Ethics in Natural Language Processing – SS 2022

Lecture 10
Language of Manipulation / Exam Preperation

TECHNISCHE UNIVERSITÄT DARMSTADT

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)

TECHNISCHE UNIVERSITÄT DARMSTADT

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	NO LECTURE – Illness
13	Language of Manipulation

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Introduction

Political Bots

Fake News & Influencing Elections

Recap: Data anonymization

TECHNISCHE UNIVERSITÄT DARMSTADT

- k-anonymity provides some anonymity, but can be vulnerable to certain weaknesses (Homogeneity, Background Knowledge)
- l-diversity improves k-anonymity by adding contraints to the diversity of the sensitive values
- t-closeness compares the distribution of sensitive values to the overall distribution (no explicit statements about eq. class size)

Recap: Differential Privacy

TECHNISCHE UNIVERSITÄT DARMSTADT

Basic idea: Introduce randomness

- Coin Toss Example: When asked about feature x for record y
  - Toss a coin: if heads give right answer
  - If tails: throw coin again, answer yes if heads, no if tails
- Still has accuracy at some level of confidence
- Still has privacy at some level of confidence (plausible deniability)

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

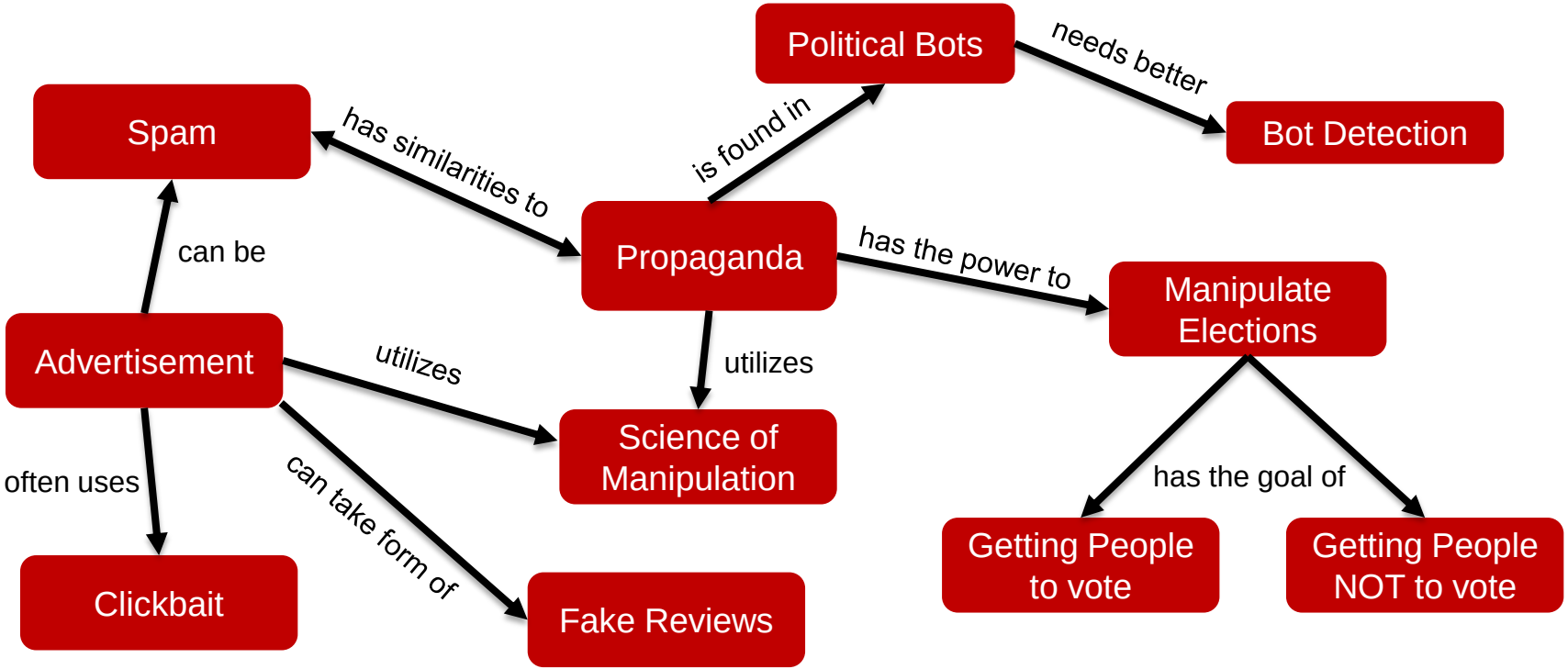
Recap

Introduction

Political Bots

Fake News & Influencing Elections

Concept Map of today's lecture



Learning Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

After hearing this lecture, you should be able to…

- Describe techniques that are used in propaganda
- Compare propaganda to spam, advertisements…
- Discuss the methods and effects in political comapigns (getting people to vote / not to vote)

History of Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Carthago delenda est!

