

**Subproject as part of SAIL:**  
***Longitudinal Analysis of Change and Variety of Natural Language Data***

# **Computational linguistic methods for modeling lexical-semantic dynamics of hate speech**

**by Sanne Hoeken**

Supervised by dr. Özge Alaçam and prof. dr. Sina Zarriß

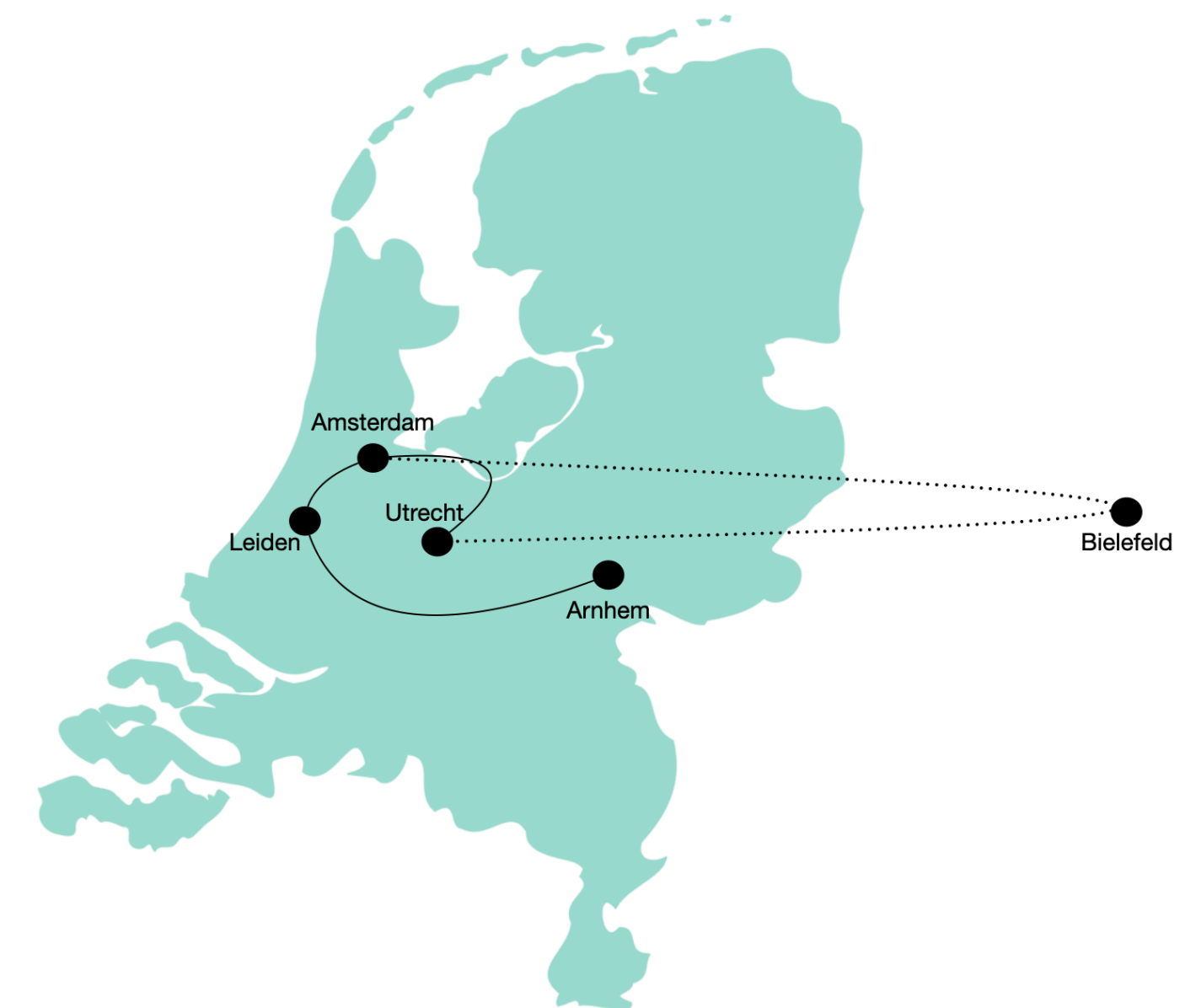
# But first...

## Who am I?

- (3rd year) PhD in Computational Linguistics - Bielefeld University
- MA Human Language Technology - Vrije Universiteit Amsterdam
- BA Linguistics - Leiden University

Besides spiralling my way into NLP,

I also love sports (gym, running, cycling, skiing, ...) and cooking (others with a named sourdough starter?)



