

FAMTAFOS

**Free automated multi-language text
anonymization for open science**

Bennett Kleinberg and Maximilian Mozes

Tilburg University + University College London

`bennett.kleinberg@tilburguniversity.edu`

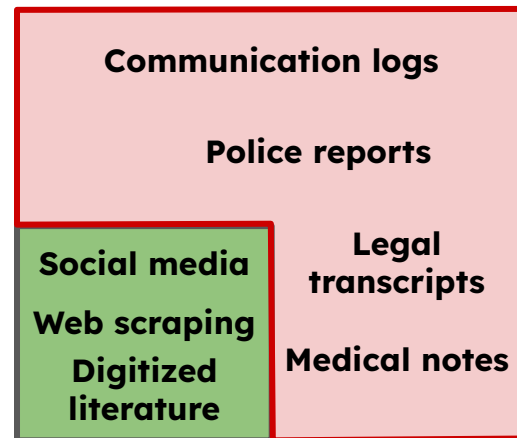
Text anonymization

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].

Bond. James Bond.

***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

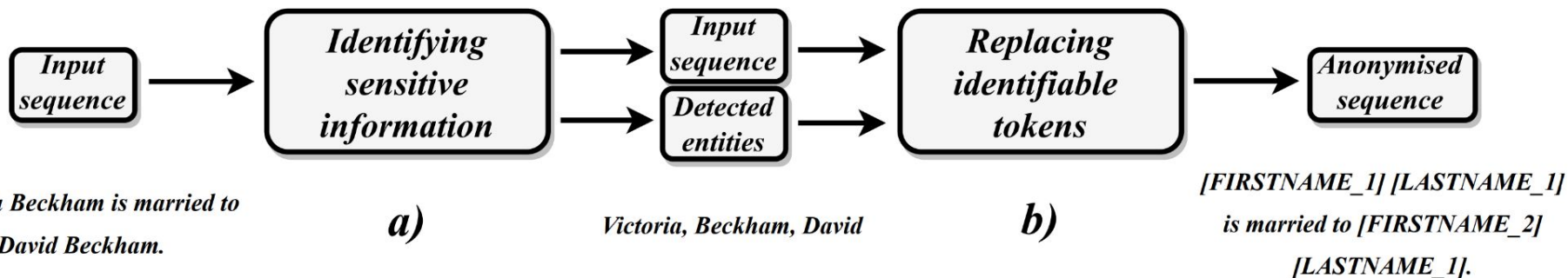
Our approach

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

Pillars:

- **Fast**
 - **Scalable**
 - **Offline**
 - **Lightweight**
-
- **Open** science-focused
 - **Research** end user in mind

Anonymization pipeline



Textwash (= current version)

Model:

- Machine learning-based text anonymization
- Model is based on BERT (Devlin et al., 2018)

→ *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

PERSON_FIRSTNAME

Jane

PERSON_LASTNAME

Smith

OCCUPATION

Doctor

LOCATION

Netherlands, behind the curtain

TIME

12.59pm, afternoon, midday

ORGANIZATION

Microsoft, NWO

DATE

01.01.1970, 3rd of November

ADDRESS

42 London Road

PHONE_NUMBER

+44XXXXXXXXXX

EMAIL_ADDRESS

jane@smith.com

OTHER

?

The importance of evaluation

Example from Mozes and Kleinberg (2021)

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.*

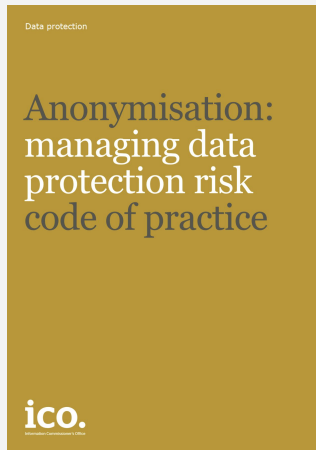
This anonymization is *almost* perfect (6 out of 7 PSI were anonymized), yet a single entity gives away identity of individual.

