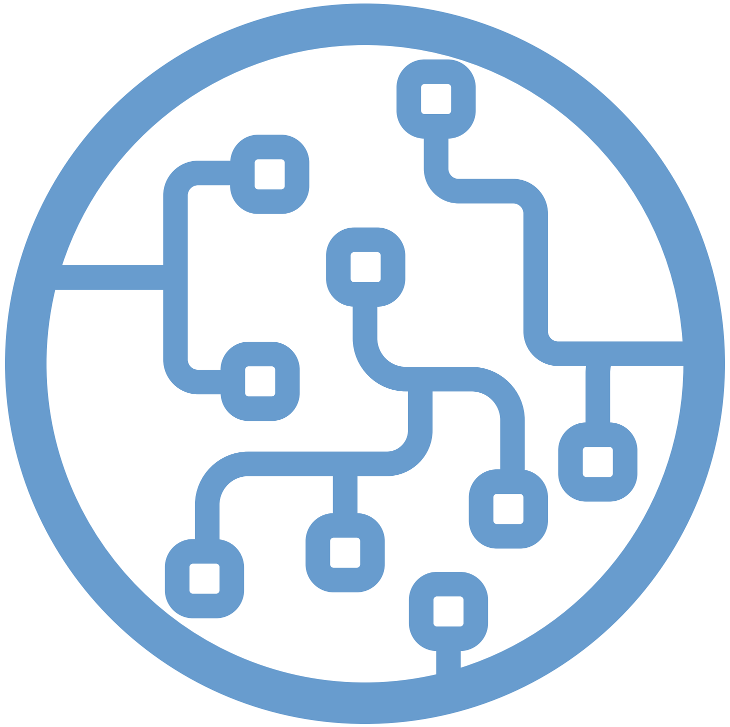
How modeling choices can inform data collection and annotation

Jana Lasser
Research group Complex Social &
Computational Systems

The interdisciplinary digital
lab of the University of Graz

Outline

- A brief primer on supervised learning
- Model choices for text classification
- Wishful thinking and data reality



"Artificial Intelligence" by Oksana Latysheva from NounProject.com, CC BY 3.0 Deed

A brief primer on supervised learning for text

Classification

Let's assume we want to answer the following research question:

How did the prevalence of hate speech on X change over the last decade?

Classification

Let's assume we want to answer the following research question:

How did the prevalence of hate speech on X change over the last decade?

some example tweets	hate?
Women aren’t capable of doing politics... see Merkel	
My god, just SHUT UP! You are not contributing anything useful to this conversation!	
I please want to know where the photo with the crying man comes from an if poss. who it depicts. Thanks upfront :)	
Politics should rather concentrate on getting rid of all those multicultural people again.	
You are unbelievably dumb. All Muslims are filth and should leave immediately.	

Adapted from example tweets in Herderich & Lasser et al., Collective moderation of hate, toxicity, and extremity in online discussions, *arXiv* (2024).

Hate speech: *insults, discrimination, or intimidation, spreading fearful, negative, and harmful stereotypes, calling for exclusion or segregation, inciting hatred, and encouraging violence against individuals or groups on the grounds of their supposed race, ethnic origin, gender, religion, or political beliefs.*

Classification

Let's assume we want to answer the following research question:

How did the prevalence of hate speech on X change over the last decade?

some example tweets	hate?
Women aren’t capable of doing politics… see Merkel	yes
My god, just SHUT UP! You are not contributing anything useful to this conversation!	no
I please want to know where the photo with the crying man comes from an if poss. who it depicts. Thanks upfront :)	no
Politics should rather concentrate on getting rid of all those multicultural people again.	
You are unbelievably dumb. All Muslims are filth and should leave immediately.	

Adapted from example tweets in Herderich & Lasser et al., Collective moderation of hate, toxicity, and extremity in online discussions, *arXiv* (2024).

Hate speech: *insults, discrimination, or intimidation, spreading fearful, negative, and harmful stereotypes, calling for exclusion or segregation, inciting hatred, and encouraging violence against individuals or groups on the grounds of their supposed race, ethnic origin, gender, religion, or political beliefs.*

Classification

Let's assume we want to answer the following research question:

How did the prevalence of hate speech on X change over the last decade?

some example tweets	hate?
Women aren't capable of doing politics... see Merkel	yes
My god, just SHUT UP! You are not contributing anything useful to this conversation!	no
I please want to know where the photo with the crying man comes from an if poss. who it depicts. Thanks upfront :)	no
Politics should rather concentrate on getting rid of all those multicultural people again.	
You are unbelievably dumb. All Muslims are filth and should leave immediately.	

Adapted from example tweets in Herderich & Lasser et al., Collective moderation of hate, toxicity, and extremity in online discussions, *arXiv* (2024).

What do we have here?

- A set of data points x
- For some data points a class label $c \in \{yes, no\}$

Classification

Let's assume we want to answer the following research question:

How did the prevalence of hate speech on X change over the last decade?

some example tweets	hate?
Women aren’t capable of doing politics... see Merkel	yes
My god, just SHUT UP! You are not contributing anything useful to this conversation!	no
I please want to know where the photo with the crying man comes from an if poss. who it depicts. Thanks upfront :)	no
Politics should rather concentrate on getting rid of all those multicultural people again.	
You are unbelievably dumb. All Muslims are filth and should leave immediately.	

Adapted from example tweets in Herderich & Lasser et al., Collective moderation of hate, toxicity, and extremity in online discussions, *arXiv* (2024).

What do we have here?

- A set of data points *x*
- For some data points a class label $c \in \{yes, no\}$

Goal: A classifier that maps data points without labels to classes, based on examples.