History of Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Carthago delenda est!
- History is written by the winners
  - So it is biased (those losers never deserved to win anyway)
- Propaganda has existed even before writing
- But with mass media it has become more refined
  - Newspapers/pamphlets
  - Radio/Movies/TV/News
  - Social Media
  - Interactive Social Media (comments)
  - Personalized Propaganda targeted specifically to you sitting quietly in the second row

Propaganda vs. Persuasion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Propaganda is designed to influence people **emotionally**
- Persuasion is designed to influence people with **rational arguments (ish)**

- But its not that easy to draw the line **objectively**
  - They use propaganda to influence
  - We use rational arguments to inform

We vs Them

We have … Army, navy and air force Reporting guidelines Press briefings	They have … A war machine Censorship Propaganda
We … Take out Suppress Dig in	They … Destroy Kill Cower in their fox holes
Our men are … Boys Lads	Their men are … Troops Hordes

Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Demonize the enemy
  - “The only good bug is a dead bug”
- Personalize your side
  - “Our good boys ...”
- Be inclusive
  - “Good people like yourself ...”
- Be exclusive
  - “Never met a good one ...”

Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Obfuscate the source
- Nazi Germany makes a BBC-like show
  - Lord Haw Haw (William Joyce) “Germany Calling”
  - Sounded like a BBC broadcast (at first)
  - Talked about failing Allied Forces
  - Personalized to local places
- Flood with misinformation
  - To hide main message
  - Discredit a legitimate source
  - Add a sex story to deflect attention

Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- Doesn’t need to be True (or False)
  - Make up stories that distract
- But you can still just be selective with the truth
  - Marketing does this all the time
  - The most popular smart phone in the world
  - The most popular smart phone platform in the world
- Maybe truth plus distraction
  - Add a hint of a financial scandal

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Introduction

Political Bots

Fake News & Influencing Elections

Computational Propaganda

TECHNISCHE UNIVERSITÄT DARMSTADT

- People still generate base stories
- Language generation is not good enough (yet)
- But automated bots can magnify attention
  - Bots can retweet
  - Add likes
  - Give a quote and a link
- Build an army of bot personas
  - Be applied to many aspects of on-line influence

Political Bots

- @Girl4TrumpUSA created on Twitter
- Generated 1,000 tweets a day
- Mostly posting comments and links to Russian news site
- Deleted by Twitter after 38,000 tweets
- Many other similar bots
  - They amplify a candidate's support
  - Forward other messages (so you see things multiple times)
  - Ask: "what do you think about 'x'" (to get responses)
  - Like and retweet articles
  - Create fake trends on hashtags
    - Astroturfing vs Grassroots
  - Manufacture consent

How Many Bots

TECHNISCHE UNIVERSITÄT DARMSTADT

- Use crowd sourcing services to do tasks
- Can buy armies of bots with existing personas
- Start a twitter account
  - Buy a following of bots
  - High number followers attracts real followers
  - Bots will get deleted
  - Keep all the real followers
- There are offers of 30,000+ personas for sale

Bot Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Not very hard (at present)
  - Bot activity over time is quite different from humans
  - Bot post contents is often formulaic (its all rule driven)
  - Do damage, get detected and deleted – damage stays!

