COURSE SUBJECTIVITY MINING

ISA MAKS

AUTOMATIC IDENTIFICATION OF ONLINE HATE SPEECH

COURSE OVERVIEW

week 1 (05-09)	Introduction hate speech identification (Isa Maks)	
week 2 (12-09)	Annotating hate speech (Isa Maks)	*assignment 1 (16/09)*
week 3 (19-09)	Hate speech datasets (Isa Maks)	*assignment 2 (01/10)*
week 4 (26-09)	Automatic classification of hate speech: methods (Ilia Markov)	
week 5 (03-10)	Automatic classification of hate speech: ensemble methods (Ilia Markov)	*assignment 3 (15/10)*
week 6 (10-10)	Guest Lecture: identification of targets of hate speech (Baran Barbarestani)	
week 7+8		*final assignment 4 (29/10)*

- dr. Isa Maks (teacher, course coordinator)
- dr. Ilia Markov (teacher)
- Harshita Choudary (teaching assistant)
- Baran Barbarestani - MA (guest lecturer

OVERVIEW OF THIS LECTURE

- What is online hate speech (some examples)
- Hate speech detection in relation to sentiment and subjectivity mining
- The problem of online hate speech (in society)
- How to approach online hate speech detection
- Conclusions

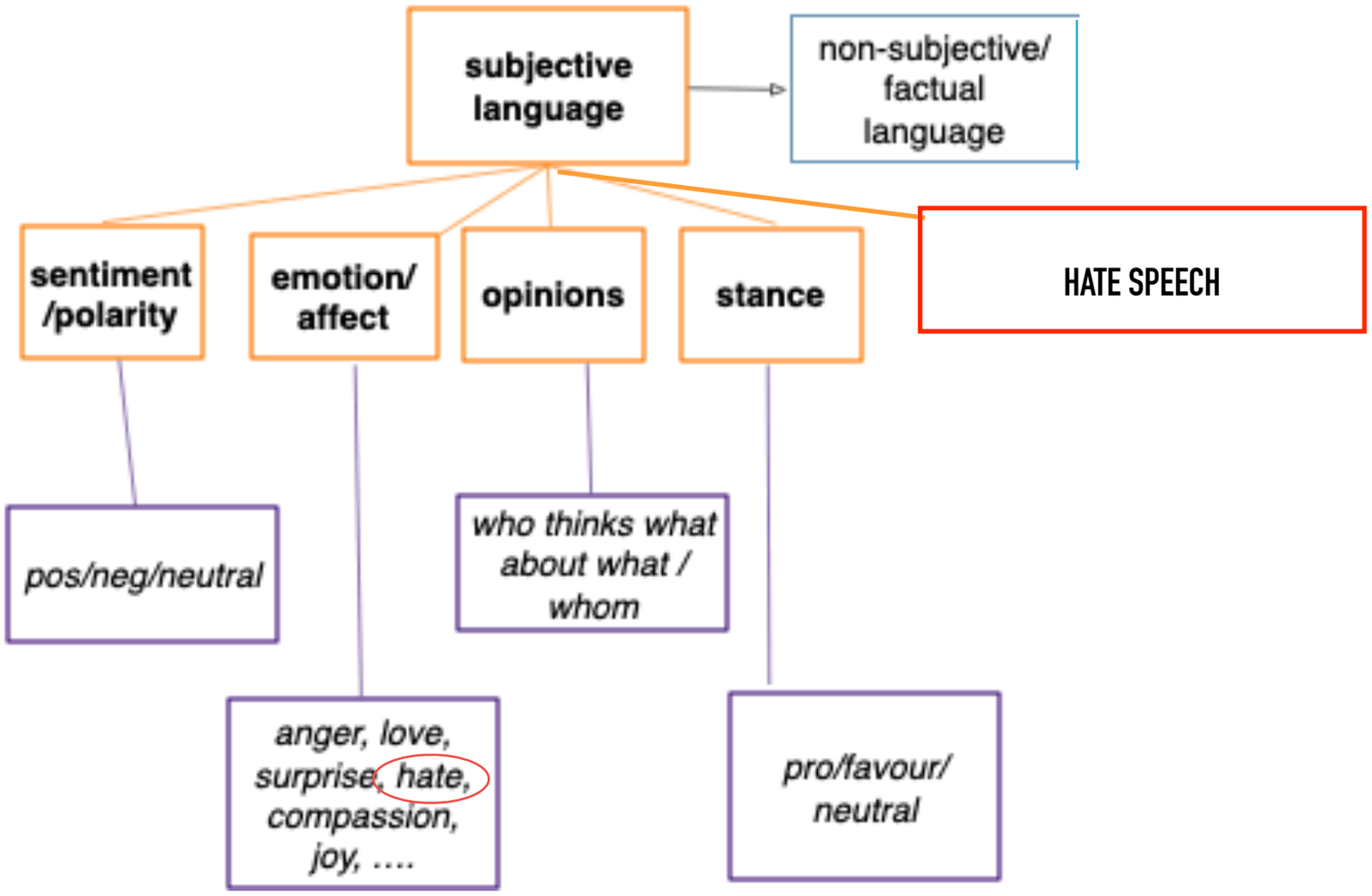
Disclaimer:

Due to the nature of the subject these slides contain offensive language. This does not reflect the opinion of the author(s)

EXAMPLES OF HATE SPEECH

- Jews and n*ggers destroy & pervert everything they touch #jewfail #niggerfail
- How is Mexico doing today? These people come here because they couldn’t build it
- Illegal immigration is a cancer which if not eliminated will bring the downfall of Europe and Europese culture
- [about a Roma family] If that kind of think happened to me I would not tell the police. I would set fire to the house . No mercy
- [about Muslims] “Poorer, dumber, breeding like rats. They’ve got it all,”. “India needs to eliminate them before they rise up,”
- Women drivers should be banned from the road

SEVERAL TASKS



WHAT IS SUBJECTIVITY MINING

- Subjectivity mining also known as sentiment analysis or opinion mining aims at understanding the underlying sentiment of unstructured content
  - Text mining : automatically extracting of information from text
    - Subjectivity: a general term that covers people's inner states i.e. their opinions, beliefs, doubts, thoughts, feelings, emotions, goals, evaluations, and judgments (Wiebe, 2005) as expressed in text and conversations
- As opposed the text mining of factual content such as dates, places, names, products, etc.

WHAT-WHO-WHEN-WHERE-WHY-FEELINGS

Groningen hit by strong earthquake as gas extraction impact continues

Where

When

When

The province of Groningen was hit by a strong earthquake in the early hours of Wednesday morning, as the ground continues to settle following the extraction of natural gas. Hundreds of people have reported feeling the quake, which hit shortly before 6am. 'The people of Groningen were shaken away,' one person said on Twitter. By 11am, officials had received 90 reports of damage, including 12 requiring immediate assessment, news agency ANP said. The quake measured 3.4 on the Richter scale, making it the third strongest in the province since the problems began. 'This will have caused damage,' a spokesman for the KNMI seismology unit said.

