Case Study: Deontological Ethics in NLP

Shrimai Prabhumoye*, Brendon Boldt*, Ruslan Salakhutdinov, Alan W Black

Carnegie Mellon University

Language Technologies Institute

Ethics

- Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

Ethics

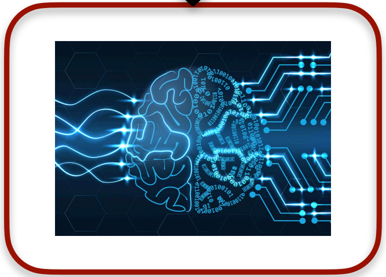
- Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

Large body of work on Ethics

Ethics

- Prior work on understanding and mitigating bias (Hovy & Prabhumoye, 2021; Blodgett et al, 2020; Shah et al, 2020; Sun et al, 2019; Zhao et al, 2019; Tatman, 2017; Bolukbasi et al, 2016)

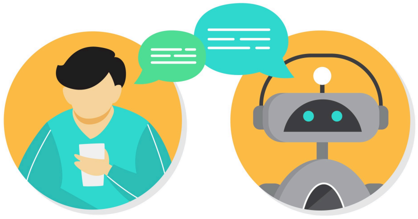
Large body of work on Ethics



How can we apply it to NLP?

Ethics

- Deontological framework for NLP
  - Generalization principle
  - Respect for Autonomy
- Reasonable, clear ethical rules, “rule of law”



Question-Answering

Machine Translation

Detecting objectionable content

Dialogue Systems

Which tasks have important ethical implications?

What factors and methods are preferable in ethically solving this problem?

Generalization Principle

Generalization Principle

An action $\mathscr{A}$ *taken for reasons* $\mathscr{R}$ *is ethical if and only if a world where all people perform* $\mathscr{A}$ *for reasons* $\mathscr{R}$ *is conceivable.*

Generalization Principle

*An action $\mathscr{A}$ taken for reasons $\mathscr{R}$ is **un**ethical if and only if a world where all people perform $\mathscr{A}$ for reasons $\mathscr{R}$ logically contradicts $\mathscr{R}$*

Detecting objectionable content

Detecting objectionable content

$\mathscr{A}$ deploying flagging systems

Detecting objectionable content

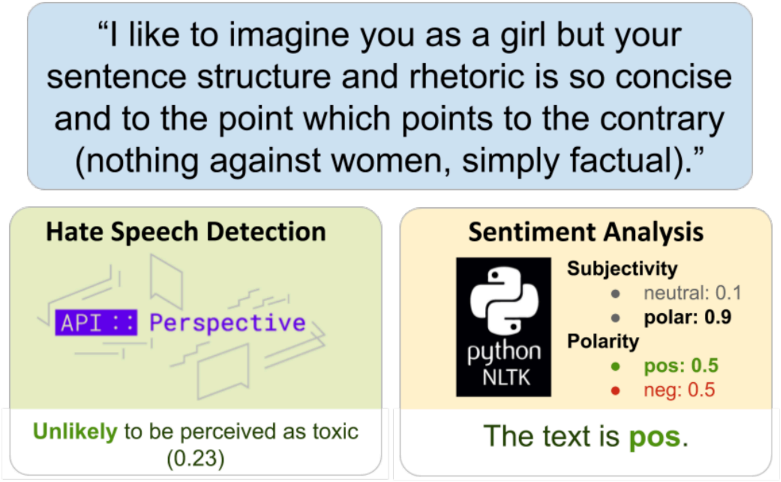
$\mathscr{A}$ deploying flagging systems

$\mathscr{R}$ ↓ burden on humans ↓ # posts that need to be seen by human eyes

Detecting objectionable content

$\mathscr{A}$ deploying flagging systems

$\mathscr{R}$ ↓ burden on humans ↓ # posts that need to be seen by human eyes



"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."

