converge, and provides a modern and superior alternative to the development and use of SAS-based de-identification scripts.

Blur started off as a data anonymization tool, where rules can be applied to variables, such as offsetting dates or elevating countries to continent. It then added a risk module to offer a mathematical calculation that illustrates the probability of a successful attack on that data based on a set of subject identifiers. Specifically, anonymization rules are applied to **I**ndividual **P**atient **D**ata (IPD); *risk* is quantified based on equivalence class sizes for each patient and compared to adequate threshold. EMA recommends using the maximum risk metric and a standard threshold of 0.09. Anonymization rules on selected Quasi-Identifiers are set to meet the threshold based on IPD and subsequently automatically propagated to the document for anonymization of corresponding PPD.

In addition to the extended document anonymization feature, image detection and manual redaction are also key features. These, combined with Blur NLP, create an automated approach to the long and tedious document anonymization task. At the same time creating a report of what has been anonymized or redacted, and where in the document that took place. Blur enables high-quality de-identified clinical trial data and reports to your internal and external collaborators at lower cost and 70% faster than any other approach available today.

PROTECTED PERSONAL DATA

Together with our partners in industry we have adopted/defined (not exhaustive) a set of identifiers or PPD described in Table 1.

Table 1 - Protected Personal Data.

IDENTIFIERS	TYPE
ANTHROPOMETRIC DATA	• Body mass index • Height • Weight
DATE	• Adverse event • Birth • Death • Test • Treatment • Visit
FREE TEXT	• Narrative text or investigator text
CONTACT INFORMATION	• Email • Fax • Phone • Uniform resource link or web address
IDENTIFICATION NUMBER	• Medical record • Study site • Subject • National
LOCATION DATA	• City • Country • Organization • Department • Postcode • State or metropolitan area • Street name
NAME	• Person • Username • Initials • Title
SENSITIVE DATA	• Drinking habit • Drug use • Findings/Interventions • History • Laboratory value • Smoking habit • Test • Sexual orientation • Religion
SOCIOECONOMIC DATA	• Profession • Qualification
DEMOGRAPHIC DATA	• Age • Ethnicity • Gender • Nationality • Race

Moreover, for each PPD an entity dimension was defined. This is necessary in order to guide the anonymization rules (or methods) propagated by *Blur Risk* (e.g., we may need to treat principal investigator information different from pharmaceutical staff members)*:*

- Aggregate or summary data
- Contract research organization or vendor
- Other persons (e.g., medical monitors, committee members, etc.)
- Other pharmaceutical staff members

- Principal or coordinating Investigator
- Sponsor non-signatory
- Sponsor signatory
- Study participant
- Study site staff
- Vendor (as principal investigator)

METHODS

Based on the defined PPD (e.g., Table 1) and entities, we modeled the Blur NLP engine using DeepCognito AI Text Analytics Platform (AI-TAP). The initial step in the NLP workflow is to convert PDF documents to text, and tables into a XML representation, both of which are prerequisites for input to the NLP engine. Further, a range of bespoke state-of-the-art methods were developed, including knowledge-driven (e.g., rules) and data-driven (**M**achine **L**earning, ML) to enable automation of identification and classification of PPD.

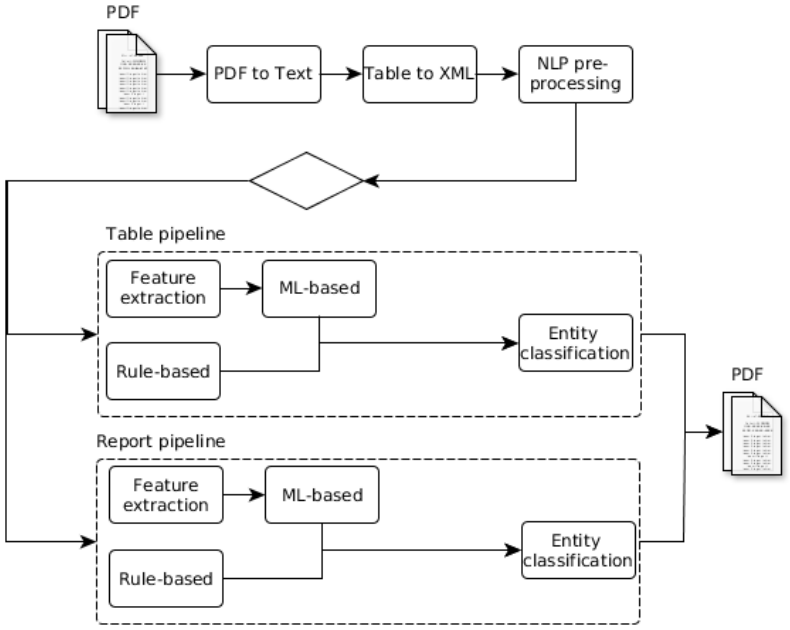

Figure 1 - NLP architecture

BLUR WORKFLOW

The following workflow shows the simple steps required to initiate the anonymization/redaction workflow using the Blur user interface:

