Ethics in Natural Language Processing – SS 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

Lecture 8
Privacy & Security I

Dr. Thomas Arnold
Aniket Pramanik


Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

Nr.	Lecture
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

Outline

Recap

What is Privacy?

Misuse of Privacy Information

Demographic Profiling

Authorship Obfuscation

Low-resource NLP

- Low-Resource NLP is hard!
- Ambiguity (word senses, part-of-speech, syntactic structure…)
- Linguistic diversity at all levels of language structure
  - Tokenization, morphology, part-of-speech, syntax, semantics, discourse…
- Paradigm shifts in NLP
  - Rule-based NLP: high precision, low recall
  - Statistical NLP: Needs more data
  - Neural NLP: Needs MORE more data!
- Most promising approach: Transfer Learning

Learning Goals

After hearing this lecture, you should be able to…

- Explain how aggregation of data can violate OR facilitate privacy
- Describe the privacy paradox
- Give examples how privacy information can be used positively or negatively. (used / misused)
- Give ideas about authorship obfuscation

Outline

Recap

What is Privacy?

Misuse of Privacy Information

Demographic Profiling

Authorship Obfuscation

1993

"On the Internet, nobody knows you're a dog."

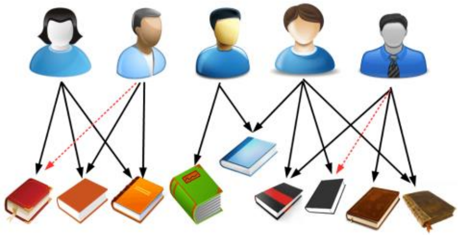
2018

TECHNISCHE UNIVERSITÄT DARMSTADT

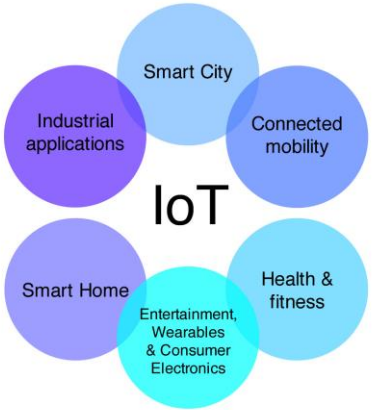
Cambridge Analytica

When is profiling used in our everyday life?

Profiling




TSA Pre✓

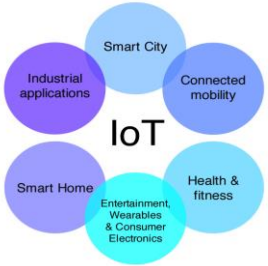

Smart City
Industrial applications
Connected mobility
IoT
Smart Home
Health & fitness
Entertainment, Wearables & Consumer Electronics


amazon


VISA
MasterCard
AMERICAN EXPRESS

Profiling

TSA Pre



The more data these systems have about people the better is their accuracy.

What is Privacy

What is Privacy

https://en.wikipedia.org/wiki/Privacy is the ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively

Privacy in Everyday Life

- Private vs public space

- Being seen vs being identified

- Being observed vs being identified, followed and watched
- Personal space
- Informational space

- Being tracked across locations and activities

Aggregation of Private Information

- Aggregation per user - violates privacy
  - Profiling, tracking, individual behaviour…
- Aggregation across users - facilitates privacy (ideally)
  - Statistics, trends, general behaviour…

Replace That Camera by a Person

- **Territorial privacy**: Public vs private space
- **Personal privacy**: Being seen vs being watched
- **Informational privacy**: Being seen vs being watched vs being tracked

→ Privacy issues arise due to personalization and aggregation of data

Do People REALLY Care About Protecting Their Privacy?

- **Information privacy paradox**: privacy attitudes vs privacy behaviors (Kokolakis ‘17)

- Surveys of internet users’ attitudes: Highly concerned about their privacy
- But easily trade their personal data
  - Revealing personal details to a shopping bot (Spiekermann et al. ‘01)
  - Trading online history for ~7 Euros (Carrascal et al. ‘13)

Do People REALLY Care About Protecting Their Privacy?