Feeling

Prime minister Mark Rutte told television show Goedemorgen Nederland that he hoped the damage had been limited. In the 1950s everyone was so optimistic about the gas find, but it has now changed into a nightmare, Rutte said. In 2012 the province was hit by a quake measuring 3.6 on the Richter scale which caused considerable damage to hundreds of homes and other buildings. Since then pressure has mounted on the government to wind down production and last year the government decided that gas extraction should stop altogether in 2030.

When

SUBJECTIVE VS. OBJECTIVE

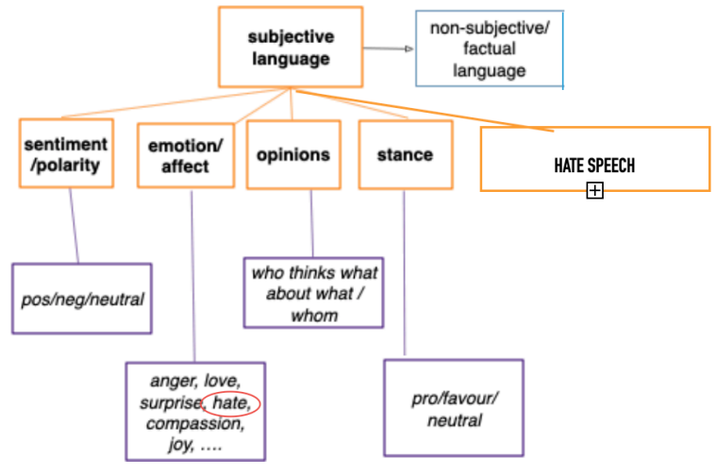
- (OBJECTIVE) Philip Roth was born in *1933* in Newark, New Jersey. He was an American novelist and short story writer . His first novel Goobdye won the National Book Award in 1960. (Named entities and dates)
- (SUBJECTIVE: Subjectivity: covers people's inner states i.e. their opinions, beliefs, doubts, thoughts, feelings, emotions, goals, evaluations, and judgments (Wiebe, 2005) as expressed in text and conversations
    - I love this movie (positive)
    - The streets are very crowded (positive?)
    - There are mice running in the office (negative?)
    - He is slamming with the doors. (He is angry?)

ANSWERS TO QUESTIONS LIKE

- Do people like the newest smartphones
- What kind of hotels people like to go to and why?
- How do people feel about recent events in the Middle-East?
- What is the range of opinions being expressed about the best course of action in Ukraine?
- Is the rhetoric from a particular group intensifying?
- How do people feel about gay rights? Do their attitudes change?
- Does the attitude towards vaccinations change over time? Why?
- What do people think about Brexit? Are there differences between Scotland, North Ireland and the UK?
- Do we find more emotions in books from the 16th or the 17th century? What type of emotions?

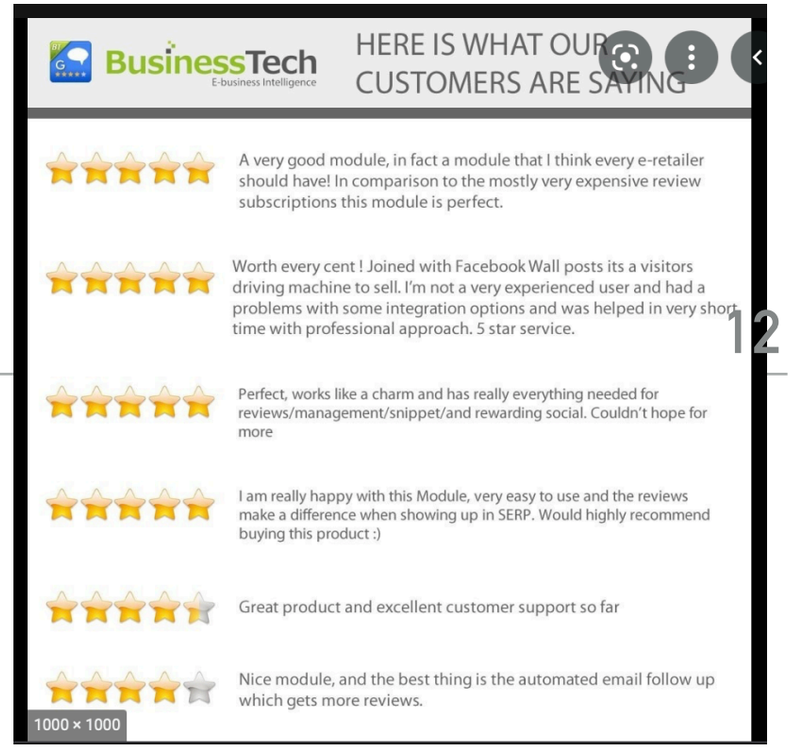
CURRENT TASKS

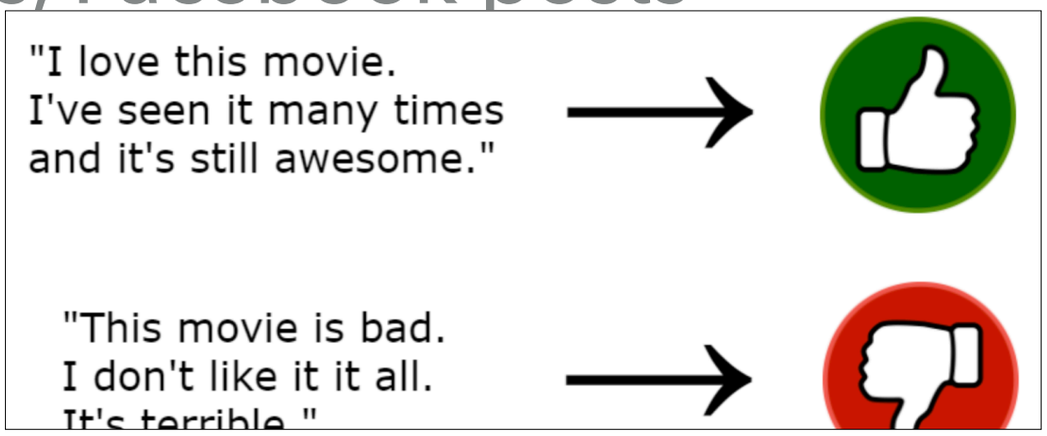
- Sentiment analysis
- Aspect-based sentiment analysis
- Stance detection
- Opinion mining
- Emotion analysis
- Hate speech detection
- ..