Classification

Let's assume we want to answer the following research question:
How did the prevalence of hate speech on X change over the last decade?

some example tweets	hate?
Women aren’t capable of doing politics... see Merkel	yes
My god, just SHUT UP! You are not contributing anything useful to this conversation!	no
I please want to know where the photo with the crying man comes from an if poss. who it depicts. Thanks upfront :)	no
Politics should rather concentrate on getting rid of all those multicultural people again.	
You are unbelievably dumb. All Muslims are filth and should leave immediately.	

Adapted from example tweets in Herderich & Lasser et al., Collective moderation of hate, toxicity, and extremity in online discussions, *arXiv* (2024).

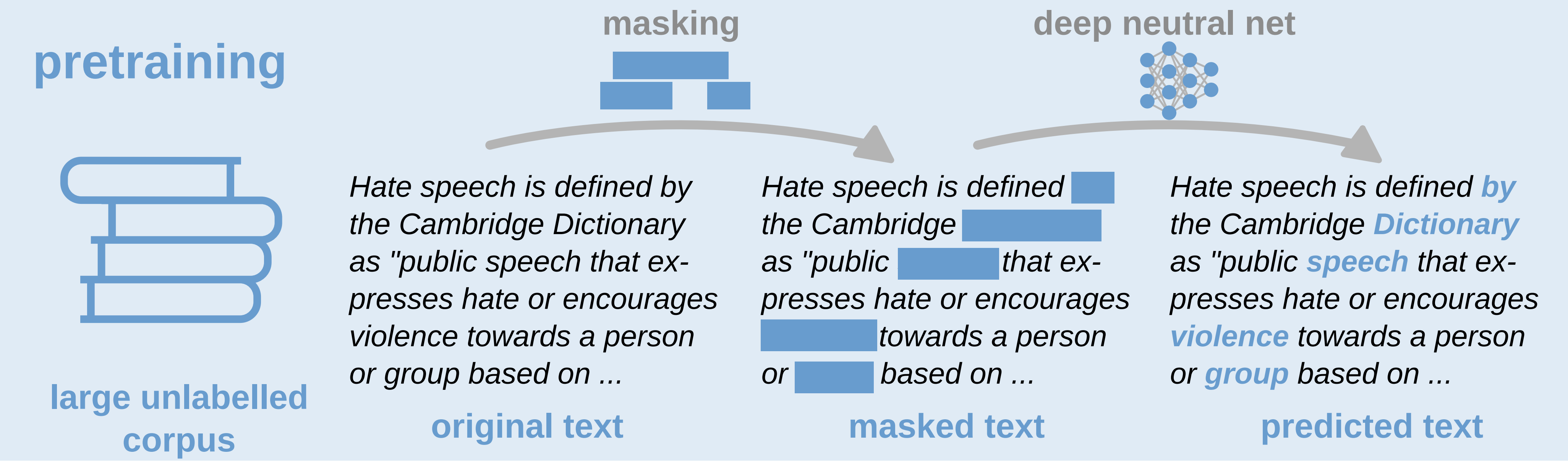
What do we have here?

- A set of data points *x*
- For some data points a class label *c* ∈ *{yes, no}*

supervised learning

Goal: A classifier that maps data points without labels to classes, based on examples.

Training a classifier for text



pretraining

masking

deep neutral net

large unlabelled corpus

Hate speech is defined by the Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on ...

original text

Hate speech is defined ▇ the Cambridge ▇ as "public ▇ that expresses hate or encourages ▇ towards a person or ▇ based on ...

masked text

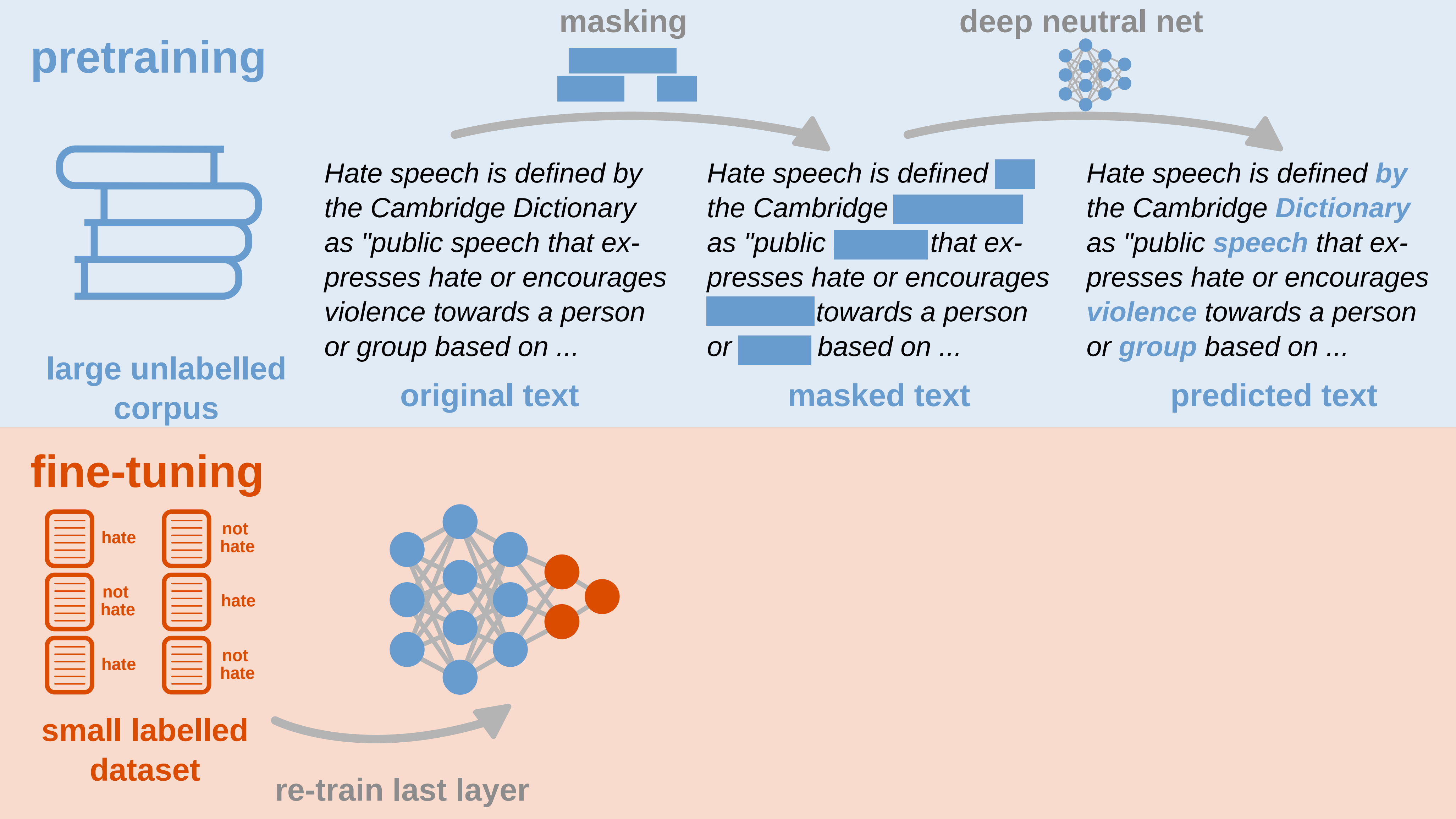
Hate speech is defined ***by*** *the Cambridge* ***Dictionary*** *as "public* ***speech*** *that expresses hate or encourages* ***violence*** *towards a person or* ***group*** *based on ...*