- When Parents Compromise Children's Online Privacy (Minkus et al. '15)

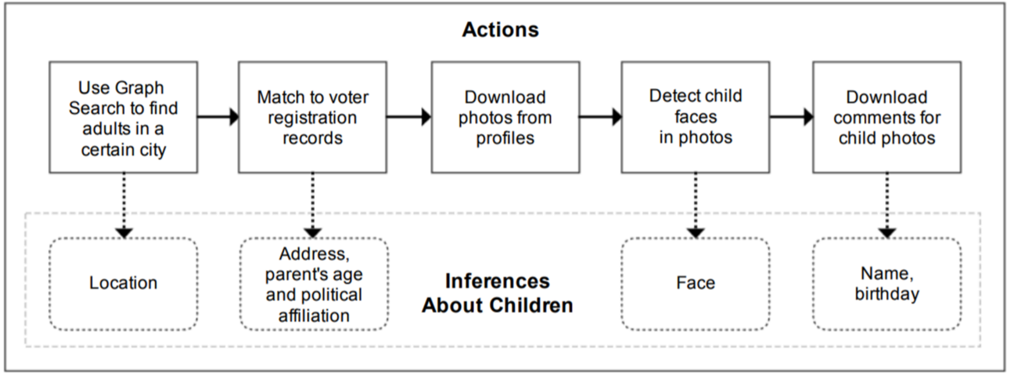


Figure 1: The process for downloading and inferring traits about children whose photos are posted on Facebook.

- How concerned are you about your children's online privacy? → 3.8 on a Likert scale [1..5]
- 35% of 2,383 users publicly shared at least 1 photo of their child
- Did you post anything embarrassing about your kids? 11% yes, 54% unsure

Privacy Perception Across Cultures

A 3-D camera then scans the customer's face to verify their identity. An additional phone number verification option is available for added security.

Alex Wong | Staff | Getty Images

An Alibaba employee demonstrates 'Smile to Pay', an automatic payment system that authorize payment via facial recognition

https://www.cnbc.com/2017/09/04/alibaba-launches-smile-to-pay-facial-recognition-system-at-kfc-china.html

Germany's Complicated Relationship With Google Street View

BY CLAIRE CAIN MILLER AND KEVIN J. O'BRIEN APRIL 23, 2013 4:39 PM 3

Street View, which Google started in 2007, is in 50 countries, including Germany. Johannes Eisele/Agence France-Presse — Getty Images

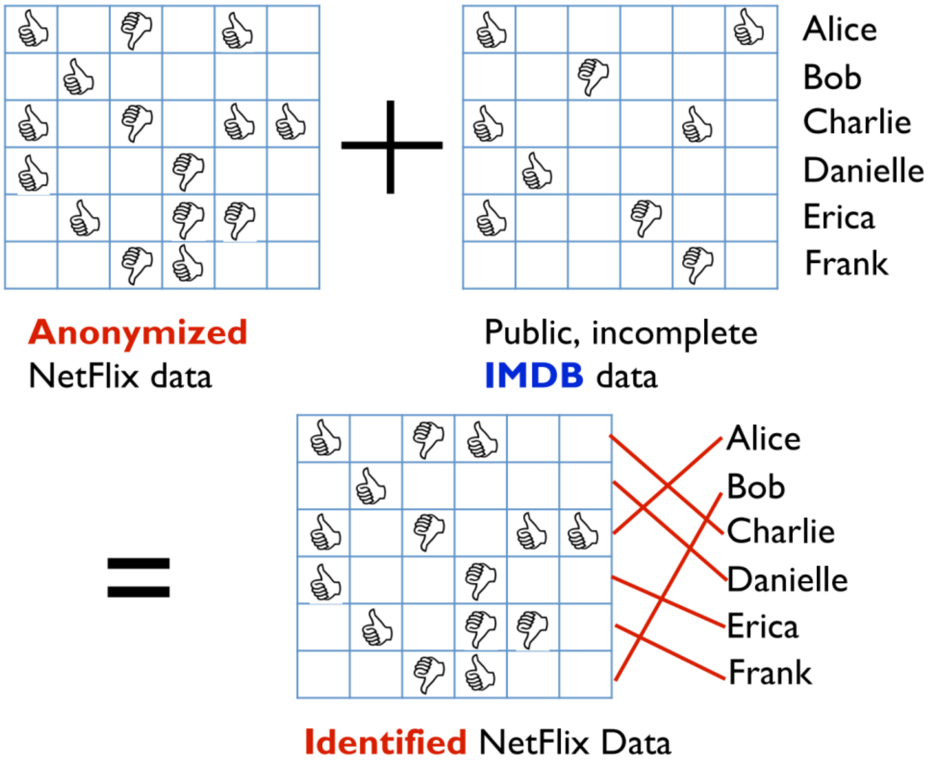
Email

Share

Germany is one of the most privacy-sensitive countries in the world. So when Google started taking pictures of buildings and homes for its Street View maps, some people were outraged, even though it was legal.