SENTIMENT ANALYSIS

- Are they positive, negative or neutral:
  - Product Reviews
  - Social media : Tweets, Facebook posts
  - letters, diaries, lyrics





SENTIMENT ANALYSIS ON LYRICS, TWEETS

Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter...I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mofette briliant! May the fourth be with you #starwarsday #starwars	Positive
Good morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

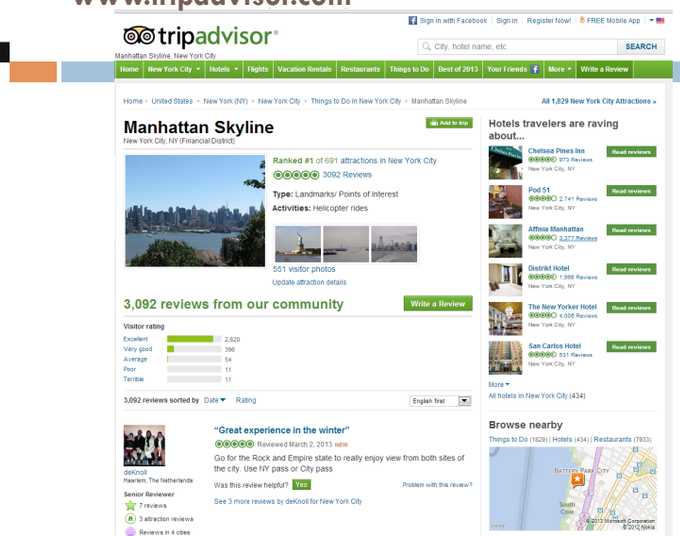
948 × 540

Artist	Neg	Neutral	Pos
Cigarettes After Sex	3.5	86.3	10.1
Eric Clapton	2.2	80.0	17.8
Damien Rice	6.1	87.6	6.3
Dire Straits	5.2	84.2	10.6
The Black Keys	6.9	85.4	7.7
Eminem	12.6	80.1	7.3
Porcupine tree	3.5	89.9	6.7
Northlane	8.3	88.3	3.4
Incubus	3.4	83.2	13.5
Radiohead	9.9	83.0	7.1

Negative lyrics¶

Eminem wins the negative lyrics contest (not surprisingly), with 12.6% of his lyrics being negative. **Radiohead** comes in second, with 9.9%. **Eric Clapton** has the least negative lyrics, at 2.2%. One surprise is **Porcupine tree**,

www.tripadvisor.com



STANCE DETECTION

- Tweets
- Ideological stance
- Stance and sentiment

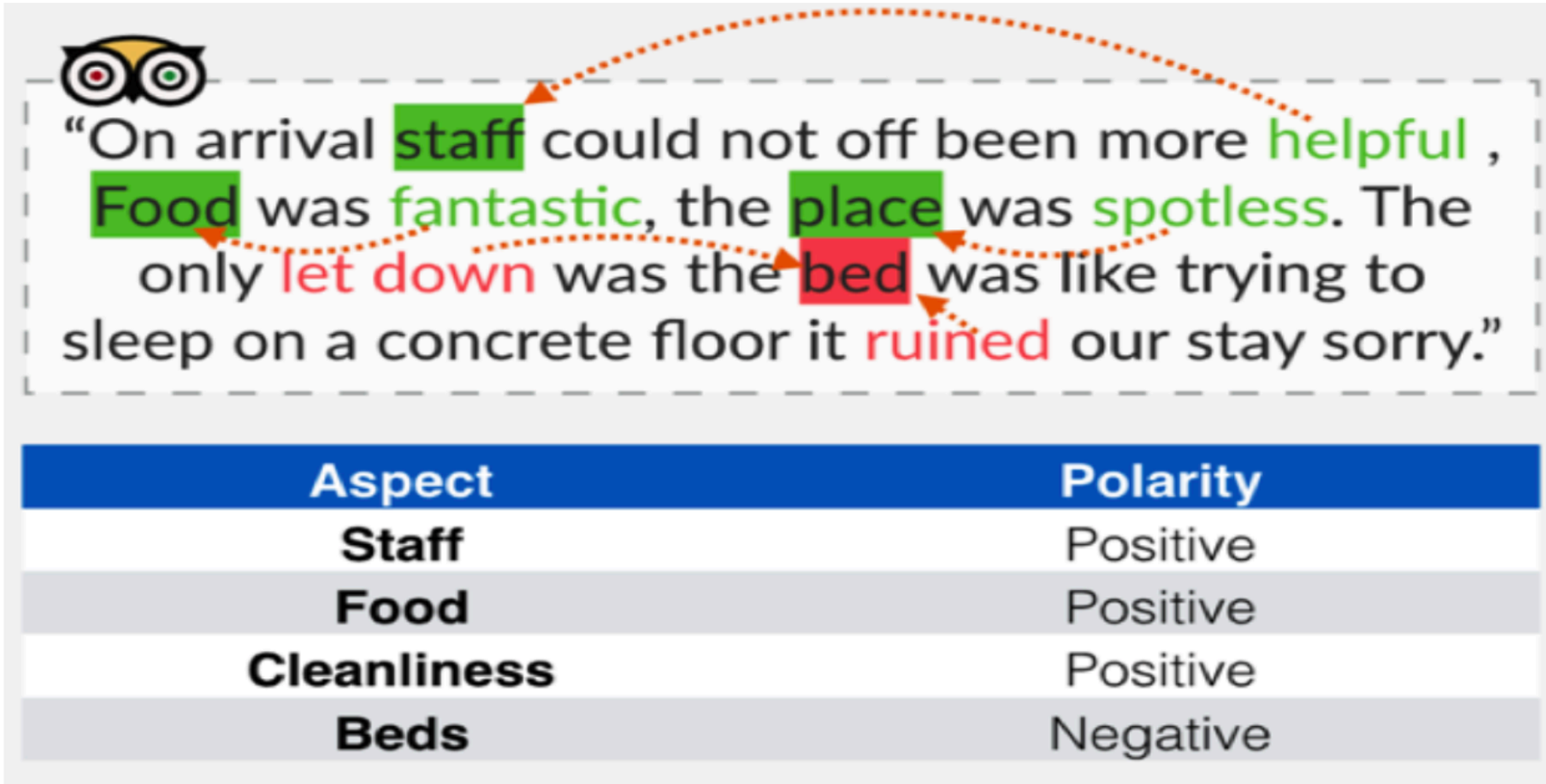
Table 1. sample of tweets illustrating the sentiment polarity of the expressed stance

#	Tweet	Target	Sentiment	Stance
1	It is so much fun having younger friends who are expecting babies. #beentheredonethat #chooselife .	Legalisation of Abortion	+	-
2	Life is sacred on all levels. Abortion does not compute with my philosophy. (Red on #OITNB)	Legalization of Abortion	0	-
3	The biggest terror threat in the World is climate change #drought #floods	Climate Change is the real concern	-	+
4	I am sad that Hillary lost this presidential race	Hillary Clinton	-	+

From: SEMEVAL-2016 (task 6): Detecting stance in tweets

ASPECT-BASED SENTIMENT ANALYSIS

- Identify target words
- Identify sentiment towards target words
- Group target words into product aspects