Evaluation

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

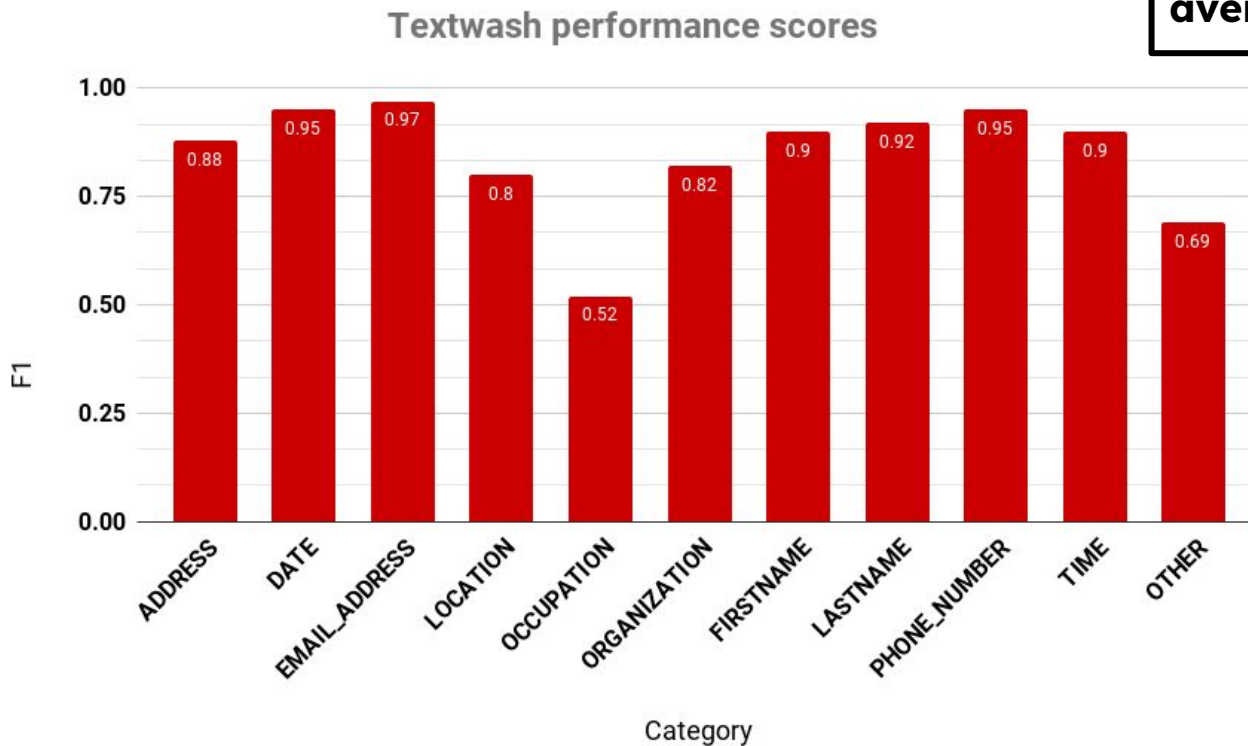
1. How many PSI does it correctly identify?
→ **Technical evaluation**
2. How does anonymization affect downstream tasks?
→ **Information loss evaluation**
3. Can individuals be identified from anonymized texts?
→ **De-anonymization (motivated intruder testing)**

Motivated intruder test



- Proposed by ICO
- Human re-identification
- Can use any resources
- No prior knowledge
- No specialist skills

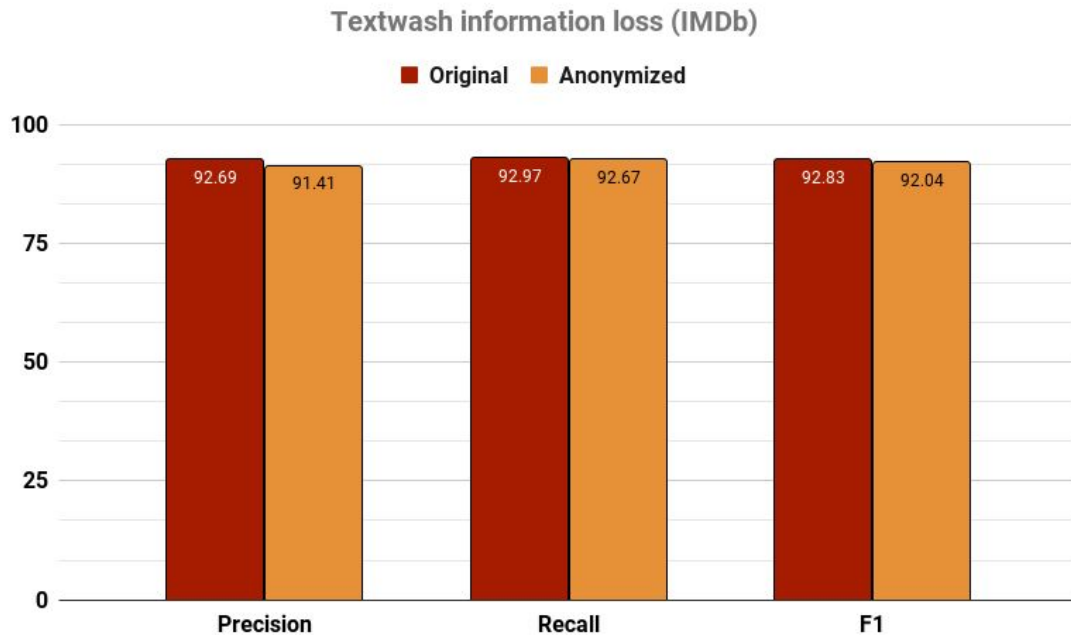
Technical evaluation



**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

	Famous	Semi-famous	Fictitious
% identified	18.25	1.01	2.01

- 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 FAMTAFOS

1. Extension to the Dutch language

- High quality annotations of Wikipedia corpora (Dutch and English)
- Completed with 2.5k documents → 783k annotated entities (21% in PSI categories)
- Addition of new categories:
 - TITLE (of a song, prize, book, etc.)
 - CULTURAL IDENTITY (e.g., religion, sexual orientation, ethnicity)

Next phase: scaled-up crowdsourced annotation

- 8k documents with ~ 2.5m annotated entities (~ 500k PSI entities)
- Training the Dutch model
- Updating the English model

Textwash becomes FAMTAFOS

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

FAMTAFOS User Interface

This is a simple UI demo for FAMTAFOS.