https://bits.blogs.nytimes.com/2013/04/23/germanys-complicated-relationship-with-google-street-view/

Privacy Attacks



Neural Models Leaking Privacy

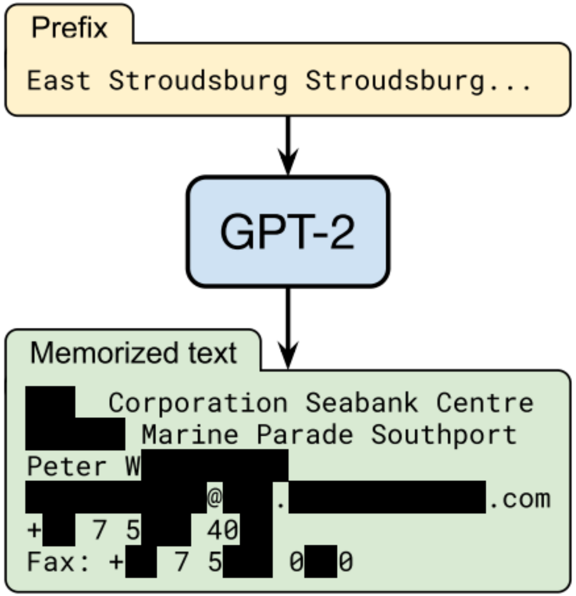


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Outline

Recap

What is Privacy?

Misuse of Privacy Information

Demographic Profiling

Authorship Obfuscation

Dangers in Mis-using Private Information

Why Do Some People Care To Protect Their Privacy?

- If the whole world will find out about my shoe size does it matter?

Why Do Some People Care To Protect Their Privacy?

- If the whole world will find out about my shoe size does it matter?
- Does it matter only if I’m doing something bad?

"If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place."

Eric Schmidt, 2009

Class discussion: do you agree with this quote?

Why Do Some People Care To Protect Their Privacy?

- If the whole world will find out about my shoe size does it matter?
- What about medical history
  - employer, insurance, airlines, etc.

- Public vs Private attributes

Dangers in Misusing Private Information

Examples of scenarios how people can be harmed

- Identity fraud with stolen SSN
- Medical records
- Private vs public accounts on social media: “People You May Know“
- Phone number, call history
- Location history
- Profile pictures across communities and social circles

Public vs Private Attributes

TECHNISCHE UNIVERSITÄT DARMSTADT

- Personally identifiable information (**PII**) or sensitive personal information (**SPI**)

“information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context”

Wikipedia

"any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."

NIST

PII

TECHNISCHE UNIVERSITÄT DARMSTADT

- Full name (if not common)
- Home address
- Email address
- National identification number
- Passport number
- IP address
- Vehicle registration plate number
- Driver's license number
- Face, fingerprints, or handwriting
- Credit card numbers
- Date of birth
- Birthplace
- Genetic information
- Telephone number
- Login name, screen name, nickname, or handle

PII

L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

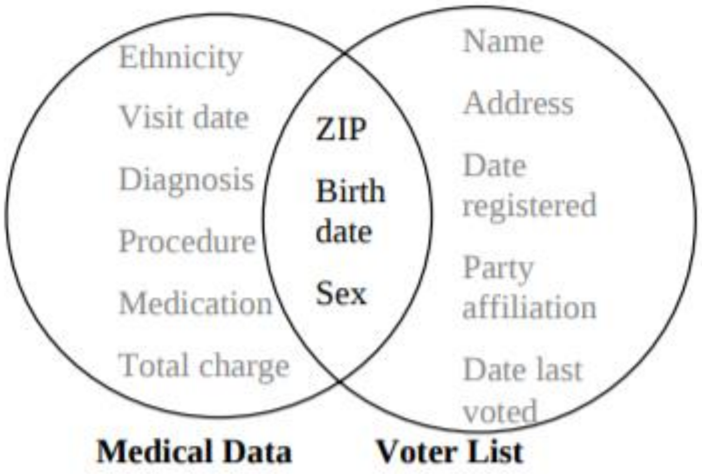


Figure 1 Linking to re-identify data

Without Security There Is No Privacy

- It is not illegal to collect PII
- Handling of PII data requires enhanced security
- End-to-end encryption

Outline

Recap

What is Privacy?