- Oxford Computational Propaganda Project
  - Published papers on bot types and detection techniques
  - They interviewed a bot maker
    - “How do you avoid your bots from being detected?”
    - “We read papers by you on what you do to detect us”

Bot Development

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most bot content is very formulaic
  - Generated from basic templates
  - Hand written
- Bot actions vs machine learning
  - Reinforcement learning
  - Send message1 to 50 people
  - Send message2 to different 50 people
  - Count number of clicks
  - Send most clicked message to 500 people
- Do this on more targeted messages to personalized interests
  - Send education message to person who mentioned education
  - Send healthcare message to person who mentioned healthcare

Automated Bot plus Humans

TECHNISCHE UNIVERSITÄT DARMSTADT

- Crowdworkers won't post propaganda for you
  - But they can still “help” with your propaganda…

- Please help with this propaganda detection problem
  - Here are 4 messages
  - Which ones are real, and which ones are bot generated:
  - “We’re the greatest”
  - “They’re the worst”
  - “Where is his birth certificate?”
  - “My granddaughter sent this link ...”
- Thank you for help with the propaganda ~~generation~~ detection problem

Comparison to Spam

TECHNISCHE UNIVERSITÄT DARMSTADT

- Spam: the mass distribution of ads (real or otherwise)
- It was successful at first (a few people clicked)
- People developed automatic spam detection algorithms
  - Mostly on usenet as that was the largest forum at the time
  - Then in email
  - Detection improved, but it is still there
- We still receive spam, though mostly we ignore it
- Other much more sophisticated marketing is now common
  - And more acceptable
  - Google links to purchasing options
  - Amazon recommendations
- So spam is contained and mostly ignored

Can Propaganda become like Spam

TECHNISCHE UNIVERSITÄT DARMSTADT

- People send spam **if** it works
  - Spam working means people “buying”
- People send propaganda **if** it works
  - Propaganda working means people … voting (?)
  - Which isn’t as important as buying the best smart phone :-(
- People may become more sophisticated with propaganda
  - Learn to ignore it (but what of those who don’t)
  - But it will become more targeted to the unsophisticated
- Propaganda messages may become more sophisticated
  - Control your news bubble/echo chamber
- Propaganda messages may drift to informative messages
  - People will learn to evaluate both sides of the issue and make informed decisions

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Introduction

Political Bots

Fake News & Influencing Elections

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam.

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam
- Buy Me! – sent to only those who might buy me
  - Hard to target that population (and you want more people to buy)

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam
- Buy Me! – sent to only those who might buy me
  - Hard to target that population (and you want more people to buy)
- Buy Me! – I'll help you with your latest endeavor
  - Try to target the interest of new people to buy me

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam
- Buy Me! – sent to only those who might buy me
  - Hard to target that population (and you want more people to buy)
- Buy Me! – I'll help you with your latest endeavor
  - Try to target the interest of new people to buy me
- Buy Me! – I'll help you with <your latest endeavor>
  - Actually personalize the message to include personalized phrases

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam
- Buy Me! – sent to only those who might buy me
  - Hard to target that population (and you want more people to buy)
- Buy Me! – I'll help you with your latest endeavor
  - Try to target the interest of new people to buy me
- Buy Me! – I'll help you with <your latest endeavor>
  - Actually personalize the message to include personalized phrases
- Buy Me! – I'll help you with <your latest endeavor>
  - "It helped my granddaughter with her latest endeavor" – John from Pittsburgh

Let's Advertise ...

TECHNISCHE UNIVERSITÄT DARMSTADT

- Buy Me!
  - People don't always respond to general spam
- Buy Me! – sent to only those who might buy me
  - Hard to target that population (and you want more people to buy)
- Buy Me! – I'll help you with your latest endeavor
  - Try to target the interest of new people to buy me
- Buy Me! – I'll help you with <your latest endeavor>
  - Actually personalize the message to include personalized phrases
- Buy Me! – I'll help you with <your latest endeavor>
  - "It helped my granddaughter with her latest endeavor" – John from Pittsburgh
- "Everybody bought me and you wont believe what happened next ..."
  - Your whole sphere seems to have bought me.

How to write / detect Fake Reviews

TECHNISCHE UNIVERSITÄT DARMSTADT

- Try to be a verified purchaser
- Be specific about the project
  - Not just … “Great product, arrived on time”
- Add some self disclosure for realism
  - “My 6 year old granddaughter loves it, “Granny, I love my Tesla K80 24GB GPU” she says.
- Generate multiple different reviews
  - Different classes of user
  - “Works great on Linux”
  - “Works on my Mac”
- But reviews are (still) best written by humans
  - They can be adapted automatically, and posted automatically
- Automatically posted when someone mentions the product

Review vs News

- “News” is perceived to be more authoritative
  - But user-written “reviews” are more genuine
- Many “news” articles also advertise the product
- Many ads are press releases designed to be quoted as news
- You can make your reviews be like news.
- You have to release them via a recognized News site
  - … or not
- Different headlines but same story
  - Looks like there is more news about X
- Generate references to the articles
  - Pay for links
  - Tweet/retweet about them

Clickbait

- Making people click on links
- Things they like
  - Kim Kardashian something something
- Things they want to know
  - Next Avengers movie will be released …
- Things left unsaid
  - Something, something, you wont believe what happened next
- All using reinforcement learning to find the best headline
  - Kardashian Avengers bitcoin deep learning, you wont believe what happened next ….