Hate Speech Detection

Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

Subjectivity

- neutral: 0.1
- **polar: 0.9**

Polarity

- **pos: 0.5**
- neg: 0.5

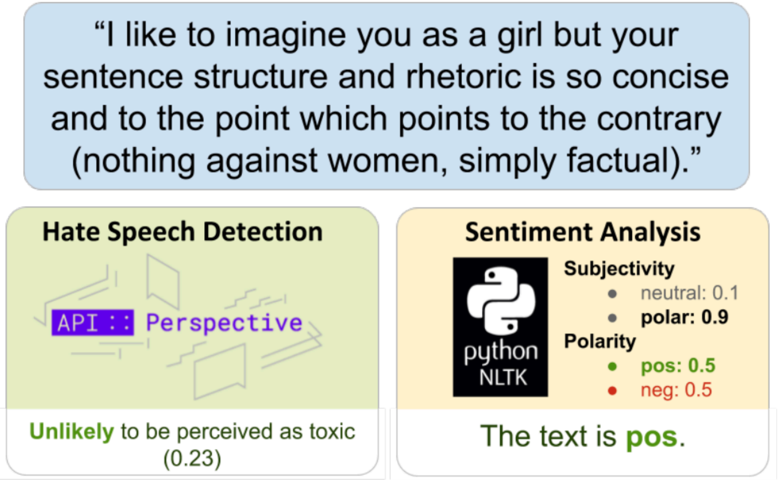
The text is **pos**.

[Breitfeller et al, EMNLP 2019]

Detecting objectionable content

$\mathscr{A}$ deploying flagging systems

$\mathscr{R}$ ↓ burden on humans ↓ # posts that need to be seen by human eyes



"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."

Unlikely to be perceived as toxic (0.23)

The text is **pos**.

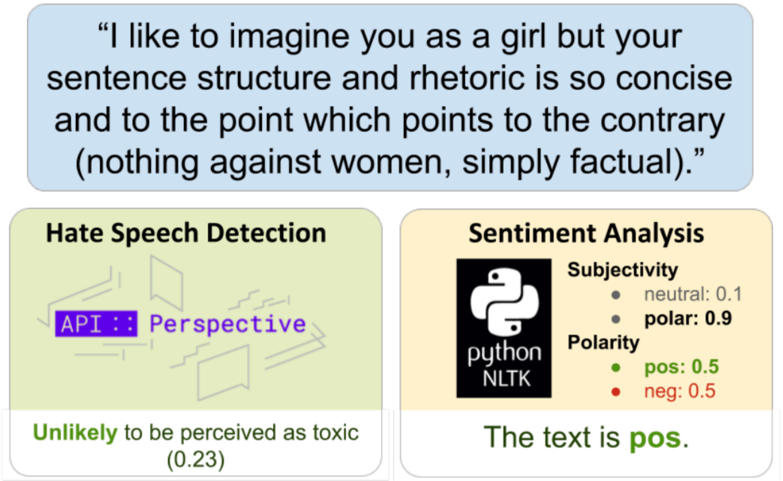
[Breitfeller et al, EMNLP 2019]

- surface level words $\Longrightarrow$ phrase the same meaning with different words

Detecting objectionable content

$\mathscr{A}$ deploying flagging systems

$\mathscr{R}$ ↓ burden on humans ↓ # posts that need to be seen by human eyes



"I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual)."

Hate Speech Detection

Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

Subjectivity
- neutral: 0.1
- **polar: 0.9**

Polarity
- **pos: 0.5**
- neg: 0.5

The text is **pos**.

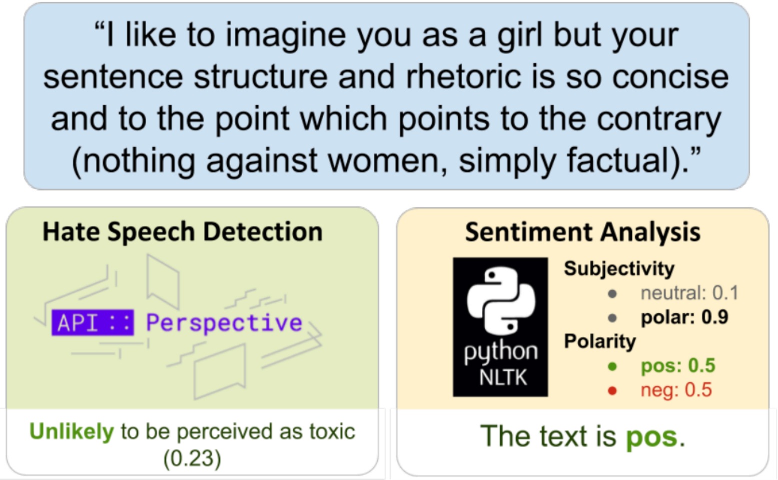
[Breitfeller et al, EMNLP 2019]

- surface level words $\Longrightarrow$ phrase the same meaning with different words
- flagging system will be unsuccessful

Detecting objectionable content

$\mathscr{A}$ deploying flagging systems

$\mathscr{R}$ ↓ burden on humans ↓ # posts that need to be seen by human eyes



“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

Hate Speech Detection

Unlikely to be perceived as toxic (0.23)

