

I want to be FAIR - How to deal with my qualitative data?

# Automated text data anonymisation Textwash and FAMTAFOS

**Bennett Kleinberg** and Maximilian Mozes

Tilburg University and University College London

[bennett.kleinberg@tilburguniversity.edu](mailto:bennett.kleinberg@tilburguniversity.edu)

*Anonymised text:* [firstname1] [lastname1] is the wife of [occupation1] [lastname2]. [pronoun] is a mother of [numeric] children; [numeric] boys and a girl. [firstname1] is educated to university level and that is where [pronoun] met [pronoun] future husband. [firstname1] dresses elegantly and is often seen carrying out charity work. However, [pronoun] is a mum first and foremost and the interactions we see with [pronoun] children are adorable. [firstname1]'s sister, [firstname2], has followed [firstname1] into the public eye. [pronoun] was born in [date1] and will soon turn [numeric]. When pregnant, [firstname1] suffers from a debilitating illness called [otherattribute1], which was little known about until it was reported that [firstname1] had it.

# Text anonymization

## The dilemma:

- Text data are very promising
- But: sensitive information often prohibit the sharing of data (e.g., through GDPR)

## Solution

Text anonymization

# Our approach

## **Automated, semantics-preserving text anonymization:**

Pillars:

- **Fast**
- **Scalable**
- **Offline**
- **Lightweight**
  
- **Open science-focused**
- **Research** end user in mind
- machine learning-based (fine-tuned token classification)

# Textwash: categories

- Current pieces of ***potentially sensitive information*** (PSI):
  - PERSON\_FIRSTNAME: a person's firstname (e.g., Jane)
  - PERSON\_LASTNAME: a person's lastname (e.g., Doe)
  - OCCUPATION: an occupation (e.g., nurse, carpenter)
  - LOCATION: a location (e.g., London, Berlin)
  - TIME: a time (e.g., 12pm, afternoon)
  - ORGANIZATION: an organisation (e.g., Google, NHS)
  - DATE: a date (e.g., 12/10/2021, yesterday)
  - ADDRESS: an address (e.g., 42 London Road)
  - PHONE\_NUMBER: a phone number
  - EMAIL\_ADDRESS: an email address
  - OTHER\_IDENTIFYING\_ATTRIBUTE: an identifying attribute that cannot be categorised into the above
  - NUMERIC: numeric values (e.g., 13, 41)
  - PRONOUN: pronouns (e.g., they, she, he)
  - TITLE: titles (e.g., Prof.)
  - MR/MS: other titles (e.g., Mr., Mrs., Miss)
  - NONE: all other tokens in the input sequence

# The importance of 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)**

# 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)***

Weighted F1 score: **0.93**

Macro average: 0.86

# 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)**

Utility loss in sentiment classification  
(50k movie reviews):

Accuracy raw (original data): 92.80%

Accuracy anonymised data: 92.00%

Utility loss is **less than 1.00%**

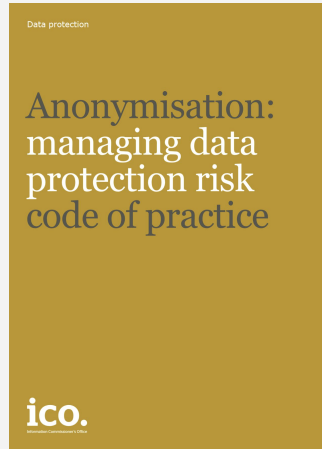


# Evaluation

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



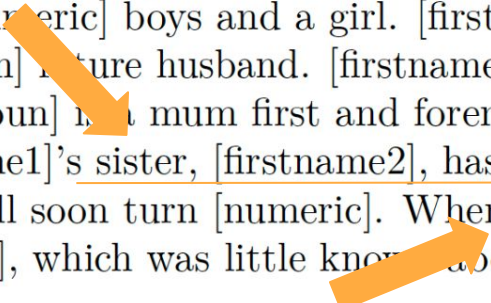
- Human re-identification
- Can use any resources
- No privileged knowledge
- No specialist skills

# The motivated intruder test

*Anonymised text:* [firstname1] [lastname1] is the wife of [occupation1] [lastname2]. [pronoun] is a mother of [numeric] children; [numeric] boys and a girl. [firstname1] is educated to university level and that is where [pronoun] met [pronoun] future husband. [firstname1] dresses elegantly and is often seen carrying out charity work. However, [pronoun] is a mum first and foremost and the interactions we see with [pronoun] children are adorable. [firstname1]'s sister, [firstname2], has followed [firstname1] into the public eye. [pronoun] was born in [date1] and will soon turn [numeric]. When pregnant, [firstname1] suffers from a debilitating illness called [otherattribute1], which was little known about until it was reported that [firstname1] had it.