Your input documents

No file chosen

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

Folder with raw files

Select the entity types that should be anonymized

- | | | | |
|--|--|---|--|
| <input checked="" type="checkbox"/> Select all | <input checked="" type="checkbox"/> PERSON_FIRSTNAME | <input checked="" type="checkbox"/> PERSON_LASTNAME | <input checked="" type="checkbox"/> OCCUPATION |
| <input checked="" type="checkbox"/> LOCATION | <input checked="" type="checkbox"/> TIME | <input checked="" type="checkbox"/> ORGANIZATION | <input checked="" type="checkbox"/> DATE |
| <input checked="" type="checkbox"/> ADDRESS | <input checked="" type="checkbox"/> PHONE_NUMBER | <input checked="" type="checkbox"/> EMAIL_ADDRESS | <input checked="" type="checkbox"/> OTHER |

Custom entity selection

Only the selected entity types will be anonymized by FAMTAFOS.

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

Tilburg University, January

White-listing terms

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

Submit

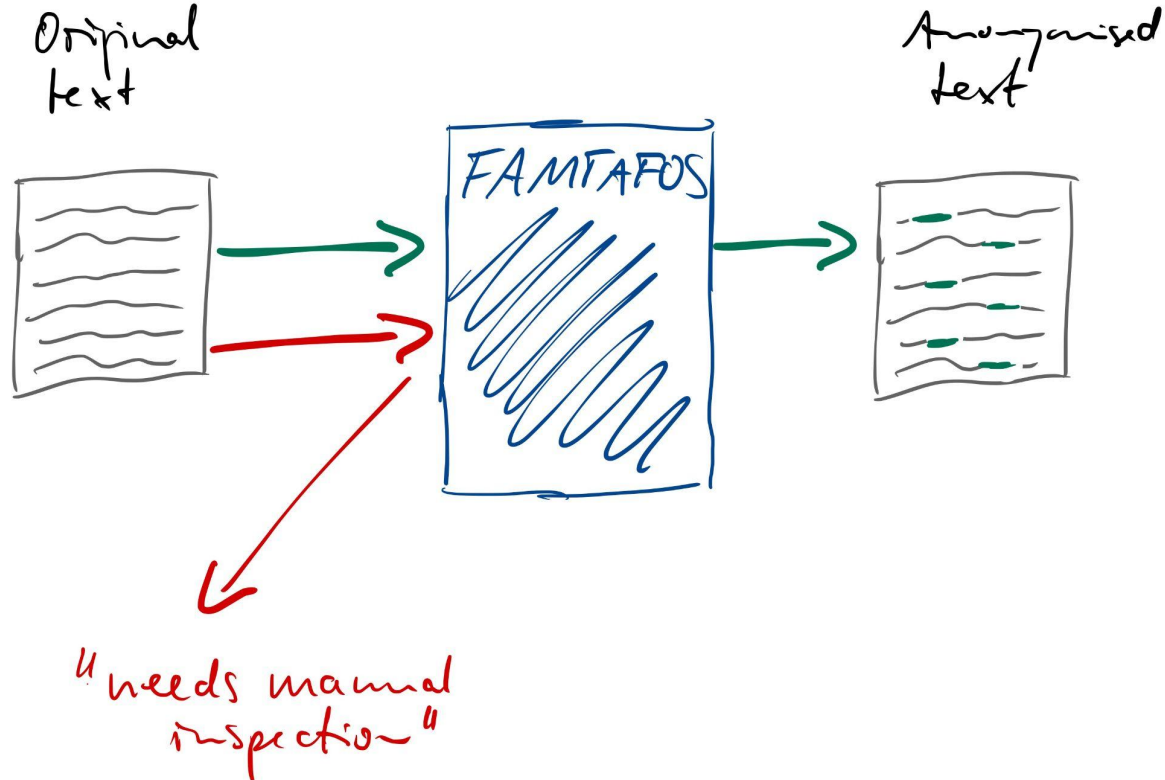
When clicking Submit, FAMTAFOS will anonymize your documents and the result will automatically be downloaded.

Textwash becomes FAMTAFOS

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 FAMTAFOS

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

Paper: <https://arxiv.org/abs/2208.13081>

GitHub: <https://github.com/maximilianmozes/textwash>

References

- Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
- 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).
- Mozes, M. and Kleinberg, B., 2021. No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization. arXiv preprint arXiv:2103.09263.