1. Anonymize CSR datasets
2. Upload CSR documents
3. Repeat these steps as necessary
   a. Refine dataset anonymization to assist with CSR anonymization
   b. Run Find PPD
   c. Anonymize & Redact PPD as appropriate
4. Generate Internal Review Package
5. (External procedure) Internal review occurs
6. Import Study Documents requiring PPD edits (preserve CCI redactions if they exist)
7. Make PPD edits
8. Generate EMA Proposal Package
9. CCI Redactions added
10. (External procedure) EMA provides feedback (consultation output from EMA)
11. Import Study Documents requiring PPD edits (preserve CCI redactions if they exist)
12. Make PPD edits
13. Generate EMA Final Package
14. (External procedure) Add supporting materials and submit Final to EMA

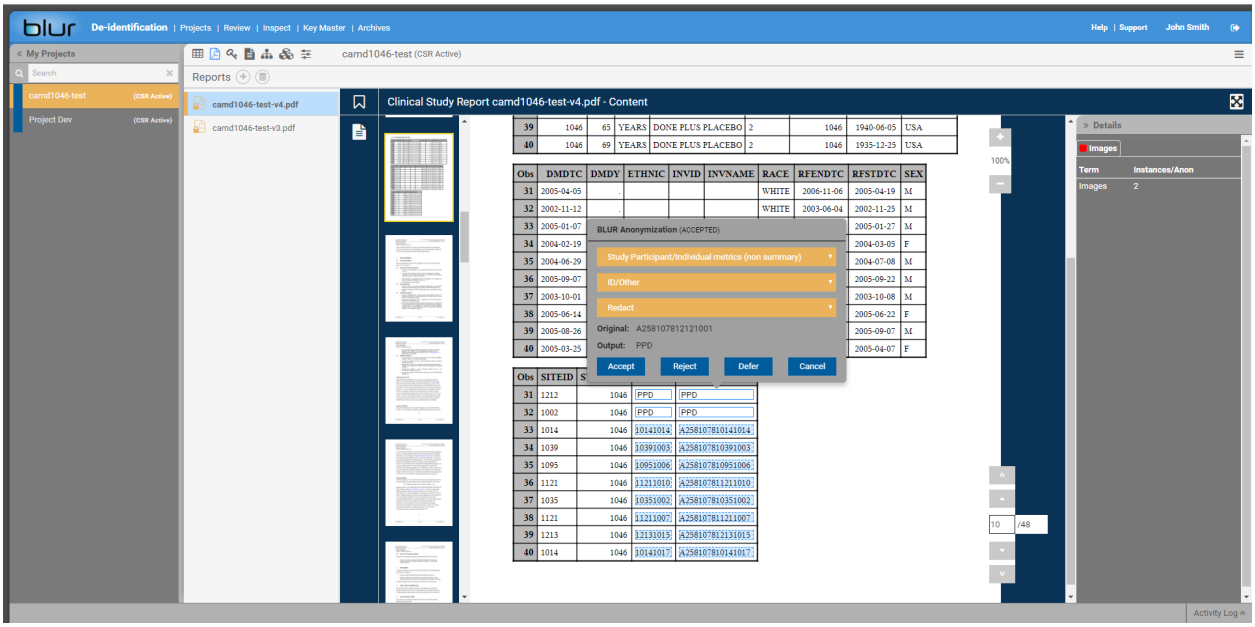

Figure 2 - Blur

RESULTS

We evaluated the NLP system using customary information extraction metrics such as (1), (2), and (3).

1) $Precision = \frac{True\ positives}{True\ positives + False\ positives}$

2) $Recall = \frac{True\ positives}{True\ positives + False\ negatives}$

3) $F_\beta\text{-}score = (1 + \beta^2)\frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$;

where β=1, as we accept equivalent importance between Precision and Recall.

Our evaluation of system performance was benchmarked against human annotators or de-identifiers. Specifically, we allowed human experts to manually analyze a set of duplicate documents and subsequently calculated a set of agreement scores (also known as inter-annotator or inter-rater agreement) across all (micro) PPD. Further, our experiments, using both k-fold cross validation and held-out datasets, showed that the automated system performs as well humans. In other words, we did not observe any significant difference between humans and Blur NLP. Both human and machine achieved 99% micro F_1-score.

DISCUSSIONS

The aforementioned metrics are widely used in information retrieval/extraction as opposed to arbitrarily defined *accuracy* which is misleading given commonly skewed label distribution with regard to *true negatives* in textual data. As an example, if we have a clinical study report which contains an average of 500 words per page across 100 pages; and let us say we have 500 subject identifiers distributed evenly across the report; this would effectively mean that without capturing a single subject identifier we could still report 99% accuracy. Therefore, the use of (1), (2), and (3) are preferred as they are not sensitive to skewed distributions, and given the aforementioned scenario, would correctly give us the score of 0%.

We are continuing to develop Blur NLP to address identifiers that were either sparse or not available in our training and evaluation datasets; in parallel we are aspiring towards a continuous learning system where the algorithms improve overtime by leveraging novel methods that extends latest research in artificial intelligence.

ACKNOWLEDGEMENT

We would like to acknowledge the support of the team with helping to draft this paper including Jean-Marc Ferran, Kris Spring, Alistair Dootson, and Clayton Leonard.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Azad Dehghan
DeepCognito Ltd.
Three Piccadilly Place
Manchester, M13BN
United Kingdom
Office Tel.: +44 (0) 161 241 2716
Email: azad.dehghan@deepcognito.com
Web: www.deepcognito.com