# An educated guess?

*Anonymised text:* [firstname1] [lastname1] is the wife of [occupation1] [lastname2]. [pronoun] is a mother of [numeric] children; [numeric] boys and a girl. [firstname1] is educated to university level and that is where [pronoun] met [pronoun] future husband. [firstname1] dresses elegantly and is often seen carrying out charity work. However, [pronoun] is a mum first and foremost and the interactions we see with [pronoun] children are adorable. [firstname1]'s sister, [firstname2], has followed [firstname1] into the public eye. [pronoun] was born in [date1] and will soon turn [numeric]. When pregnant, [firstname1] suffers from a debilitating illness called [otherattribute1], which was little known about until it was reported that [firstname1] had it.



*Information mentioned by intruder:* “Suffered from an illness in pregnancy and has a famous sister.”

# An educated guess?

*Original text:* Kate Middleton is the wife of Prince William. She is a mother of 3 children; 2 boys and a girl. Kate is educated to university level and that is where she met her future husband. Kate dresses elegantly and is often seen carrying out charity work. However, she is a mum first and foremost and the interactions we see with her children are adorable. Kate's sister, Pippa, has followed Kate into the public eye. She was born in 1982 and will soon turn 40. When pregnant, Kate suffers from a debilitating illness called Hyperemesis Gravidarum, which was little known about until it was reported that Kate had it.



# Motivated intruder testing

## Findings

Table 3: Cosine similarities between the true person name and the participant choice (M, SD) and (un)successful de-anonymisations per type.

Item type	M	SD	% identified	SE % identified
famous	0.41	0.36	18.25	1.93
fict	0.04	0.13	1.01	0.50
semifamous	0.13	0.20	2.01	0.70

Honest performance:

**1-2% re-identification rate**

(findings replicated in a new dataset)

# 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

## Free Automated Multi-language Text Anonymization For Open Science GUI

### Improved base model

- Annotation of an additional 1.2k documents (Wikipedia biographies)
- Updating the English model

### Extension to the Dutch language

- Annotations of Dutch Wikipedia articles + 2.5k newspaper articles
- Ca. 1,000k annotated entities (21% in PSI categories)
- New categories:
  - TITLE (of a song, prize, book, etc.)
  - CULTURAL IDENTITY (e.g., religion, sexual orientation, ethnicity)

# FAMTAFOS User Interface

*This is a simple UI demo for FAMTAFOS.*

## 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.*

**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

**Allowlisting 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.*



# Where are we?

- v0.01: Netanos (2017-2019)
- v0.02: Textwash precursors (2019)
- v1.00: Textwash core (2020-)
- v1.10: FAMTAFOS (2022-)

## What's next?

- Improved base model (based on new set of 1,000k annotations)
- 18 categories
- Multi-language models: Dutch, German
- GUI
- Feature updates

Release: end of October 2023

# Anticipated questions/comments

1. Do large language models not make this all redundant?
2. But there is also tool XYZ that can do this (not question, more of a comment)
3. If you're not using word lists, how does the model know what a first name/an organisation/etc. is?