Misuse of Privacy Information

Demographic Profiling

Authorship Obfuscation

Demographic Profiling

Do We Really Need PII To Identify A Person?

- The absence of PII data does not mean that the remaining data does not identify individuals
- How many bits are needed to identify a person?
    - 8 billion people
        - Approximate gender
        - Approximate age, 3 buckets
        - Do they speak English?
        - Country? Occupation? Do they like hockey?
        - Chrome or Safari? Mac or Windows? Internet provider?

How Much Do We Need To Know To Identify A Person?

- The absence of PII data does not mean that the remaining data does not identify individuals
- How many bits are needed to identify a person?
  - 8 billion people
    - Approximate gender → 4 billion
    - Approximate age, 3 buckets → 1.5 billion
    - Do they speak English? → 800M
    - Country? Occupation? Do they like hockey? → 10M
    - Chrome or Safari? Mac or Windows? Internet provider? → 1M

To encode 8 billion people we need 35 Yes/No questions

For minorities you need much less, so these are more vulnerable!

What Can We Reveal?

“Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including:

- sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. “

Kosinski M., Stillwell D., and Graepel T. (2013) **Private traits and attributes are predictable from digital records of human behavior.** *PNAS*

1 Users' Facebook Likes

55,814 Likes →

58,466 Users ↓

	art	cnn.com	(...)	BMW
User 1	**1**	**1**	...	0
User 2	0	**1**	...	**1**
User 3	**1**	0	...	0
(...)	...	...	...	...
User n	**1**	**1**	...	0

User – Like Matrix
(10M User-Like pairs)

2 Singular Value Decomposition

100 Components →

58,466 Users ↓

	$Comp_1$	$Comp_2$	(...)	$Comp_{100}$
User 1	**1.5**	**.7**	...	**-.9**
User 2	**.3**	**-.4**	...	**-.2**
User 3	**-.6**	**.1**	...	**4.7**
(...)	...	...	...	...
User n	**1.2**	**1**	...	**-.6**

User – Components Matrix

3 Prediction Model

Using Logistic or Linear Regression
(with 10-**fold cross validation**)

e.g. $age=\alpha+\beta_1 C_1 + ... + \beta_n C_{100}$

Predicted variables

Facebook profile: **age, gender, political and religious views, relationship status, proxy for sexual orientation,** social network size and density

Profile picture: ethnicity

Survey / test results: BIG5 Personality, **intelligence, satisfaction with life,** substance use, parents together?