**Subproject as part of SAIL:**  
***Longitudinal Analysis of Change and Variety of Natural Language Data***

# **Computational linguistic methods for modeling lexical-semantic dynamics of hate speech**

**by Sanne Hoeken**

Supervised by dr. Özge Alaçam and prof. dr. Sina Zarriß

Subproject as part of SAIL:  
*Longitudinal Analysis of Change and Variety of **Natural Language Data***

# Computational linguistic methods for modeling **lexical-semantic** dynamics of **hate speech**

by Sanne Hoeken

Supervised by dr. Özge Alaçam and prof. dr. Sina Zarriß

Subproject as part of SAIL:  
*Longitudinal Analysis of Change and Variety of Natural Language Data*

# Computational linguistic methods for modeling lexical-semantic dynamics of hate speech

by Sanne Hoeken

Supervised by dr. Özge Alaçam and prof. dr. Sina Zarriß

Subproject as part of SAIL:  
*Longitudinal Analysis of **Change and Variety** of Natural Language Data*

# Computational linguistic methods for modeling lexical-semantic **dynamics** of hate speech

by Sanne Hoeken

Supervised by dr. Özge Alaçam and prof. dr. Sina Zarriß

*Change and Variety*

**dynamics**

*Change* over time

dynamics



*Change* over time

**dynamics**

→ the evolution of  
hateful word meanings

*Variety* across different contexts

dynamics

individual *Variety* across different contexts

dynamics

individual *Variety* across different contexts

Computational linguistic methods for modeling  
lexical-semantic dynamics of hate speech

# Hateful Word in Context Classification

**Sanne Hoeken<sup>1</sup>, Sina Zarriß<sup>1</sup> and Özge Alaçam<sup>1,2</sup>**

<sup>1</sup>Computational Linguistics, Department of Linguistics, Bielefeld University, Germany