predicted text

transformer model
can "speak" the language

huggingface
model hub

Training a classifier for text



pretraining

masking

deep neutral net

Hate speech is defined by the Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on ...

original text

Hate speech is defined ▇ the Cambridge ▇ as "public ▇ that expresses hate or encourages ▇ towards a person or ▇ based on ...

masked text

Hate speech is defined ***by*** *the Cambridge* ***Dictionary*** *as "public* ***speech*** *that expresses hate or encourages* ***violence*** *towards a person or* ***group*** *based on ...*

predicted text

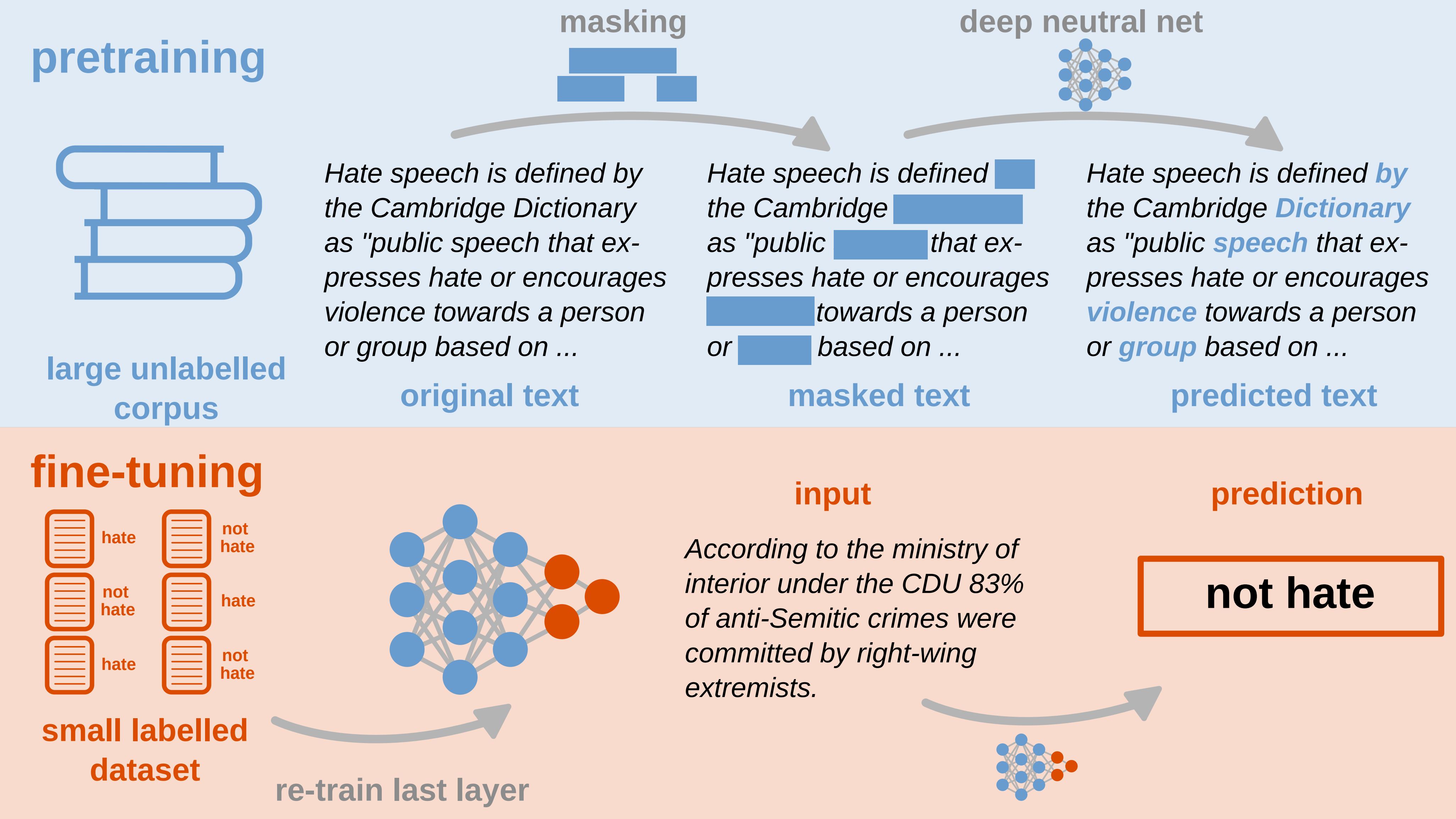
large unlabelled corpus

fine-tuning

small labelled dataset

re-train last layer

Training a classifier for text



pretraining

masking

deep neutral net

large unlabelled corpus

Hate speech is defined by the Cambridge Dictionary as "public speech that expresses hate or encourages violence towards a person or group based on ...

original text

Hate speech is defined █ the Cambridge █ as "public █ that expresses hate or encourages █ towards a person or █ based on ...

masked text

Hate speech is defined ***by*** *the Cambridge* ***Dictionary*** *as "public* ***speech*** *that expresses hate or encourages* ***violence*** *towards a person or* ***group*** *based on ...*

predicted text

fine-tuning

small labelled dataset

re-train last layer

input

According to the ministry of interior under the CDU 83% of anti-Semitic crimes were committed by right-wing extremists.