# Thank you

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

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

*Questions, interested in using it or a demo? Reach out to us:*

*Bennett Kleinberg ([bennett.kleinberg@tilburguniversity.edu](mailto:bennett.kleinberg@tilburguniversity.edu)).*

# 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.

# Textwash (= current version)

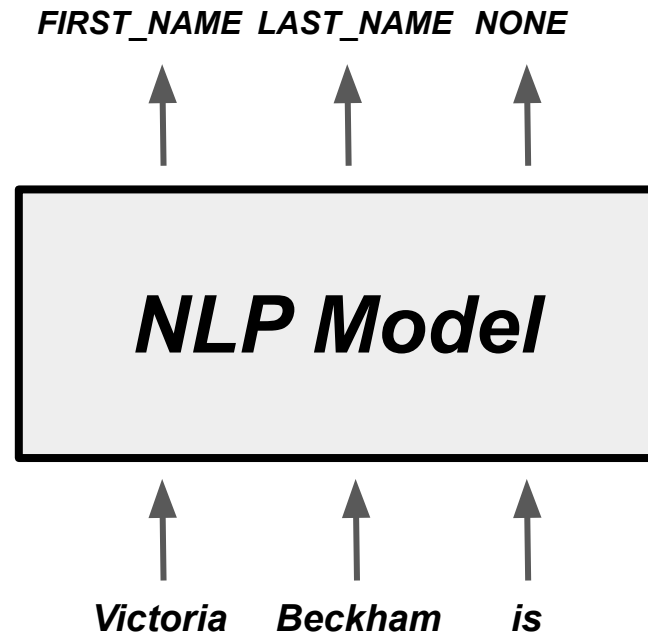
## Model:

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

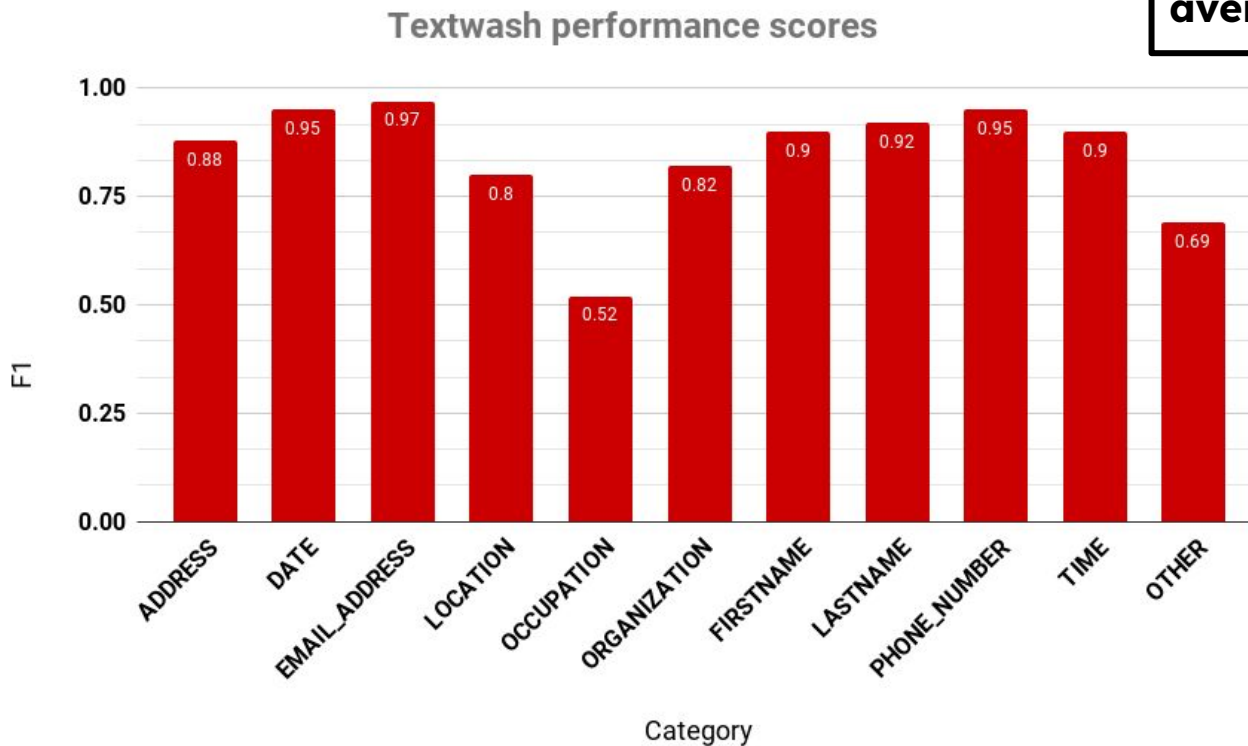
→ *Fine-tuned with a token classification objective*

## Data:

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



# 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