<sup>2</sup>Centre for Information and Language Processing, LMU Munich, Germany

{sanne.hoeken, sina.zarriess, oezge.alacam}@uni-bielefeld.de

*The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*

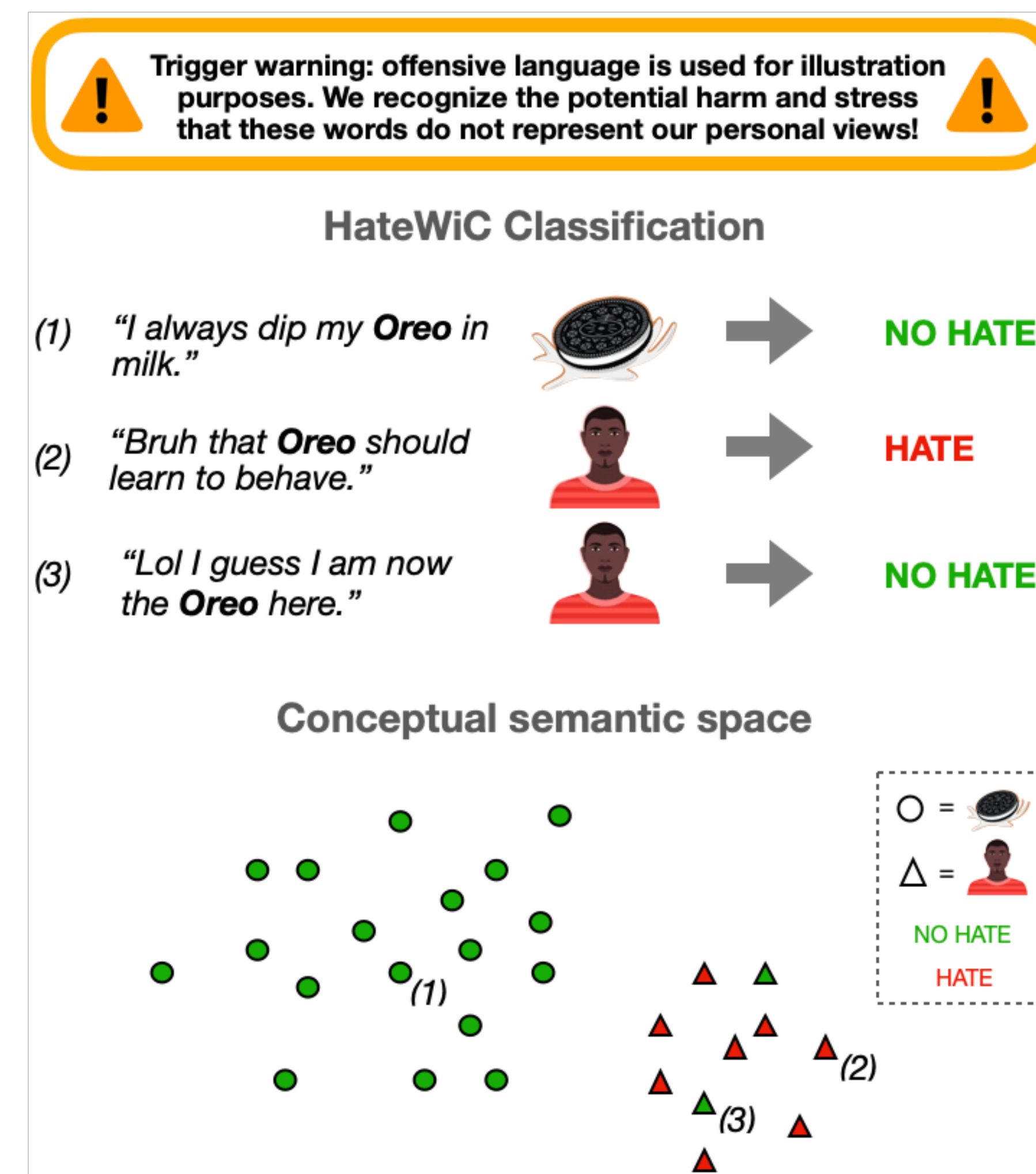
# Table of contents

1. Why Hateful Word in Context (HateWiC) Classification?
2. HateWiC dataset
  - with Wiktionary data and crowd-sourced annotations
3. HateWiC classification
  - with various word sense and annotator representations
4. Results
5. Final remarks

# HateWiC classification

## because hateful senses are not...

- ... enough in focus within HSD research
  - Predominant focus on entire utterances  
(e.g. Waseem & Hovy, 2016; Davidson et al., 2017)
- ... descriptive only, but highly subjective
  - Hateful connotation depends on contextual factors  
(Frigerio & Tenchini, 2019)
  - Current HSD data typically reflect single perspectives  
(e.g. Zampieri et al., 2020; Mathew et al., 2020)



# The HateWiC dataset

## Starting with Wiktionary...

- 1087 entries with at least one sense labeled with category *offensive* or *derogatory*

After cleaning:

- 826 terms
- 1888 sense definitions
- 4029 examples

term →

sense definition →

example →

example →

sense definition →

example →


sense category labels

### Oreo

**Etymology** [ [edit](#) ]

Brand name of unknown origin, trademarked by [National Biscuit Company](#) on 14 March 1912. See the [Etymology](#) section of [Wikipedia](#)'s Oreo article for various theories. In reference to well-assimilated blacks, derived from the slur that they are "black on the outside, white on the inside".

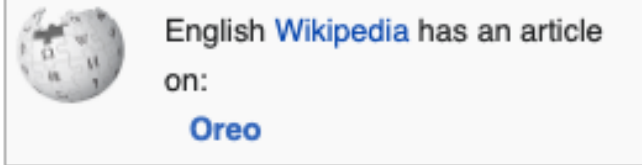

**Pronunciation** [ [edit](#) ]

- ([US](#)) [IPA](#)<sup>(key)</sup>: /ˈɔɪ.ɹi.əʊ/
- Audio ([General Australian](#)): 

**Noun** [ [edit](#) ]

**Oreo** (*plural* **Oreos**)

- (*foods*) A [cookie](#) made of [two wafers](#) joined with a [sugary filling](#), particularly a [Nabisco cookie](#) with [two alkalized cocoa-powder wafers](#) around a [white creme filling](#). [ [quotations](#) ▲ ]
  - 1984 December**, Michael Norman, "Junk Food", [Cincinnati](#) ⓘ, p. 67:  
The **Oreo** is really an antidepressant drug disguised as a cookie. You lost your job? Eat a bag of **Oreos**... There are two ways to eat **Oreos**. Kids like to pull the chocolate wafers apart and eat the sweat, creamy middle first. Adults dunk them in milk.
  - 2022 April 19**, Jennifer Chu, "[MIT Engineers Introduce the Oreometer](#) ⓘ", [MIT News](#):  
While **Oreo** cream may not appear to possess fluid-like properties, it is considered a "yield stress fluid"—a soft solid when unperturbed that can start to flow under enough stress... Curious as to whether other had explored the connection between **Oreos** and [rheology](#), Owens found mention of a 2016 [Princeton University](#) study in which physicists first reported that indeed, when twisting **Oreos** by hand, the cream almost always came off on one wafer.
- ([US](#), [derogatory](#), [ethnic slur](#)) A black person considered to have overly assimilated to white culture, a black race traitor. [ [quotations](#) ▲ ]
  - 1971**, Iceberg Slim, *The Naked Soul of Iceberg Slim*, Holloway House:  
She's a pure **Oreo**. You know, like the cookie, black outside and white inside.