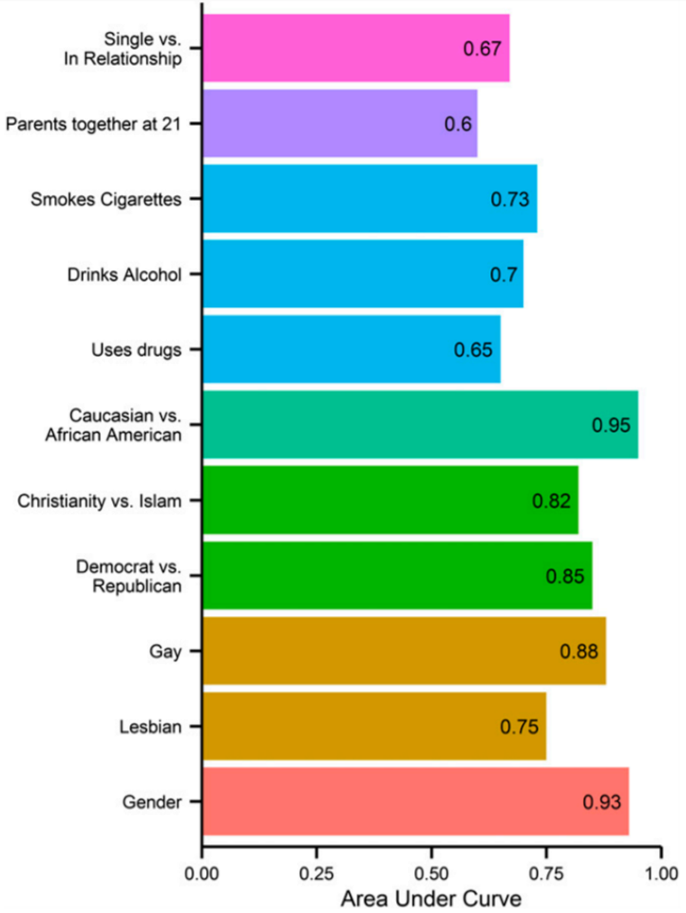


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

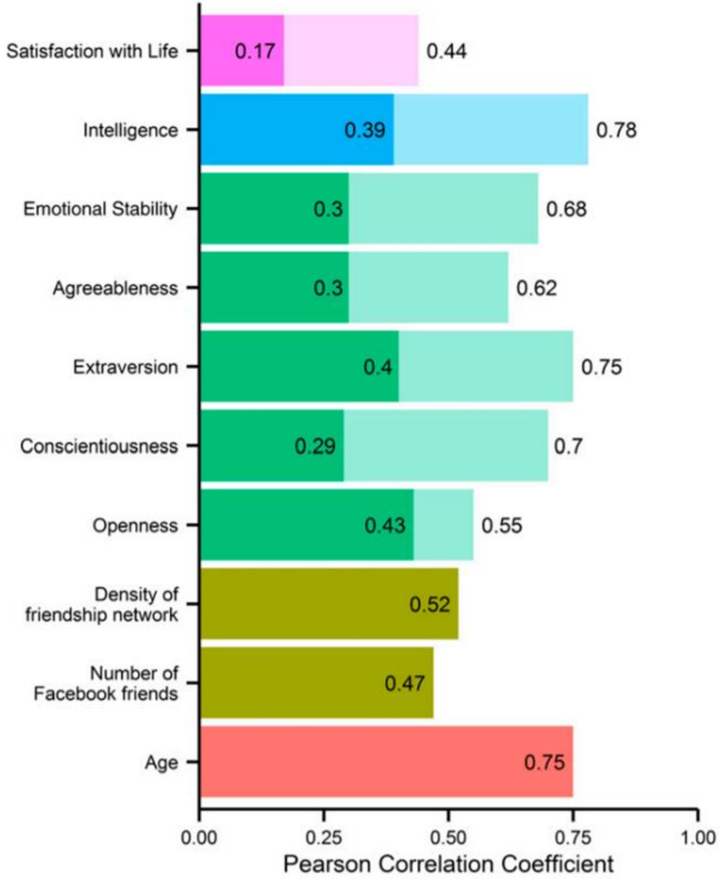


Fig. 3. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the $P < 0.001$ level. The

*58K people

What Can We Reveal From User's Language?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gender; Age; Location; Religion; Ethnicity; Social class; Diet; Personality type

Dong Nguyen, A. Seza Dogruöz,Carolyn P. Rosé, and Franciska de Jong (2016) **Computational Sociolinguistics: A Survey.** *Computational Linguistics*
- Sec. 3

Nina Cesare, Christan Grant, and Elaine O. Nsoesie (2017) **How well can machine learning predict demographics of social media users?** *https://arxiv.org/pdf/1702.01807.pdf*

Gender on Twitter

TECHNISCHE UNIVERSITÄT DARMSTADT

John D. Burger, John Henderson, George Kim and Guido Zarrella (2011)

Discriminating Gender on Twitter. *EMNLP’11*

Data

- Data collection
  - 213 million tweets from 18.5 million users, in many different languages
- Fields used
  - Screen name (e.g., *jsmith92, kingofpittsburgh*)
  - Full name (e.g., *John Smith, King of Pittsburgh*)
  - Location (e.g., *Earth, Paris*)
  - URL (e.g., *the user's web site, Facebook page, etc.*)
  - Description (e.g., *Retired accountant and grandfather*)
- Annotation of gender labels
  - tracking of user's labels across their accounts on social media platforms

Classification Results

Baseline (F)	54.9%
One tweet text	67.8
Description	71.2
All tweet texts	75.5
Screen name (e.g. *jsmith92*)	77.1
Full name (e.g. *John Smith*)	89.1
Tweet texts + screen name	81.4
Tweet texts + screen name + description	84.3
All four fields	92.0

Figure 5: Development set accuracy using various fields

Condition	Train	Dev	Test
Baseline (F)	54.8%	54.9	54.3
One tweet text	77.8	67.8	66.5
Tweet texts	77.9	75.5	74.5
All fields	98.6	92.0	91.8