Aspect	Polarity
Staff	Positive
Food	Positive
Cleanliness	Positive
Beds	Negative

Example of Aspect based sentiment analysis — Source: https://medium.com/seek-blog/your-gui sentiment-analysis-344d43d225a7

OPINION MINING

Opinion triples:

- Whose opinion (opinion holder)
- Positive or negative
- Strong or weak
- About what (opinion target)

- … and why .(argument mining)
  she (OH) didn't like Greece(OT) as it was too hot (WHY) for her

- *News articles*

What is an opinion (Kim, Hovy 2006)

	sentence1	sentence2
Opinion holder	Speaker/writer	Bush
Opinion expression	negative	negative
Opinion target	Bush (for the economy)	Obama's behaviour

judgment

(1) Bush is bad for the economy

(2) Bush is angry about Obama's behaviour

emotion -> judgment

Who is **negative/positive** about **what**

OPINION MINING IN NEWS PAPERS

Washington (AFP) - President Barack Obama said Thursday that he had included openly gay athletes in the US Olympic delegation to show the **United States** would **not abide discrimination** in sport or anywhere else. 32

His comments, in an interview with NBC on the eve of the Sochi Winter Olympics, came after a **senior Russian official warned** ?**athletes or spectators?** should **not promote gay rights** during the Olympics following the passage of controversial anti-gay legislation in Russia. **Obama picked** several **openly gay former athletes** in the US delegation to the opening of the games and pointedly did not dispatch a cabinet-level official or a member of his family to Sochi

20 10 4

United States (Opinion Holder) not abide (negative) discrimination (opinion target)

MAIN ISSUES FOR (AUTOMATIC) ANALYSIS OF SUBJECTIVE EXPRESSIONS(1)

- Definition:
  - Is it possible to detect somebody's "inner state" (he slammed with the door, the bar was crowded) {definition)
  - What exactly is positive, negative , neutral .
- Annotation schemas"
  - What emotions are expressed in text; (the was sad about the loss of his father).
  - Opinion triplets
- Extra knowledge needed for interpretation (annotation problems)
  - Context needed for interpretation: *the rooms are small, but the hotel is very close to the city center*
  - Domain knowledge: *the rooms are small, the iPhone is big*
  - World knowledge: *mice are running through the office*
- Diversity in genre, domain, topic

MAIN ISSUES FOR (AUTOMATIC) ANALYSIS OF SUBJECTIVE EXPRESSIONS(2)

- Creative words (social media): *bedrijfspoedel (office dog)*
- Implicit language: *luckily these guys cannot have children (hateful)*
- Sarcasm : *my favourite thing to do at 4AM in the morning is going to the airport. How about you?*
- Words are string cues but polysemous: alarm (device vs. Feeling);
- Diversity in surface forms: I think this is alright/ OK/ Do that again,

ONLINE HARMFUL BEHAVIOUR

- Online ontspoord (2021) Rathenau instituut

What is the nature and scale of harmful and immoral behaviour online in the Netherlands, what are the underlying mechanisms and causes, and what options for action are available to the ministry, the government and society as a whole, for limiting harmful and immoral behaviour online?

TAXONOMY OF ONLINE HARMFUL BEHAVIOUR

- All phenomena that an internet user may encounter sooner or later
- The internet has certain characteristics that end to inspire, facilitate and catalyse harmful behaviour online. *A person who would never insult a passer-by on the street may have no trouble doing so on Twitter*

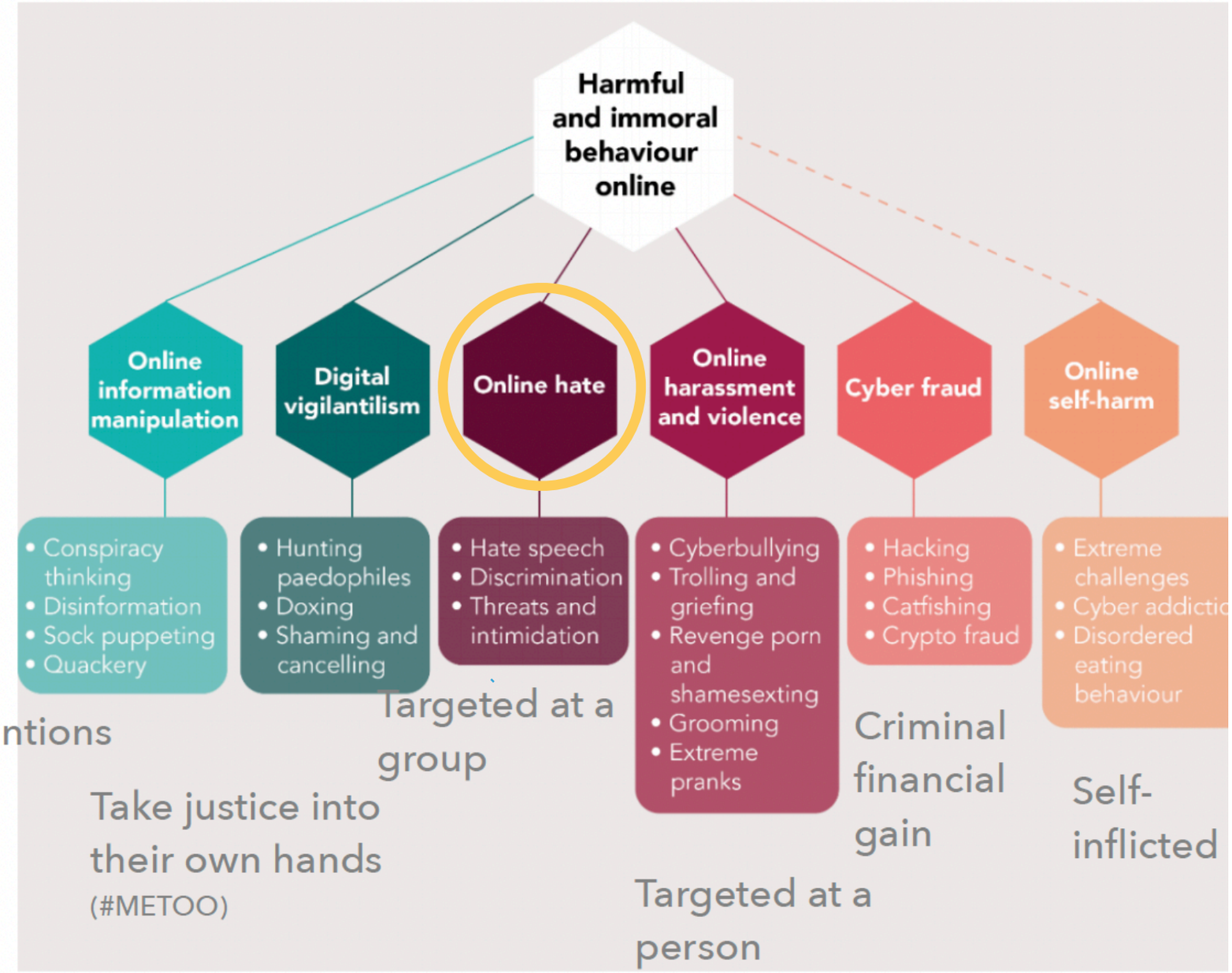


Figure 1 **Taxonomy of harmful and immoral behaviour online[2]. Source: Rathenau Instituut**

Source: Online ontspoord (2021) Rathenau instituut

MECHANISMS

- The study identified a total of 18 online properties and mechanisms that play a role in inspiring, facilitating and driving harmful and immoral behaviour online
- The case studies in the report show that the same mechanisms can play a role in very different phenomena, and that the mechanisms occur in combination.