So what happened to Truth?

TECHNISCHE UNIVERSITÄT DARMSTADT

- It maybe never was there …
  - News reports about things I know about are always wrong in the details, I'm just pleased that all the other news is correct
- We could fact check everything
  - "flat earth" 500m google hits vs "spherical earth" 200m
- Identify "good" sources of facts
  - But we actually want opinion too
  - Who decides truth?

Trustworthiness

- Jeff Pasternack and Dan Roth at UIUC/UPenn
- Identify sources for fact checking
- Idea: Present multiple views when searching
  - “Is milk good for you?”
  - Gave side-by-side search results for and against
  - This was preferred by most subjects (sometimes)
- But probably wont work when people are already charged in one direction

https://argumentsearch.com/ (early prototype)

Confirmation Bias

TECHNISCHE UNIVERSITÄT DARMSTADT

- Humans see things to confirm their biases
  - “Well that’s probably only one example” vs
  - “I bet there are many more examples like this”

- Arguments are rarely actually rational debates
  - Besides your just clearly wrong anyway ...

Exploiting Human Behavior for Gain

TECHNISCHE UNIVERSITÄT DARMSTADT

- You probably can't (really) change peoples views
- But you can amplify them

- I'm a democrat but my vote doesn't really count
  - Healthcare will still be too expensive under either party
  - News: "Democrats will cut healthcare costs"
  - Okay maybe I will vote

Getting People to Vote

TECHNISCHE UNIVERSITÄT DARMSTADT

- Rayid Ghani, Chief Scientist of Obama campaign 2012
  - Masters from MLD, now at University of Chicago leading “Data Science for Social Good”

- Amplifying Activism
  - Find marginal constituencies (regions / states with unclear winners)
  - Find registered democrats in the area
  - Identify their key interests (education, healthcare, etc)
  - Send them messages about their key interests
  - Ask for donations
  - Measure success in sending messages
  - Do it again

Getting People to Vote

TECHNISCHE UNIVERSITÄT DARMSTADT

- Attenuating Apathy
  - Find marginal constituencies (regions / states with unclear winners)
  - Find registered democrats in the area
  - Identify their key interests (education, healthcare etc)
  - Send them messages about their key interests
  - Get them worked up about the election
  - Get them to vote
  - Best to do that <24h before the election
- It doesn’t take much to change an election result

Getting People Not to Vote

Getting People Not to Vote

Getting People Not to Vote

TECHNISCHE UNIVERSITÄT DARMSTADT

- Deflect voters
  - Its not worth voting
  - Poll estimates show X is overwhelmingly winning
  - “Trump does not have a chance”…
- Mislead voters
  - Vote by text to ….
  - Vote early on March 9th (but its actually March 6th)
  - You need government ID to vote

Misleading Voters Through News

TECHNISCHE UNIVERSITÄT DARMSTADT

- Show relevant News stories
  - Stories of interest to the particular voter
  - No longer a general editor/newspaper
  - Only see things in your news feed
  - Overwhelmed with obviously fake stories so ignore everything
  - Add fake facts to real stories
  - Question objectivity itself
  - Call “Fake News” for anything you don’t like

Can this be stopped?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Companies and Countries already do that
  - “Russia did it all”, “It was North Korea’s fault”
  - Could be an excuse, true, or just misinformation
- Where to draw the line?
  - What is the difference between political campaigns and Cambridge Analytica?
- Can you ever define legality
  - You must allow people to campaign
  - You have to avoid creating unfair laws about campaigning
  - You want to stop unfair vote manipulation
- Does microtargerting actually work?
  - Depends on who you ask…

Science of Manipulation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Marketing and Advertising
  - We want to influence people
- Public Service Announcements
  - Influencing the populace to do “good” things
- Psychology
  - Studying human behavior
  - Modeling group behavior
- Manipulation for good/bad