Figure 6: Accuracy on the training, development and test sets

Simple Classifiers Perform Better than Humans

Baseline	54.9
Average response	60.4
Average worker	68.7
Average worker (100 or more responses)	62.2
Worker ensemble, majority vote	65.7
Worker ensemble, EM-adjusted vote	67.3
Winnow all-tweet-texts classifier	75.5

Figure 10: Comparing with humans on the all tweet texts task

Highly Weighted Features

Rank	MI	Feature f	$P(Female\|f)$
1	0.0170	!	0.601
2	0.0164	_:	0.656
3	0.0163	_lov	0.687
4	0.0162	love	0.680
5	0.0161	lov	0.676
6	0.0160	_love	0.689
7	0.0160	!_	0.618
8	0.0149	:)	0.697
9	0.0148	y!	0.687
10	0.0145	**my**	0.637
11	0.0143	love_	0.691
12	0.0143	haha	0.705
13	0.0141	my_	0.634
14	0.0140	_my	0.637
15	0.0140	_:)	0.697
16	0.0139	_my	0.634
17	0.0138	!_i	0.711
18	0.0138	hah	0.698
19	0.0137	_hah	0.714
20	0.0135	_so	0.661
21	0.0134	_haha	0.714
22	0.0132	**so**	0.661
23	0.0128	_i	0.618
24	0.0127	ooo	0.708
25	0.0126	!_i	0.743
26	0.0123	i_lov	0.728
27	0.0120	ove_	0.671
28	0.0117	ay!	0.718
29	0.0116	aha	0.678
30	0.0116	<3	0.856
31	0.0115	_cute	0.826
32	0.0114	i_lo	0.704
33	0.0114	:)$	0.701
34	0.0110	:(	0.731
35	0.0109	_:)$	0.701
36	0.0109	!$	0.614
37	0.0107	ahah	0.716
38	0.0106	_<3	0.857

464	0.0051	_ht	♂ 0.506
465	0.0051	hank	0.641
466	0.0051	too_	0.659
467	0.0051	_yay!	0.818
468	0.0051	_http	♂ 0.506
469	0.0051	_htt	♂ 0.506
624	0.0047	Googl	♂ 0.317
625	0.0047	ing!_	0.718
626	0.0047	hair_	0.749
627	0.0047	_b	0.573
628	0.0047	y_:	0.725
629	0.0046	Goog	♂ 0.318

Outline

Recap

What is Privacy?

Misuse of Privacy Information

Demographic Profiling

Authorship Obfuscation

Authorship Obfuscation

Authorship Obfuscation

- Obfuscation is the adversary task to identification
- Render identification of authors (or certain characteristics) impossible
- Obfuscation software should be:
  - Safe: Text cannot be attributed to original author
  - Sound: Text is paraphrase of original text
  - Sensible: Text is well-formed and unsuspicious

Authorship Obfuscation

- Remove most identifiable words/n-grams
  - “So” → “Well”, “wee” -> “small”, “If its not too much trouble” → “do it”
- Reddy and Knight 2016
  - Obfuscating Gender in Social Media Writing
  - “*omg I’m soooo excited!!!*”
  - “*dude I’m so stoked*”

Authorship Obfuscation

- Most gender related words (Reddy and Knight 16)

Twitter	
Male	bro, bruh, game, man, team, steady, drinking, dude, brotha, lol
Female	my, you, me, love, omg, boyfriend, miss, mom, hair, retail
Yelp	
Male	wifey, wifes, bachelor, girlfriend, proposition, urinal, oem corvette, wager, fairways, urinals, firearms, diane, barbers
Female	hubby, boyfriend, hubs, bf, husbands, dh, mani/pedi, boyfriends bachelorette, leggings, aveda, looooove, yummy, xoxo, pedi, bestie

Authorship Obfuscation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Learning substitutions
  - Mostly individual words/tokens
  - Spelling corrections “goood” → “good”
  - Slang to standard “buddy” → “friend”
  - Changing punctuation
- But
  - Although it obfuscates, a new classifier might still identify differences
  - It really only does lexical substitutions (authorship is more complex)

Privacy vs. Utility

Image: Mostly AI

Next Class

TECHNISCHE UNIVERSITÄT DARMSTADT

Anonymization and Privacy protection techniques

- Database Anonymization
  - k-anonymity
  - l-diversity
  - t-closeness
- Differential Privacy

Next Lecture

TECHNISCHE UNIVERSITÄT DARMSTADT

Privacy & Security II