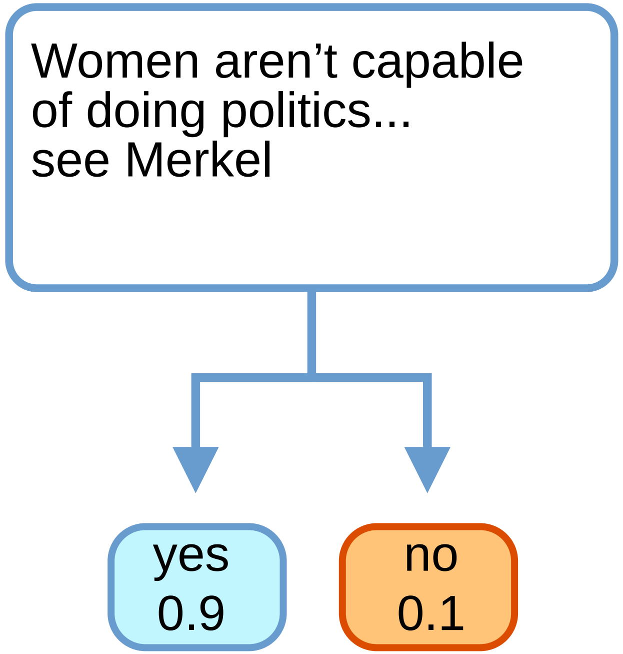
prediction

not hate

Model choices for text classification

Types of classification tasks

Binary Classification



Women aren’t capable of doing politics... see Merkel

yes 0.9

no 0.1

Classes

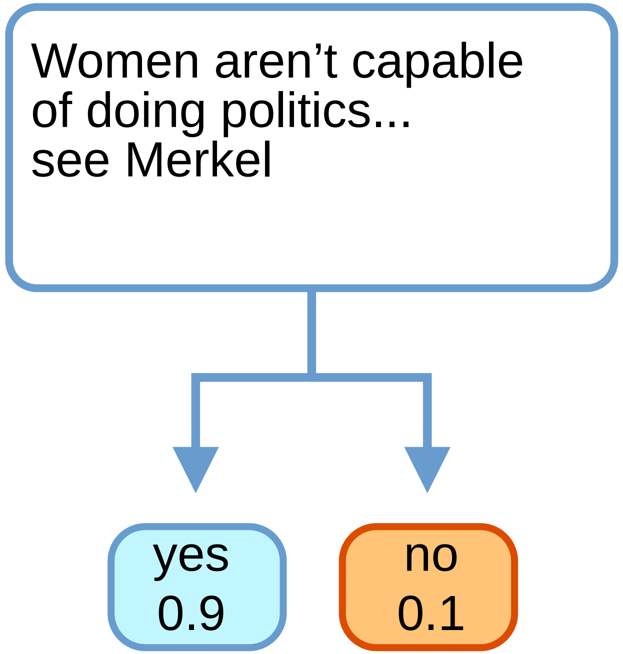
(is this hate speech?)

☑ yes

□ no

Types of classification tasks

Binary Classification



Women aren’t capable
of doing politics...
see Merkel

yes
0.9

no
0.1

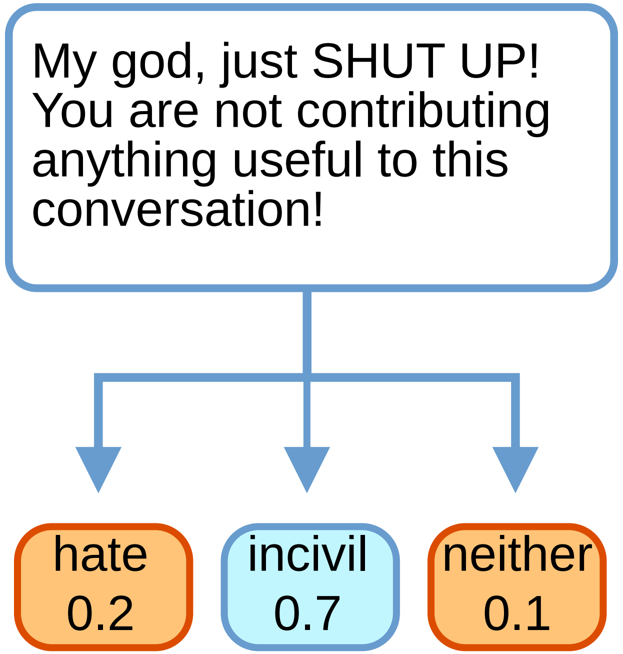
Classes

(is this hate speech?)

