Bennett Kleinberg and Maximilian Mozes
Tilburg University + University College London
bennett.kleinberg@tilburguniversity.edu
Text anonymization

Problem:

● Sensitive information often prohibit text datasets to be shared publicly (e.g., through GDPR)

→ Limits progress and collaboration

→ NLP research biased towards available data

Solution:

● Text anonymization
● Identify and redact potentially sensitive information (PSI) in text
Who is this?

[PRONOUN] is an [LOCATION_1] film actor known for playing [OTHER_1] in the [OTHER_2] series of films. Since [DATE_1], [PRONOUN] has been playing the character but [PRONOUN] confirmed that [OTHER_3] would be [PRONOUN] last [OTHER_1] film. [PRONOUN] was born in [LOCATION_2] on [DATE_2] of [DATE_3] in [DATE_4]. [PRONOUN] moved to [LOCATION_3] when [PRONOUN] parents divorced and lived there until [PRONOUN] was [NUMERIC] years old. [PRONOUN] auditioned and was accepted into the [ORGANIZATION_1] and moved down to [LOCATION_4].
He is an English film actor known for playing James Bond in the 007 series of films. Since 2005, he has been playing the character but he confirmed that No Time to Die would be his last James Bond film. He was born in Chester on 2nd of March in 1968. He moved to Liverpool when his parents divorced and lived there until he was sixteen years old. He auditioned and was accepted into the National Youth Theatre and moved down to London.
(Automated) text anonymization

- Existing efforts to text anonymization often involve manual work to redact potentially sensitive information (PSI)
  → Slow and cost-intensive process
- Others resort to hand-crafted rules, without preserving text’s semantics (e.g., UK Data Service)

Automated, semantics-preserving text anonymization:

- Use NLP methods to automatically identify and replace PSI in text
- Replace PSI in a meaningful way to preserve semantics
Our approach

- Automated text anonymisation 
  - *old school* vs learning-based

Pillars:

- Fast
- Scalable
- Offline
- Lightweight

- Open science-focused
- Research end user in mind
Anonymization pipeline

Victoria Beckham is married to David Beckham.

a) Identifying sensitive information

b) Replacing identifiable tokens

[FIRSTNAME_1] [LASTNAME_1] is married to [FIRSTNAME_2] [LASTNAME_1].
Textwash (= current version)

Model:

- Machine learning-based text anonymization
- Model is based on BERT (Devlin et al., 2018)
  
  \[ \text{Fine-tuned with a token classification objective} \]

Data:

- Textwash is built on 3.7k human-annotated documents (British National Corpus, Enron emails, Wikipedia articles)
## Textwash: categories

Textwash supports 11 categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON_FIRSTNAME</td>
<td>Jane</td>
</tr>
<tr>
<td>PERSON_LASTNAME</td>
<td>Smith</td>
</tr>
<tr>
<td>OCCUPATION</td>
<td>Doctor</td>
</tr>
<tr>
<td>LOCATION</td>
<td>Netherlands, behind the curtain</td>
</tr>
<tr>
<td>TIME</td>
<td>12.59pm, afternoon, midday</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>Microsoft, NWO</td>
</tr>
<tr>
<td>DATE</td>
<td>01.01.1970, 3rd of November</td>
</tr>
<tr>
<td>ADDRESS</td>
<td>42 London Road</td>
</tr>
<tr>
<td>PHONE_NUMBER</td>
<td>+44XXXXXXXXXXX</td>
</tr>
<tr>
<td>EMAIL_ADDRESS</td>
<td><a href="mailto:jane@smith.com">jane@smith.com</a></td>
</tr>
<tr>
<td>OTHER</td>
<td>?</td>
</tr>
</tbody>
</table>
Margaret Thatcher, nicknamed the “Iron Lady”, served as Prime Minister of the United Kingdom from 1979 to 1990.

FIRSTNAME_1 LASTNAME_1, nicknamed the “Iron Lady”, served as OCCUPATION_1 of the LOCATION_1 from DATE_1 to DATE_2.

