Anonymisation models for text data:
State-of-the-art, challenges and future directions

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet & Lilja Øvrelid

PrivateNLP workshop
NAACL 2021
Access to text documents with (sensitive) *personal data* crucial for many scientific fields

- Medicine, social sciences, legal studies, etc.
- Consent often difficult to obtain

Can we (semi-) automatically *mask* personal information from text data?
Plan

► What is anonymisation?
► Existing methods
► Limitations & case study
► Three challenges
► Sketch of future model
What is anonymisation?
(in the GDPR sense of the word)

= Complete & irreversible removal from the data of all information that may lead (directly or indirectly) to an individual being identified

But also quasi-identifiers that do not identify a person in isolation, but may do so when combined (with background knowledge): places, organisations, dates, demographic attributes, etc.

Must filter out all direct identifiers: names, bank accounts, mobile phones, etc.
What is anonymisation?
(in the GDPR sense of the word)

= Complete & irreversible removal from the data of all information that may lead (directly or indirectly) to an individual being identified

→ Removal of predefined categories of entities (like done in NER) is not enough!

→ Must consider how each textual element may influence the disclosure risk

(& the remaining data utility)
NLP methods

Based on sequence labelling:

- Handcrafted patterns or neural nets + domain adaptation

- Largest application domain: clinical data
  - Notably the 2014 i2b2/UTHealth shared task (diabetic patient records) & the 2016 CEGS –NGRID shared task (psychiatric intake)

Meystre et al. (2010)
Aberdeen et al., 2010
Yogarajan et al. (2018)
Dernoncourt et al. (2017)
Liu et al. (2017)
Hartman et al. (2020)

Stubbs and Uzuner (2015), Stubbs et al. (2017)
NLP methods

+ obfuscation methods to conceal particular personal attributes (gender, ethnicity, sexual orientation, etc.)

- Either from the text itself, or from latent representations derived from it
- Lexical substitution
- Adversarial learning
- Reinforcement learning
- Encryption

Reddy and Knight (2016)
Elazar and Goldberg (2018)
Friedrich et al (2019)
Xu et al. (2019)
Mosallanezhad et al. (2019)
Huang et al., 2020
Privacy-preserving data publishing (PPDP)

= Privacy-first approach that explicitly reasons over disclosure risk based on a privacy model (often $k$-anonymity and its variants)

- K-safety
  - Chakaravarthy et al. (2008)
- K-confusability
  - Cumby and Ghani (2011), Anandan et al. (2012)
- t-plausibility
  - Anandan et al. (2012)
- C-sanitize
  - Sánchez and Batet (2016, 2017)
Privacy-preserving data publishing (PPDP)

C-sanitize:

**Inputs:**
- Document $d$ (defined as a collection of terms)
- List of individuals/entities $C$ to protect in $d$
- Background knowledge $K$

**Output:**
Edited document $d'$ such that the remaining terms no longer identify any individual/entity in $C$

- Information-theoretic approach based on pointwise mutual information (PMI)
- PMI estimated from web occurrence counts
Case study

- **Task**: anonymise 8 Wikipedia biographies of famous scientists
  - 5 human annotators
  - 3 systems: NER, C-sanitize & Presidio

- Low agreement between the 5 annotators
  - Average of 0.68 on (binary) token decisions
  - *But remember*: anonymisation is a problem that allows for multiple solutions!
Case study

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NER</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOB-Exact</td>
<td>0.5</td>
<td>0.49</td>
<td>0.47</td>
</tr>
<tr>
<td>IOB-Partial</td>
<td>0.61</td>
<td>0.48</td>
<td>0.54</td>
</tr>
<tr>
<td>Binary</td>
<td>0.64</td>
<td>0.51</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Presidio</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOB-Exact</td>
<td>0.63</td>
<td>0.22</td>
<td>0.33</td>
</tr>
<tr>
<td>IOB-Partial</td>
<td>0.74</td>
<td>0.24</td>
<td>0.36</td>
</tr>
<tr>
<td>Binary</td>
<td>0.76</td>
<td>0.25</td>
<td>0.38</td>
</tr>
<tr>
<td><strong>C-sanitise</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOB-Exact</td>
<td>0.51</td>
<td>0.66</td>
<td>0.57</td>
</tr>
<tr>
<td>IOB-Partial</td>
<td>0.57</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>Binary</td>
<td>0.58</td>
<td>0.69</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 2: Micro-averaged scores for NER, C-sanitise and Presidio over all texts for annotators a1, a4, a5.

**Main takeaway:**
No method really solves the task appropriately
(see paper for details on error analysis)
Limitations

NLP methods:
- Does not remove enough (restricted to predefined categories)
- Removes too much (no account of disclosure risk)
- Focus on detection, not editing

PPDP methods:
- Documents reduced to “bags of terms”
- Restricted types of semantic inferences
- Scalability issues

Can we somehow «combine» those two families of approaches?
Challenge 1: inferences

- Must model how an attacker can infer the identity of a person by combining text elements with background knowledge
  - In C-sanitize: web co-occurrence counts
  - Good start, but far from sufficient

- Most harmful inferences in text documents are semantic (Montserrat Batet & David Sánchez, 2018)
  = they are based on the actual meaning expressed in the texts instead of their statistical distributions
Challenge 2: masking

► Most text anonymisation methods simply «black out» text spans
  ▪ Loss of data utility!

► Alternative: edit text spans instead of deleting them
  ▪ Ex: «surgeon» → «health professional»
  ▪ But how to we find the right generalisation?
  ▪ Good starting point: ontologies

(Anandan et al., 2012; Sánchez and Batet, 2016)
Challenge 3: evaluation

- Current systems often evaluated with IR-based metrics: precision, recall, $F_1$.
- But not all identifiers are equally important!
  - Idea: provide separate recall measures for e.g. direct & quasi-identifiers.
- Those metrics also exclusively focus on the detection, not the editing.
- Human evaluations also very useful.
  - (For instance: re-identification attacks)