Sentiment Analysis

Subjectivity
- neutral: 0.1
- **polar: 0.9**

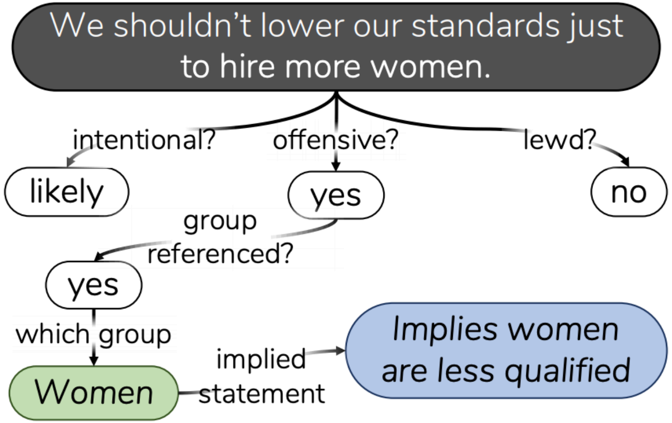
Polarity
- ● **pos: 0.5**
- ● neg: 0.5

The text is **pos.**

[Breitfeller et al, EMNLP 2019]

- surface level words $\Longrightarrow$ phrase the same meaning with different words
- flagging system will be unsuccessful
- logically contradicts the premise

Detecting objectionable content



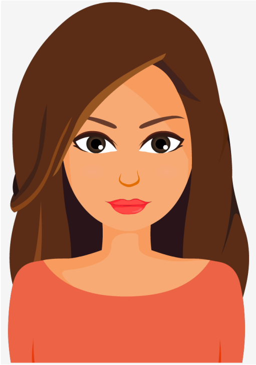
[Sap et al, ACL 2020]

- Underlying intent, offensiveness, and power differentials between different social groups.
- Generate consequences and implications
- Does not lead to an arms race between objection content generation and detection

Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***

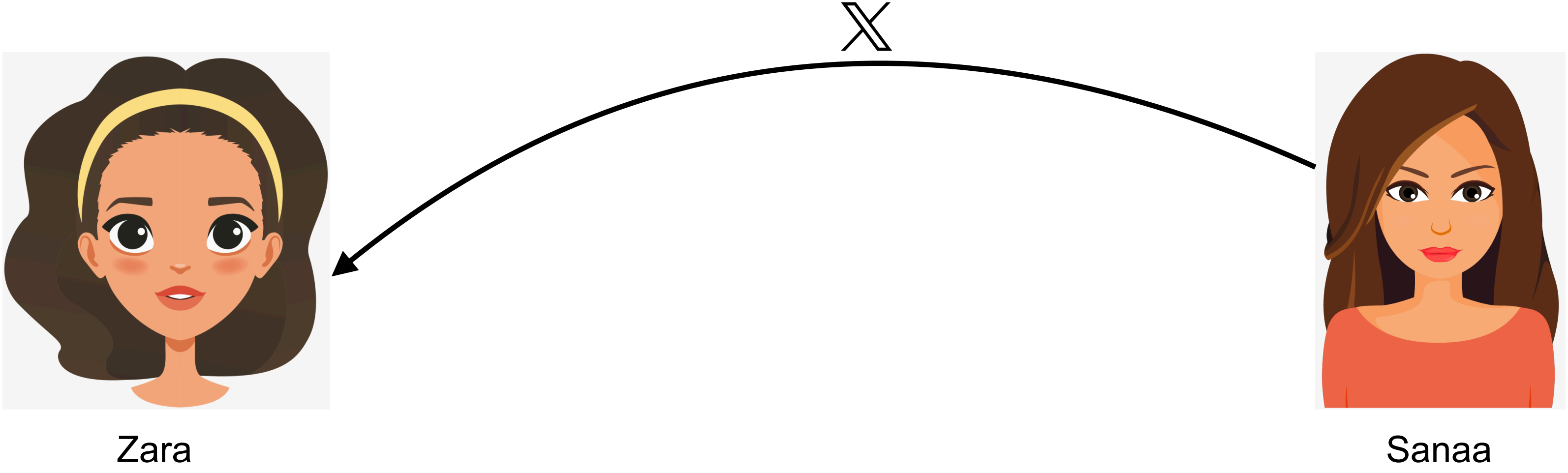
Zara



Sanaa

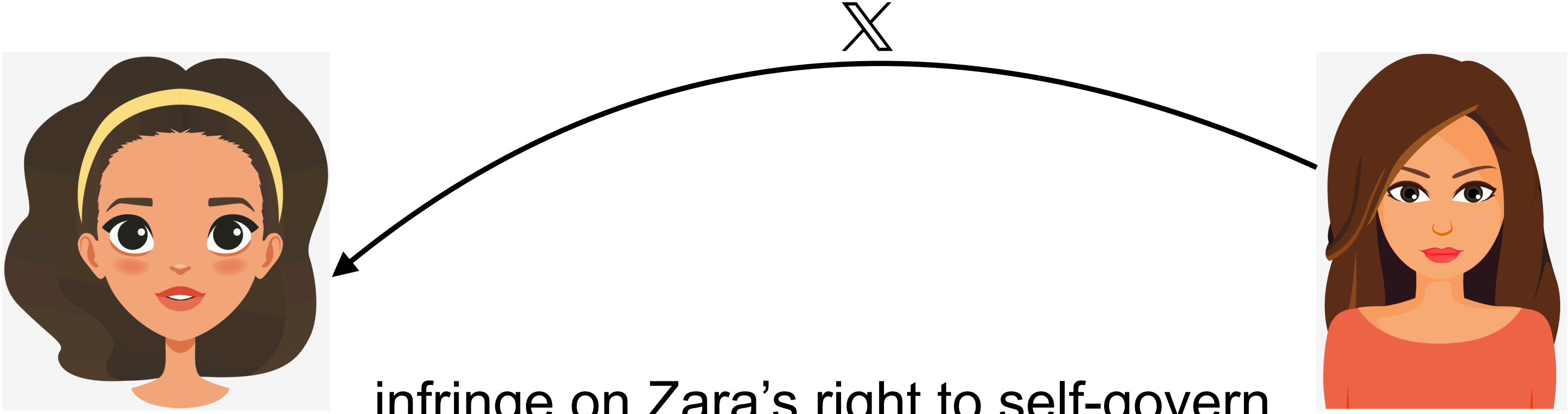
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***



Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***



𝕏

infringe on Zara’s right to self-govern

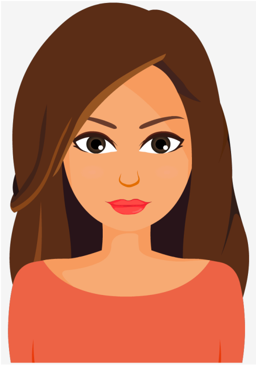
Zara

Sanaa

Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***

Zara



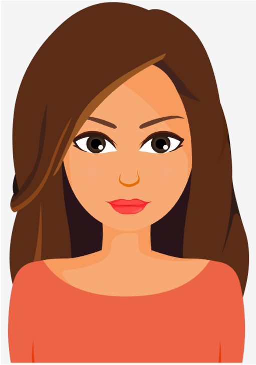
Sanaa

Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***

Zara

Zara must be sufficiently informed about $\mathbb{X}$



Sanaa

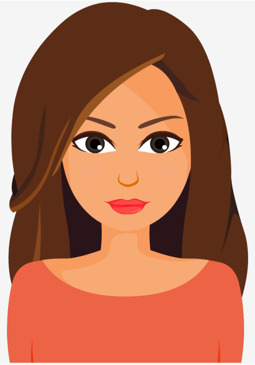
Respect for Autonomy

- Addresses the right of a person to make decisions which directly pertain to themselves.
- ***Informed consent***

Zara

Zara must be sufficiently informed about $\mathbb{X}$

Zara *herself* makes the decision to allow Sanaa to do $\mathbb{X}$

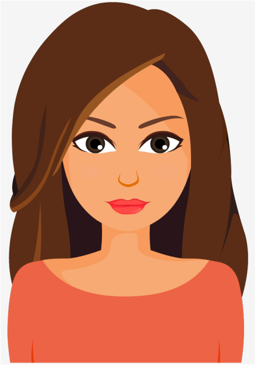


Sanaa

Translation

Translator

Zara



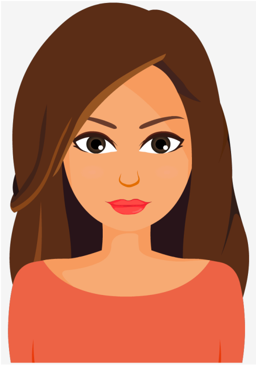
Sanaa

Translation

Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

Translator

Zara



Sanaa

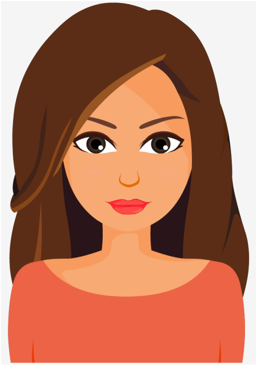
Translation

Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

Translator

Zara

There might be a formal contract of how Sanaa is to act



Sanaa

Translation

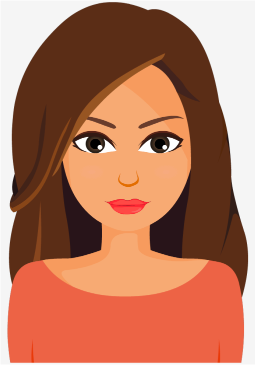
Zara consents to Sanaa serving as an *ad hoc* representative for what she would like to say.