☑ yes

□ no

Multiclass Classification



My god, just SHUT UP!
You are not contributing
anything useful to this
conversation!

hate
0.2

incivil
0.7

neither
0.1

Classes

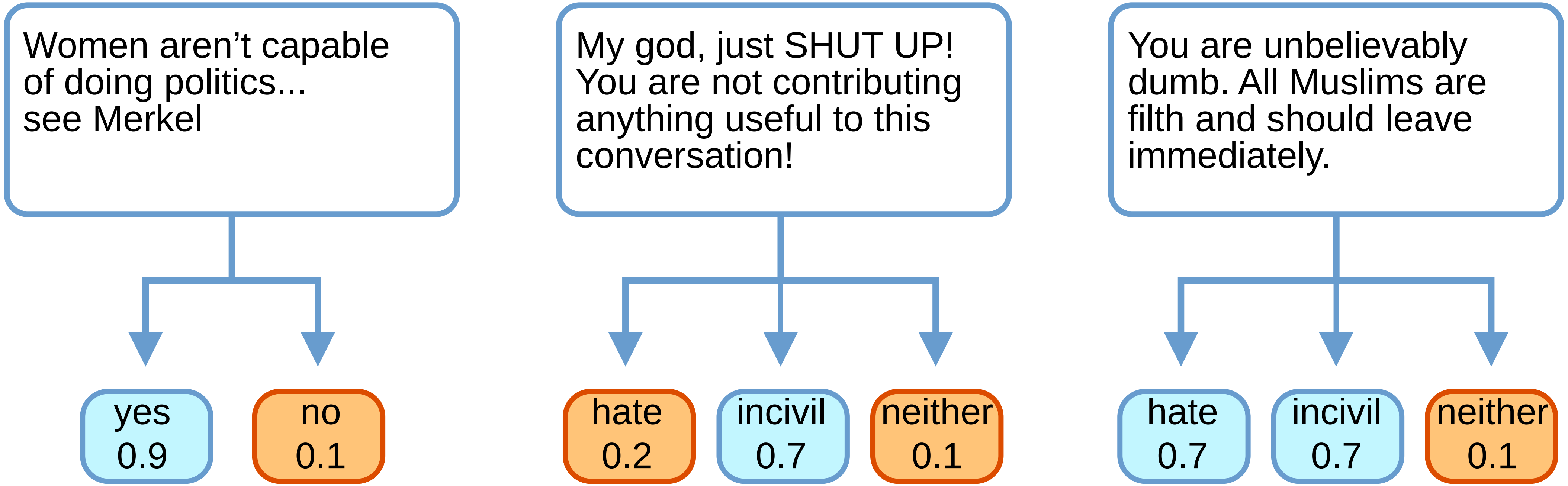
(pick one class)

□ hate speech

☑ incivil language

□ neither

Types of classification tasks



Binary Classification

Women aren’t capable
of doing politics...
see Merkel

yes 0.9

no 0.1

Classes

(is this hate speech?)

☑ yes

□ no

Multiclass Classification

My god, just SHUT UP!
You are not contributing
anything useful to this
conversation!

hate 0.2

incivil 0.7

neither 0.1

Classes

(pick one class)

□ hate speech

☑ incivil language

□ neither

Multilabel Classification

You are unbelievably
dumb. All Muslims are
filth and should leave
immediately.

hate 0.7

incivil 0.7

neither 0.1

Classes

(pick all that apply)

☑ hate speech

☑ incivil language

□ neither

Model choice overview

choice	**consequences**
fine-tuning or few-shot LLM	training data size, biases, control

Model choice overview

choice	consequences
fine-tuning or few-shot LLM	training data size, biases, control
type of classification task	structure of labelling task

Model choice overview

choice	consequences
fine-tuning or few-shot LLM	training data size, biases, control
type of classification task	structure of labelling task
difficulty of classification task	qualifications of annotators, cost

Model choice overview

choice	consequences
fine-tuning or few-shot LLM	training data size, biases, control
type of classification task	structure of labelling task
difficulty of classification task	qualifications of annotators, cost
pretraining corpus	language fit, domain fit

Model choice overview

choice	consequences
fine-tuning or few-shot LLM	training data size, biases, control
type of classification task	structure of labelling task
difficulty of classification task	qualifications of annotators, cost
pretraining corpus	language fit, domain fit
model size (# parameters)	hardware requirements

Wishful thinking and data reality

Class imbalances in real data

Wouldn't it be nice to have much finer-grained labels for classes?

opinion

providing information

pointing out consequences

providing correction

personal insult

insult of protected cat.