This anonymization is almost perfect (6 out of 7 PSI were anonymized), yet a single entity gives away identity of individual.
Evaluation

Using the **TILD criteria** (Mozes and Kleinberg, 2021)

1. How many PSI does it correctly identify?
   → **Technical evaluation**

2. How does anonymization affect downstream tasks?
   → **Information loss evaluation**

3. Can individuals be identified from anonymized texts?
   → **De-anonymization (motivated intruder testing)**

---

Technical evaluation

Textwash performance scores

Weighted average F1: 0.93
Information loss

- RoBERTa (Liu et al., 2019) fine-tuned on IMDb (Maas et al., 2011)
  → Original dataset
  → Anonymized dataset

- Performance differences are small
  → Preserves utility
Motivated intruder testing

Human participants are asked to identify individuals on three levels

- Famous individuals (e.g., Emma Watson, Daniel Craig)
- Semi-famous individuals (e.g., Kenny Kramm)
- Fictitious individuals

Collecting descriptions

- Each category consisted of 10 items
- \( n=401 \) participants wrote 3 descriptions each (total of 1202 descriptions)

Intruder testing

- \( n=366 \) participants, each judged 10 items in a single group
De-anonymization rates

<table>
<thead>
<tr>
<th></th>
<th>Famous</th>
<th>Semi-famous</th>
<th>Fictitious</th>
</tr>
</thead>
<tbody>
<tr>
<td>% identified</td>
<td>18.25</td>
<td>1.01</td>
<td>2.01</td>
</tr>
</tbody>
</table>

- Rates are highest for famous individuals (expected)

Replication study

- Repeated intruder testing for 20 famous individuals
- Collection: $n=200$ participants wrote five descriptions each
- Testing: $n=222$ individuals, 10 texts each
- Results: de-anonymization rate of 26.39% (very famous celebrities)
Using Textwash

- Currently available on GitHub
- Supports txt files, runs smoothly on CPU

```
$ python3 anon.py --input_dir examples --output_dir anonymized_examples
```

Docs and guidelines on GitHub.
Textwash becomes FAMTAFOSS

1. Extension to the Dutch language
   - High quality annotations of Wikipedia corpora (Dutch and English)
   - Completed with 2.5k documents $\rightarrow$ 783k annotated entities (21% in PSI categories)
   - Addition of new categories:
     - TITLE (of a song, prize, book, etc.)
     - CULTURALIdentity (e.g., religion, sexual orientation, ethnicity)

Next phase: scaled-up crowdsourced annotation

- 8k documents with $\sim$ 2.5m annotated entities ($\sim$ 500k PSI entities)
- Training the Dutch model
- Updating the English model
Textwash becomes FAMTAFAOS

1. Extension to the Dutch language
2. Graphical user interface (GUI) for non-programmers
FAMTAFO User Interface

This is a simple UI demo for FAMTAFO.

Your input documents

Choose files  No file chosen

Please drag and drop (or select) a folder of documents (or a zip file containing documents) that should be anonymized.

Select the entity types that should be anonymized

- Select all
- LOCATION
- PHONE_NUMBER
- EMAIL_ADDRESS
- OTHER

Only the selected entity types will be anonymized by FAMTAFO.

Please enter any terms (comma-separated) that should under no circumstances be anonymized

Tilburg University, January

FAMTAFO will ensure that these terms will not be changed in your submitted text documents.

Submit

When clicking Submit, FAMTAFO will anonymize your documents and the result will automatically be downloaded.
Textwash becomes FAMTAFOŠ

1. Extension to the Dutch language
2. Graphical user interface (GUI) for non-programmers
3. Custom white-listing
4. Custom black-listing
5. Risk score model
Textwash becomes FAMTAFOSS

1. Extension to the Dutch language
2. Graphical user interface (GUI) for non-programmers
3. Custom white-listing
4. Custom black-listing
5. Risk score model
Where are we?

- **Netanos** software → 2016-2018
- Textwash software → 2019-2020
- Textwash validation studies → 2021
- **FAMTAFOS** development
  - Dutch model (phase 1) → complete
  - Large-scale annotation → 12/2022
  - GUI → prototype
  - Validation studies → 1/2023
  - Feature wishlist → 2/2023

Famtafos will be available from March 2023 onward.
Thank you

GitHub: https://github.com/maximilianmozes/textwash
References

- Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C., 2011, June. Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).