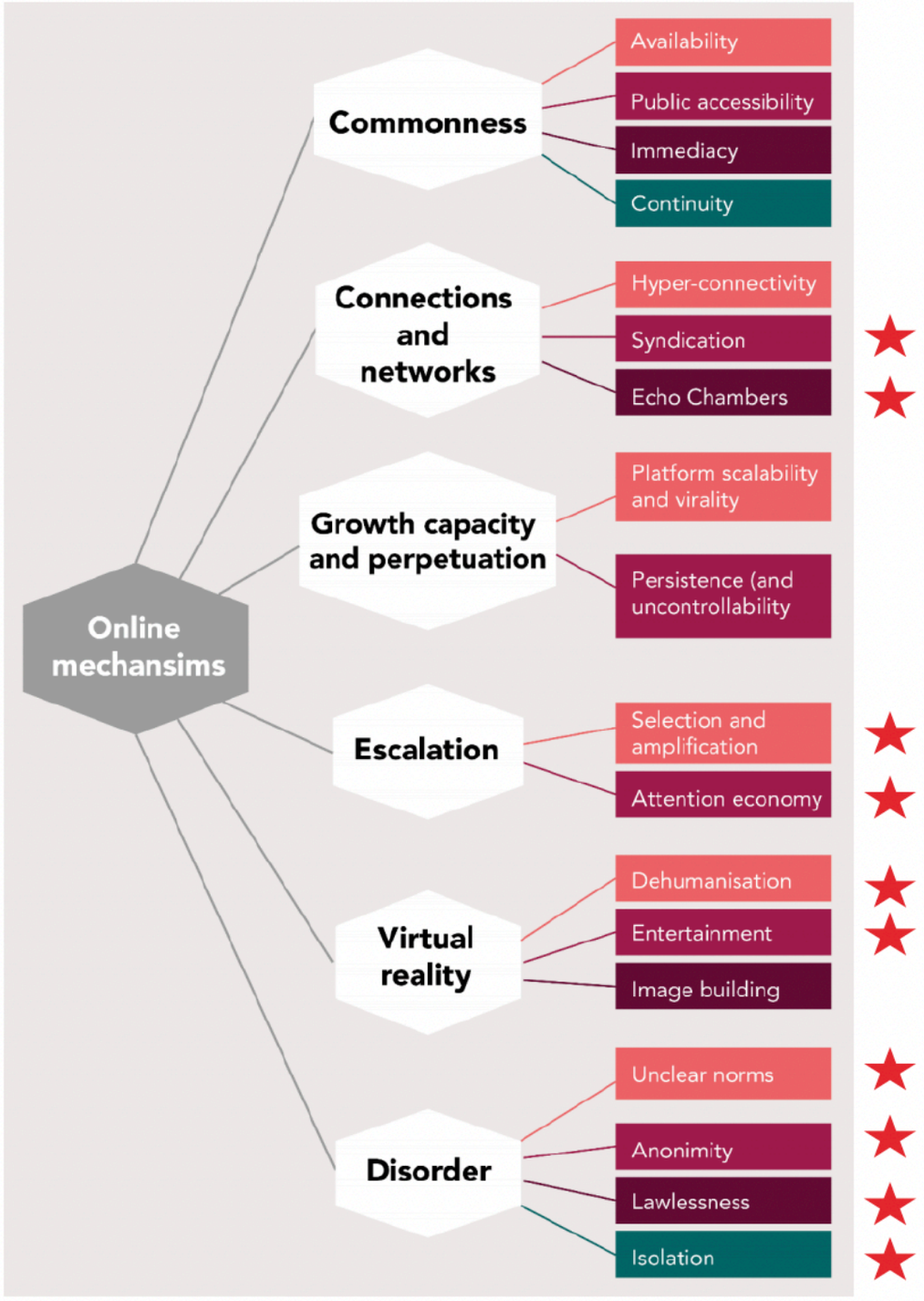


Figure 2 Overview of online mechanisms. Source: Rathenau Instituut

INTERVENTIONS

- Government:
  - legislation (helping victims, punishing offenders)
  - Make companies responsible
- Industry:
  - Change revenue models
  - Label, monitor text
- Society:
  - creating awareness
  - reach out to victims