other

sarcasm

asking question

pointing out inconsistencies

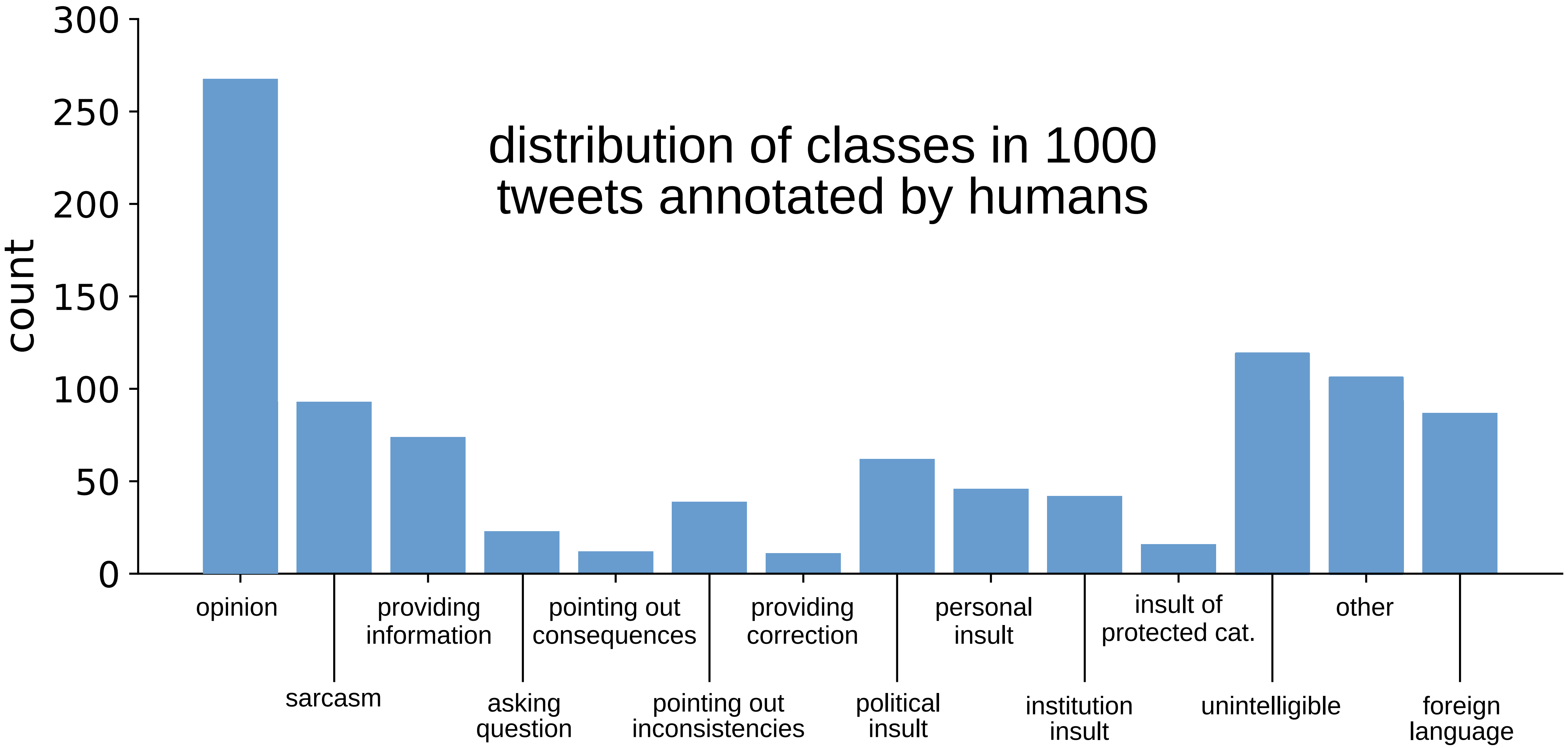
political insult

institution insult

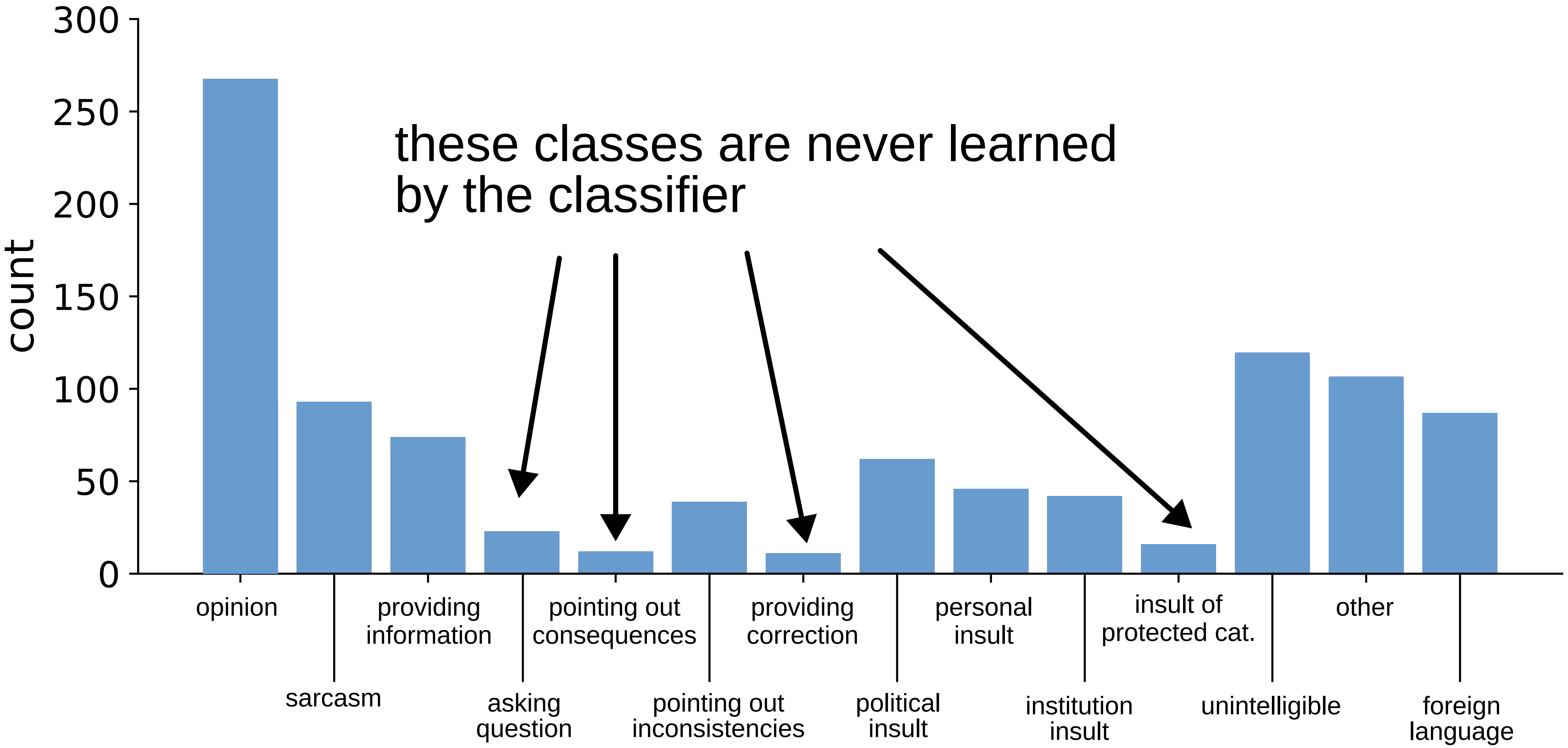
unintelligible

foreign language