Translator

Zara

There might be a formal contract of how Sanaa is to act

Zara relies on Sanaa's paralinguistic conduct



Sanaa

Machine Translation

Zara

Translator

Machine Translation

Machine Translation

MT system is speaking for Zara

Translator

Zara

Machine Translation

Machine Translation

MT system is speaking for Zara

Zara must be ***informed*** of ambiguities so that she can ***consent*** to the message which the system ultimately conveys.

Zara

Translator

Machine Translation

Machine Translation

MT system is speaking for Zara

Zara must be ***informed*** of ambiguities so that she can ***consent*** to the message which the system ultimately conveys.

Zara must also be ***informed*** of the failure cases in the MT system.

Zara

Translator

Machine Translation

Machine Translation

Ms. Hashimoto …

Zara

Translator

English to Japanese Machine Translation

-san? or
-sensei?…

Machine Translation

Zara must be notified that such an ambiguity needs to be resolved because there is a risk of offending the Japanese speaker.

Ms. Hashimoto …

Zara

Translator

English to Japanese Machine Translation

-san? or
-sensei?…

Machine Translation

My *aunt* is coming home tomorrow.

Zara

Translator

English to Hindi Machine Translation

Is it maternal or paternal aunt? They have different words in Hindi.

Aadil

Machine Translation

MT system can ask a follow up question to Zara.

My *aunt* is coming home tomorrow.

Zara

Translator

English to Hindi Machine Translation

Is it maternal or paternal aunt? They have different words in Hindi.

Aadil

Machine Translation

Zara

My *aunt* is coming home tomorrow.

Translator

English to Hindi Machine Translation

I am unable to translate the sentence in its current form. Can you please rephrase it?

Aadil

NLP methods for Ethics

NLP methods for Ethics

Machine Translation: understand social context, control formality, politeness, author attributes, voice

NLP methods for Ethics

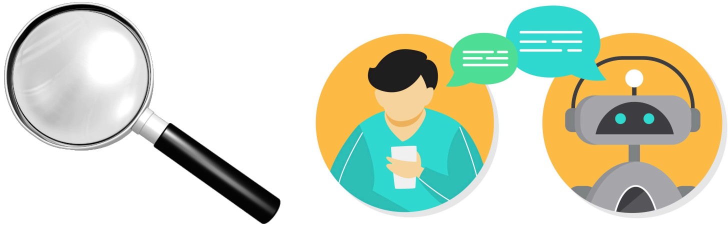
Machine Translation: understand social context, control formality, politeness, author attributes, voice

Detecting objectionable content: generate consequences and implications

NLP methods for Ethics

Machine Translation: understand social context, control formality, politeness, author attributes, voice

Detecting objectionable content: generate consequences and implications

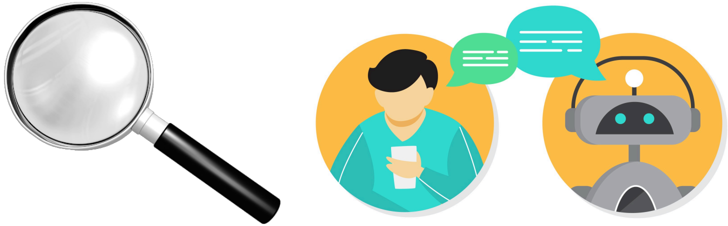


Question-Answering: transparency, dynamic graph generation for answers

NLP methods for Ethics

Machine Translation: understand social context, control formality, politeness, author attributes, voice

Detecting objectionable content: generate consequences and implications

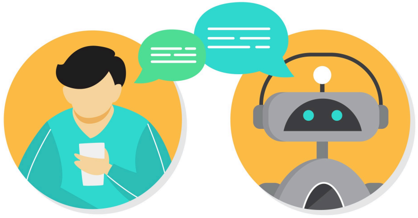


Question-Answering: transparency, dynamic graph generation for answers

Dialogue Systems: control topics, style, content, persona

Summary

- Deontological framework for NLP
  - Generalization principle
  - Respect for Autonomy
- Four case studies
- Discussion



Question-Answering

Machine Translation

Detecting objectionable content

Dialogue Systems