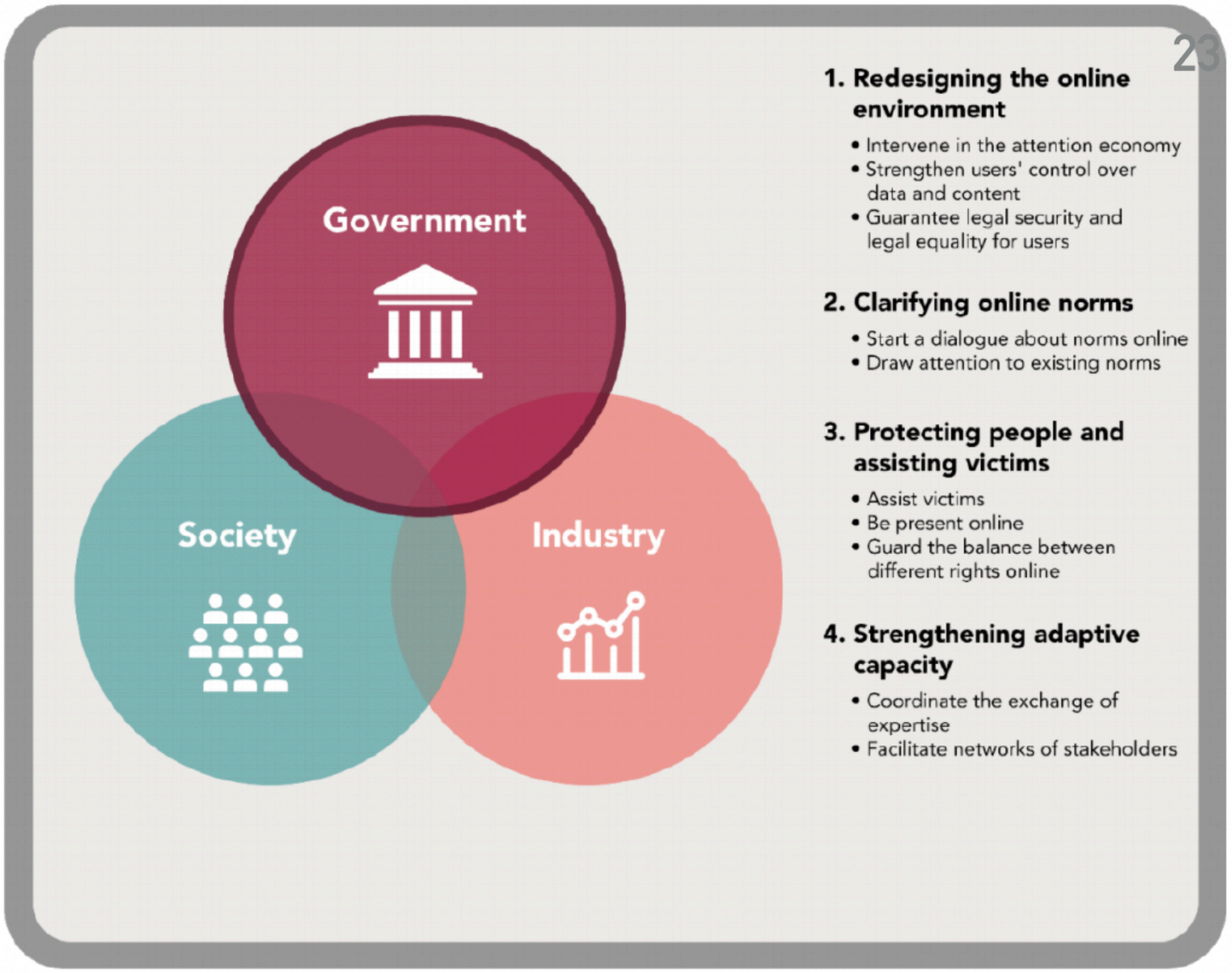


Figure 3 Strategic agenda. Source: Rathenau Instituut

Labelling of harmful content is done by hand as automatic labelling is unreliable

WHAT IS THE PROBLEM

- It can lead to hate crime (the online, virtual world is connected to the offline real world)
- It means that people are not free and safe any more. “It is a menace to democratic values and peace” (UN strategy plan)
- It is moving into the mainstream, i.e. it not an isolated phenomenon
- In some countries certain forms are illegal, but in many cases there is no legislation at all <-> freedom of speech

ACTION POINTS UNITED NATIONS ON HATE SPEECH

- Monitoring and analysing hate speech
- Addressing causes, drivers and actors (raising awareness)
- Engaging the victims (show solidarity, ensure rights)
- Engaging with new and traditional media (define values of tolerance)
- Using education
- ***Using technology*** , (*but existing technology is unreliable*)

https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech

ISSUES LEARNT FROM SUBJECTIVITY MINING

- **Lack of proper terminology and definitions**
- Lack of a common and unified annotation schema
- Task is difficult for annotators
- Diversity of text genres , topics, and domains (Problem of data collection)
- Social media language and elusive use of language : rapidly changing, diverse, implicit, use of slurs
- **Context of technology (who will use it for what purpose) // explainable AI

Is existing technology unreliable .... Why?

TERMINOLOGY

- Harmful language
- Toxic language
- Offensive language
- Hate speech
- Abusive language
- Aggressive language
- Discriminatory language

DEFINITIONS OF HATE SPEECH FROM DIFFERENT DOMAINS

- Any speech that causes some offence to **others** (Lewis, 2012)
- Any speech that is disparaging of certain gender , religion, race, and sexual orientation (Lewis., 2012)
- Any speech that disparages a **target group of people** based on such characteristics such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic (Nockelby, 2000. -
- all forms of expressions which spread, incite, promote or justify r*acial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism,* discrimination and hostility against **minorities, migrants and people of immigrant origin** (Council of Europe)
- Hate speech is defined as bias-motivated, hostile, malicious speech **aimed at a person or a group of people** because of some of their actual or perceived innate characteristics. It expresses discriminatory, intimidating, disapproving, antagonistic, and/or prejudicial attitudes towards those characteristics, which include gender, race, religion, ethnicity, color, national origin, disability or sexual orientation. Hate speech is intended to i*njure, dehumanize, harass, intimidate, debase, degrade* and victimize the targeted groups, and to foment insensitivity and brutality against them. (Cohen-Almagor)
- Verbal aggression could be understood as any kind of linguistic behaviour which intends to **damage the social identity of the target person** and lower their status and prestige (Barron and Richardson, 1994

=> [1] offensive and harmful language , [2] aimed at persons or a group of people, [3[because of aspects of their identity

UNITED NATIONS DEFINITION

UNITED NATIONS **STRATEGY AND PLAN OF ACTION ON HATE SPEECH**

What is hate speech?

There is no international legal definition of hate speech, and the characterization of what is 'hateful' is controversial and disputed. In the context of this document, the term hate speech is understood as any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor. This is often rooted in, and generates intolerance and hatred and, in certain contexts, can be demeaning and divisive.

Legal implications?

With these considerations in mind, courts in the United States have found that expression generally cannot be punished based on its content or viewpoint. Thus, although hate speech, alone, receives constitutional protection, any expression that constitutes a true threat, incitement to imminent lawless action, discriminatory harassment or defamation can be punished by UWM for those reasons.

ISSUES LEARNT FROM SUBJECTIVITY MINING

- Lack of proper terminology and definitions
- **Lack of a common and unified annotation schema**
- Task is difficult for annotators
- Diversity of text genres , topics, and domains (Problem of data collection)
- Social media language and elusive use of language : rapidly changing, diverse, implicit, use of slurs
- **Context of technology (who will use it for what purpose) // explainable AI

TOWARDS AN ANNOTATION SCHEMA

- How to turn these notions and/or different definitions in an annotation schema
  - That can be reliably annotated
  - That is clear enough to train an automatic classifier on
  - That addresses the phenomenon properly

BINARY/TERNARY ANNOTATION SCHEMES

Definition: a deliberate attack , directed towards a specific group of people; motivated by actual or perceived aspects that form the group's identity

Unit of annotation: document (tweet, comment, sentence of forum post, etc.)

- Davidson et al (2017) Automated Hate Speech Detection and the Problem of Offensive Language. Hate/Offensive/Neither

- Gibert et al. (2018) Internet forum posts: Sentences extracted from Stormfront.org, a white supremacist forum. (largest online community of white nationalists) (hate vs. No-hate)

HIERARCHICAL ANNOTATION SCHEMES

Definition: a deliberate attack , directed towards a specific group of people; motivated by actual or perceived aspects that form the group's identity

Unit of annotation: document (tweet, comment, sentence of forum post, etc.)