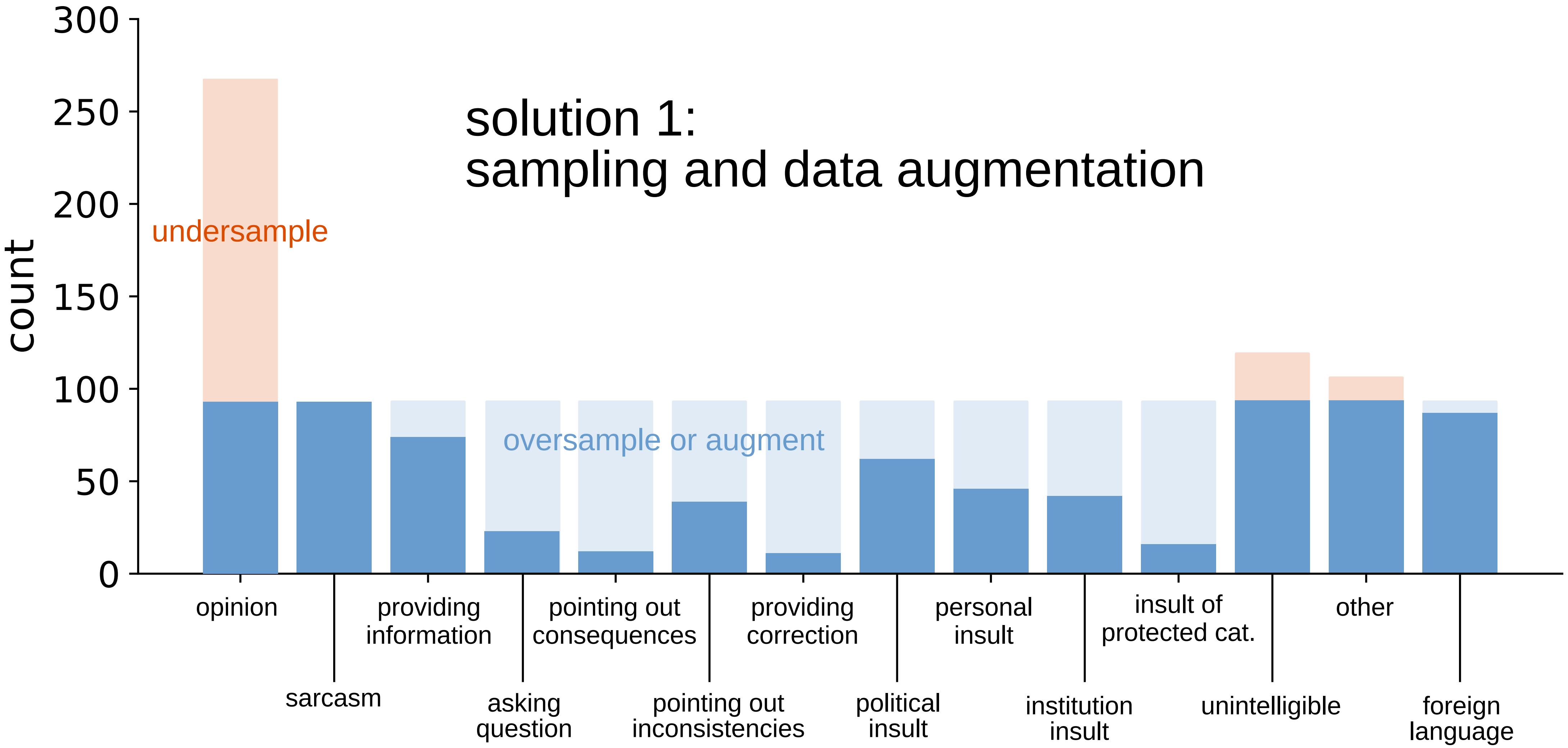
Class imbalances in real data



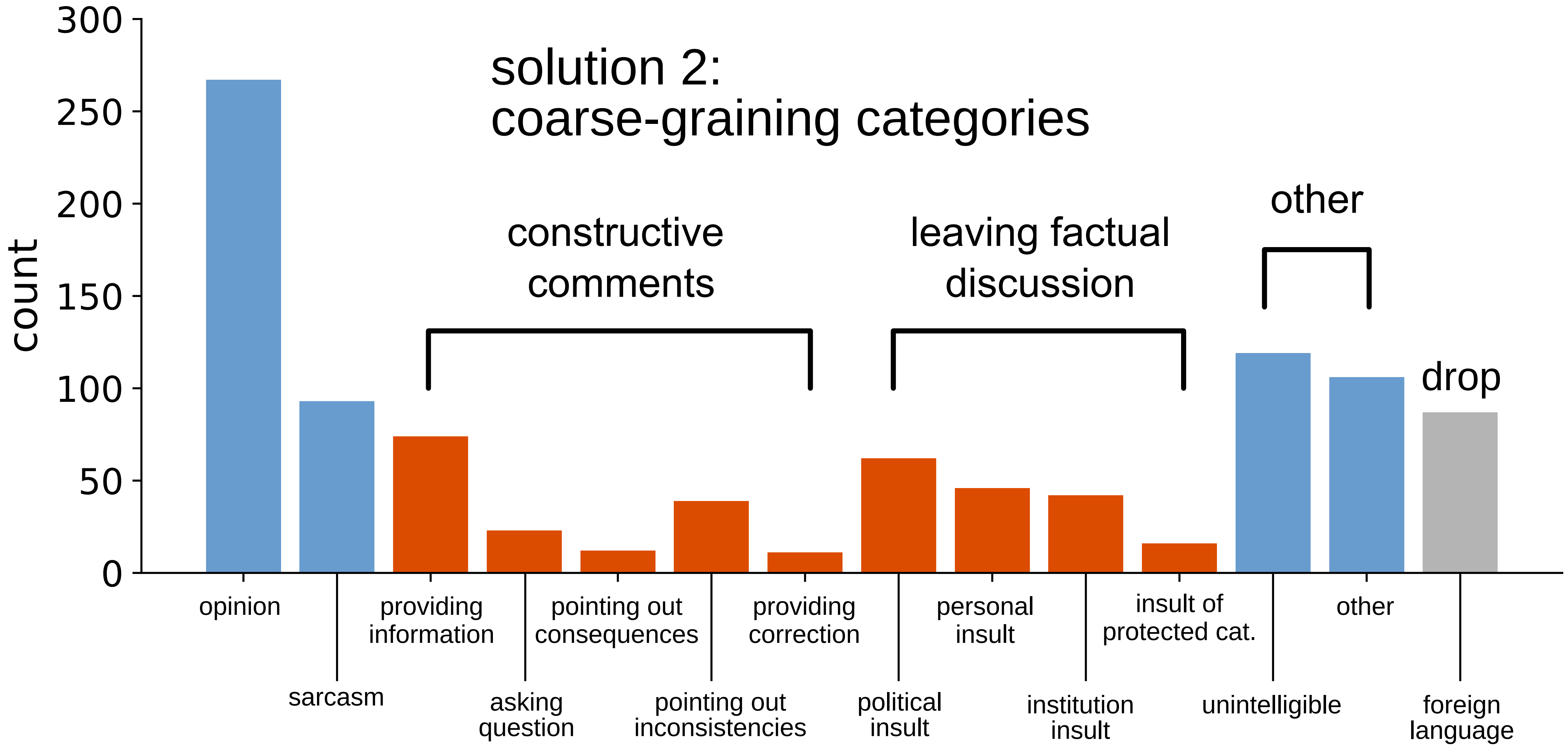
Class imbalances in real data



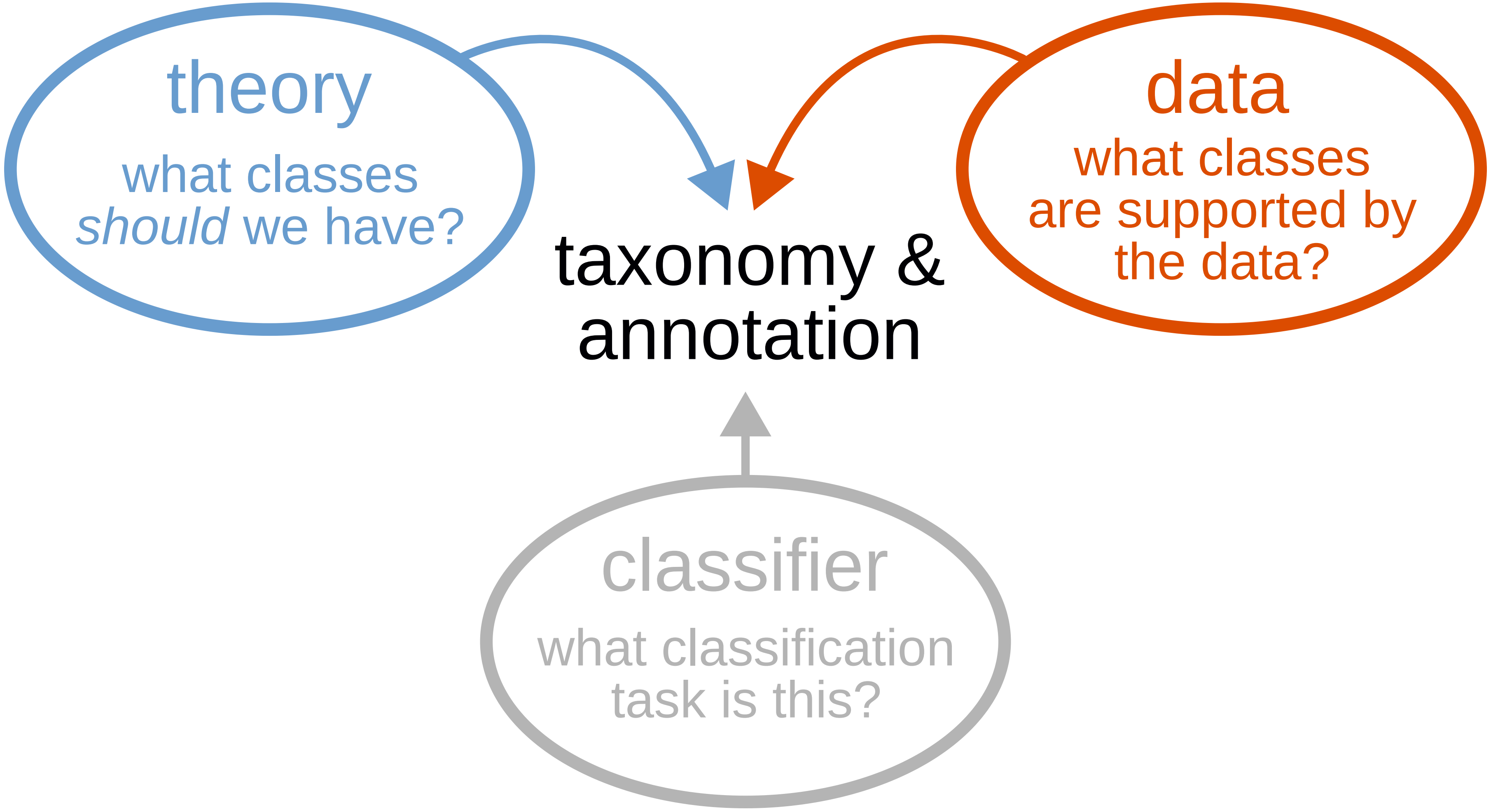
Class imbalances in real data



Class imbalances in real data



Summary



theory

what classes *should* we have?

data

what classes are supported by the data?

taxonomy & annotation

classifier

what classification task is this?