Unseen Consequences

TECHNISCHE UNIVERSITÄT DARMSTADT

- Its not just about deliberate/opportunistic manipulation
- Access to diverse information flow
  - Allows personalization of choice of interests
  - Moves your information flow to areas of interest
- But with personalization comes limitations
  - You only see the areas you want to see
  - Your own information bubble
  - “But everyone I talk to online likes My Little Pony”
  - You never see people liking other things, so your “normal” changes

Exam information

Exam Formalities

TECHNISCHE UNIVERSITÄT DARMSTADT

- September 14th, 2022
- Rooms will be announced
- Time: 10:00 – 12:00

- Be sure that you are registered for the exam in TuCan

- **Student id and photo id needed!**

Exam Formalities

TECHNISCHE UNIVERSITÄT DARMSTADT

- No books, notes, or other auxiliary material may be used.
- Problems are stated in English.
- The questions may be answered in either German or English.
- You have 90 minutes to complete the exam.
- Write you answers in the specified answer fields and only use the extra answer sheets provided by the examiners.
- After the exam, the full question book must be returned, including the additional answer sheets.

Additional Info

TECHNISCHE UNIVERSITÄT DARMSTADT

- No electronic devices of any type are allowed.
  - That also means no calculators! (If you can multiply simple fractions, you are fine)
- Basic NLP and machine learning terms should be known.
  - Examples: Tokenization, precision, recall…
- Counterexamples (things you do not need to know):
  - The specifics of Voice Identification tool x, Deep Learning Model y, Data set z…

What can we expect?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Knowledge questions
  - Definitions
  - Explainations of model X or term Y
  - Give examples for concept Z
- Discussion questions
  - Assess ethical questions / concerns for a use-case
- Small calculations
  - Precision / Recall, Data Anonymization

What can we expect?

TECHNISCHE UNIVERSITÄT DARMSTADT

- One or two questions from the exercises

For example:

  - Small code snippet, „What is the output?“
  - Questions that need the input from the exercises (Word Embeddings, Sentiment…)

Some example questions

TECHNISCHE UNIVERSITÄT DARMSTADT

True or false:

Irrational, negative attitude towards a group of people is called **Stereotype Threat**.

Answer: False
(It is called Bias)

Some example questions

Given an abstract for a made-up AI system, assess the fictional scenario from an ethical perspective by **asking and answering** three questions to evaluate the system.

e.g.:

Who could benefit? Who could be harmed? What is the "cost of misclassification"?

Could the data be biased? Could sharing the data effect people's lives?

Some example questions

TECHNISCHE UNIVERSITÄT DARMSTADT

What are Implicit Biases?

Answer:

Implicit biases are irrational attitudes that are distressingly pervasive, operate largely unconsciously, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

Some example questions

Give two examples for characteristics of Hate Speech.

Answer:

Hate Speech is **targeted** towards a person or group on the basis of some characteristic. Hate Speech is used **intentional** to humiliate or insult members of this group. Hate Speech is used for the **effect** to threaten or insult. Hate Speech is **motivated and reasoned** by the offender's bias towards an aspect of a group of people.

Some example questions

Male		Predicition	
		True	False
Reality	True	50	0
	False	0	150

Female		Predicition	
		True	False
Reality	True	50	0
	False	100	50

A system predicts a characteristic for 200 male and 200 female students.

Do the system predictions, across male/female subgroups, have…

a) Equal precision?
b) Equal recall?
c) Equal precision AND recall?
d) Neither equal precision NOR recall?

Some example questions

Male		Predicition	
		True	False
Reality	True	50	0
	False	0	150

Female		Predicition	
		True	False
Reality	True	50	0
	False	100	50

A system predicts a characteristic for 200 male and 200 female students.

Do the system predictions, across male/female subgroups, have…

a) Equal precision? – NO, precision is 1.0 for male, 0.33 for female
b) Equal recall? – YES, recall is 1.0 for both
c) Equal precision AND recall? – NO
d) Neither equal precision NOR recall? – NO

Additional Info

TECHNISCHE UNIVERSITÄT DARMSTADT

- Other questions about organisation or content: Use moodle forum!

- If you need further consultation:

arnold@ukp.informatik.tu-darmstadt.de