- Zampieri et al. (2019) Predicting the Type and Target of Offensive Posts in Social Media (hierarchical annotation schena)

Tweet	A	B	C
@USER He is so generous with his offers.	NOT	—	—
IM FREEEEE!!!! WORST EXPERIENCE OF MY FUCKING LIFE	OFF	UNT	—
@USER Fuk this fat cock sucker	OFF	TIN	IND
@USER Figures! What is wrong with these idiots? Thank God for @USER	OFF	TIN	GRP

Table 1: Four tweets from the dataset, with their labels for each level of the annotation schema.

MORE COMPLEX TAXONOMIES

Primary	Secondary	Example
Identity-directed	Derogation	Muslims cant speak English, they're savages
Identity-directed	Animosity	I dont think black people face any discrimination
Identity-directed	Threatening	Gotta kick those immigrants out... now!
Identity-directed	Glorification	Adolf had it right, yknow?
Identity-directed	Dehumanization	Travellers are nothing but trash
Affiliation-directed	Derogation	So sick of these undesirable leftist fools
Person-directed	To them	Fuck off @user
Person-directed	About them	Trump is a massive bellend
Non-hateful Slurs	/	Y'all see me and my n*ggas come in here?
Counter Speech	Identity-directed	Sorry but that's just factually incorrect
Counter Speech	Affiliation-directed	You should be nicer about the mods, they do alot for us
Counter Speech	Person-directed	That's no way to talk to someone!
Neutral	/	I've had a right bloody day of it

Table 1: Indicative examples of the categories.

Vidgen et al. (2021) Introducing CAD : the Contextual abuse dataset

Task A: Misogyny	Train	Test	Task B: Category	Train	Test	Task B: Target	Train	Test
Non-misogynous	2215	540	0	2215	540	0	2215	540
Misogynous	1785	460	Discredit	1014	141	Active	1058	401
			Sexual harassment	352	44	Passive	727	59
			Stereotype	179	140			
			Dominance	148	124			
			Derailing	92	11			

Table 1: Distribution of tweets in the dataset

Alhuwalia et al. (2020) Hate speech against women in English tweets

MORE INFORMATION: TARGET GROUPS

Text	Dad should have told the muzrat whore to fuck off, and went in anyway
Label	Hate
Targets	Islam
Text	A nigress too dumb to fuck has a scant chance of understanding anything beyond the size of a dick
Label	Hate
Targets	Women, African
Text	Twitter is full of tween dikes who think they're superior because of "muh oppression." News flash: No one gives a shit.
Label	Offensive
Targets	Gay

Table 5: Examples from our dataset. The highlighted portion of the text represents the annotator's rationale.

Target groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic
Religion	Buddhism, Christian, Hindu, Islam, Jewish
Gender	Men, Women
Sexual Orientation	Heterosexual, Gay
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others

Table 3: Target groups considered for the annotation.

Mathew et al. (2021) HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection

TOXIC SPAN DETECTION TASK (SEMEVAL-2021): RATIONALES

OFFSETS	POST
{}	I will stand by my statement that liberalism is what smart looks like to stupid people. The letter posted yesterday states the real problem clearly enough. "It doesn't matter if it was protected by the First Amendment." I have legal news flash for the University of Oregon. Because the University is public and receives government tax dollars it is indeed bound to abide by the US constitution with All OF IT'S PROTECTION INCLUDED. Which should tell everyone right off the signatories are anti-American, and perhaps even anti-America in their twisted PC nonsense.
{}	Fits the despicable piece of trash like a glove.
{}	No he shot someone so he shouldn't be free. He should be locked up like He is for the next decade. You holla like its no big deal but what if it was your little brother who was shot? Then your dumb ass wouldnrlt be screaming free ace g or whatever gay nickname he has. That is the lamest gangster name ever. Do you think that name scares people? Ace clown is what i will now call him.
{55, 56, 57, 58}	Time to give the opposition weapons and overthrow this jerk.
{31, . . . , 39}	you are a paid liberal hack of stupidity shore living the liberal lie every day
{9, . . . , 13}	Youre an idiot.
{7, 8, . . . , 11, 14, 15, . . . , 18}	what a dummy, dummy has no regard for the law. you cannot brutalize a suspect. he has complete lack of respect for any law and is acting like a dictator. he is trying to emulate putin.
{12, . . . , 17, 94, . . . , 102}	People make stupid decisions and then expect the gov't to bail them out. There is no cure for stupidity.
{14, . . . , 20, 29, . . . , 35}	Nah, the only asshole is the asshole firing a rifle within city limits.

Table 2: Examples of toxic test posts and their ground truth toxic spans (shown in red). The left column shows the character offsets of the toxic spans. The top three posts have no toxic spans, the next three have one each, while the remaining three posts have two toxic spans each.

Pavlopoulos et al.(2021) SemEval-2021 Task 5: Toxic Spans Detection (https://aclanthology.org/2021.semeval-1.6.pdf)

ANNOTATION TASK IS DIFFICULT

- Who is the best annotator: crowd, experts, ..
- What background knowledge do you need
- Age, gender, ...?
- How many annotators

ISSUES LEARNT FROM SUBJECTIVITY MINING

- Lack of proper terminology and definitions
- Lack of a common and unified annotation schema
- Task is difficult for annotators
- `**Problem of data collection:** Diversity of text genres , topics, and domains
- Social media language and elusive use of language : rapidly changing, diverse, implicit, use of slurs
- **Context of technology (who will use it for what purpose) // explainable AI

DATASETS

- Overview of datasets on hate speech https://hatespeechdata.com/
- > 120 datasets (2017-2021/22)
- > 30 languages
- Platforms: YouTube, instagram, twitter*, Facebook, reddit, newspaper comments, Gab, Stormfront, Civil comments, Wikipedia comments
- Target groups: Women, Trans people, Black people, Gay people, Disabled people, Muslims, Immigrants
- Cause of hate: Gender, Sexual orientation, Religion, Disability

Vidgen et al. (2021) Directions in Abusive Language Training Data: Garbage In, Garbage Out.

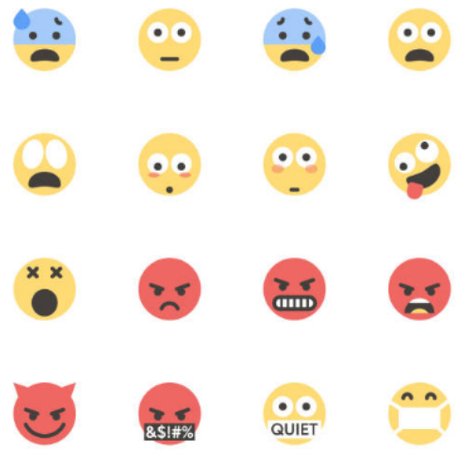
Poletto et al. (2020) Resources and benchmark corpora for hate speech detection: a systematic review'

DATASETS

- How to collect them?
- Not all data is available
  - Keyword based
  - Topic/blog based
- Not all data is freely available
  - +Wikipedia comments, news comments, tweets (not to share), Reddit, gab
  - - FB, news comments

ISSUES LEARNT FROM SUBJECTIVITY MINING

- Lack of proper terminology and definitions
- Lack of a common and unified annotation schema
- Task is difficult for annotators
- `Problem of data collection: Diversity of text genres , topics, and domains
- **Social media language and elusive use of language : rapidly changing, diverse, implicit, use of slurs**
- **Context of technology (who will use it for what purpose) // explainable AI



LANGUAGE CHARACTERISTICS OF HATE SPEECH

- Rapidly changing: Jews and n*ggers destroy & pervert everything they touch #jewfail #niggerfail
- Social media characterics: ungrammatical, hashtags, emoji's
- Elusive language:

language processing techniques, according to r/India moderators. Muslims are referred to using coded language such as "Abduls," "Mull@s," "K2as," or, derisively, "Peace loving" people. Christians are referred to as "Xtians"; while Pakistan is called "Porkistan."

ISSUES LEARNT FROM SUBJECTIVITY MINING

- Lack of proper terminology and definitions
- Lack of a common and unified annotation schema
- Task is difficult for annotators
- `Problem of data collection: Diversity of text genres , topics, and domains
- Social media language and elusive use of language : rapidly changing, diverse, implicit, use of slurs
- ****Context of technology (who will use it for what purpose) // explainable AI**

METHODS

- Rule and lexicon based
  - Still widely used: no training data needed, limited domain depend, transparent
- Traditional machine learning (SVM, NB, LR, ..)
  - Still widely used: relatively small sets of training data
- Deep learning (CNN, RNN)
  - Word order and syntactic context can be maintained
- Transformer models (BERT, RoBERTa, hateBERT,toxicBERT)
  - High performance but opaque
- Hybrid methods
  - Combining lexicon based approaches with machine learning approaches (e.g. hurtBERT)
- Ensemble method with voting

DASHBOARD

Bunde et al, (2021) AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach

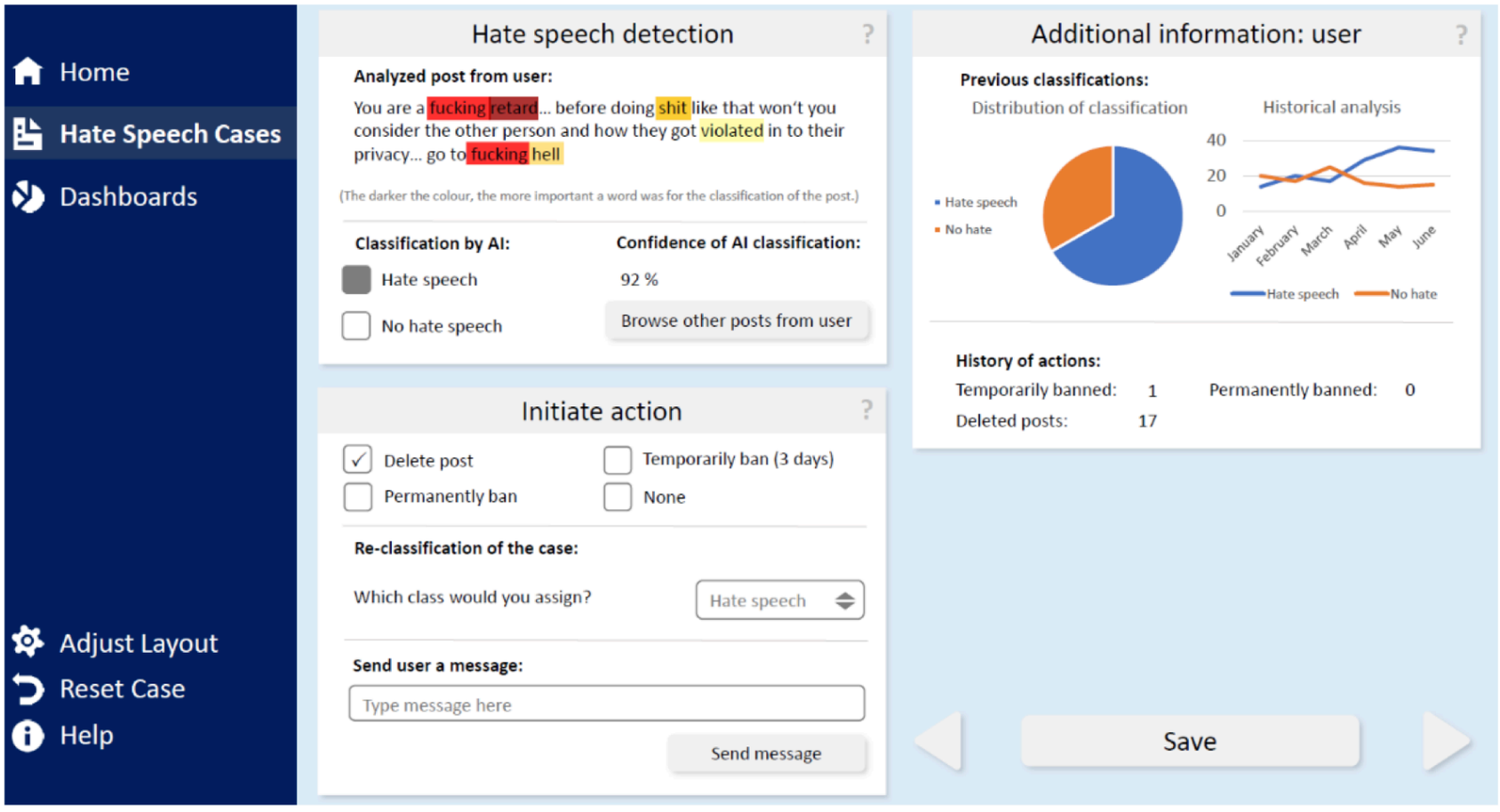


Figure 7. Dashboard with all DFs.

Bunde(2021) https://scholarspace.manoa.hawaii.edu/bitstream/10125/70766/0125.pdf

CONCLUSIONS

- Hate speech is a serious societal problem that needs to be addressed in various ways among which automatic detection.
- Automatic detection has many challenges:
  - The concept of hate speech is not well defined
  - Hate speech on the Internet is like a needle
  - Many different sets of annotation guidelines
  - Availability of (annotated data)
  - Continuously changing forms of hate speech, target groups, topics
  - Many methods are developed
  - High expectations: 100% recall-precision, explainability
- The community (WE!!) is working on solutions

TERM IS DIFFICULT TO DEFINE

- Hateful comments toward a specific group or target (Walker, 1994)
- Language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group
- Schmidt: abusive, hostile, flames, cyberbullying, offensive language, profanity-related offensive content