Oreo cookies



# The HateWiC dataset

## Annotation

Pending
Filters
Sort
43 of 4021

Submitted

Example

That numskull will never learn how to compose a letter.

Term

numskull

Definition

A person who refuses to learn or grow mentally.

Annotation guidelines

How would you rate the hatefulness of the meaning of the target term within the specific example text? \*

1 Not hateful

2 Weakly hateful

3 Strongly hateful

4 Cannot decide

Discard

Save as draft

Submit

# The HateWiC dataset

## Annotation

- Crowd-sourced annotations using Prolific
- Three annotations per instance; 250 instances per annotator
  - 48 annotators (with diverse backgrounds)
  - 12442 individual annotations (48% hate and 52% non-hate ratings)
- Inter-annotator agreement of 0.33 (three-class) and 0.45 (binary)
  - inherent subjectivity of the task!

# The HateWiC dataset

## Annotation

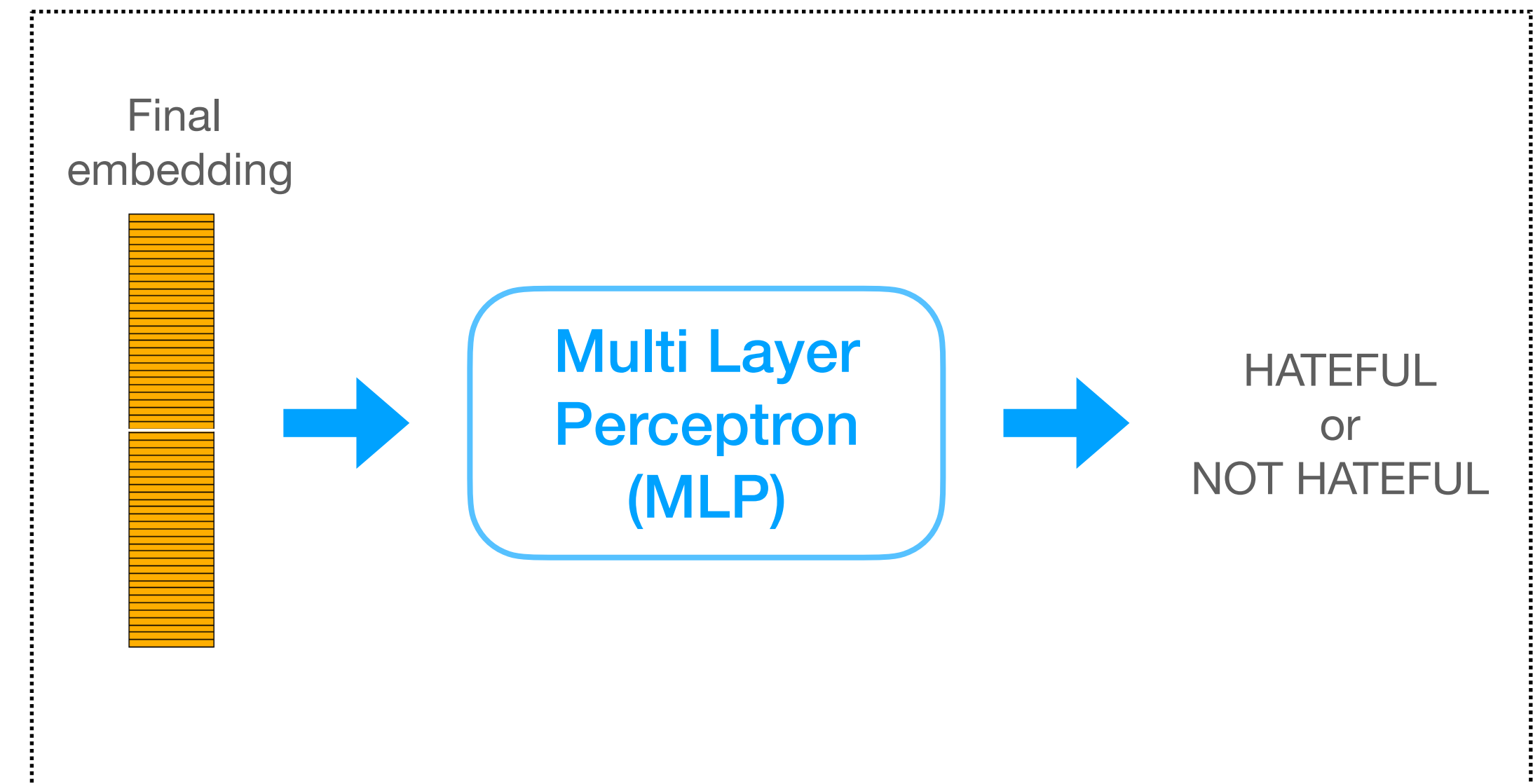
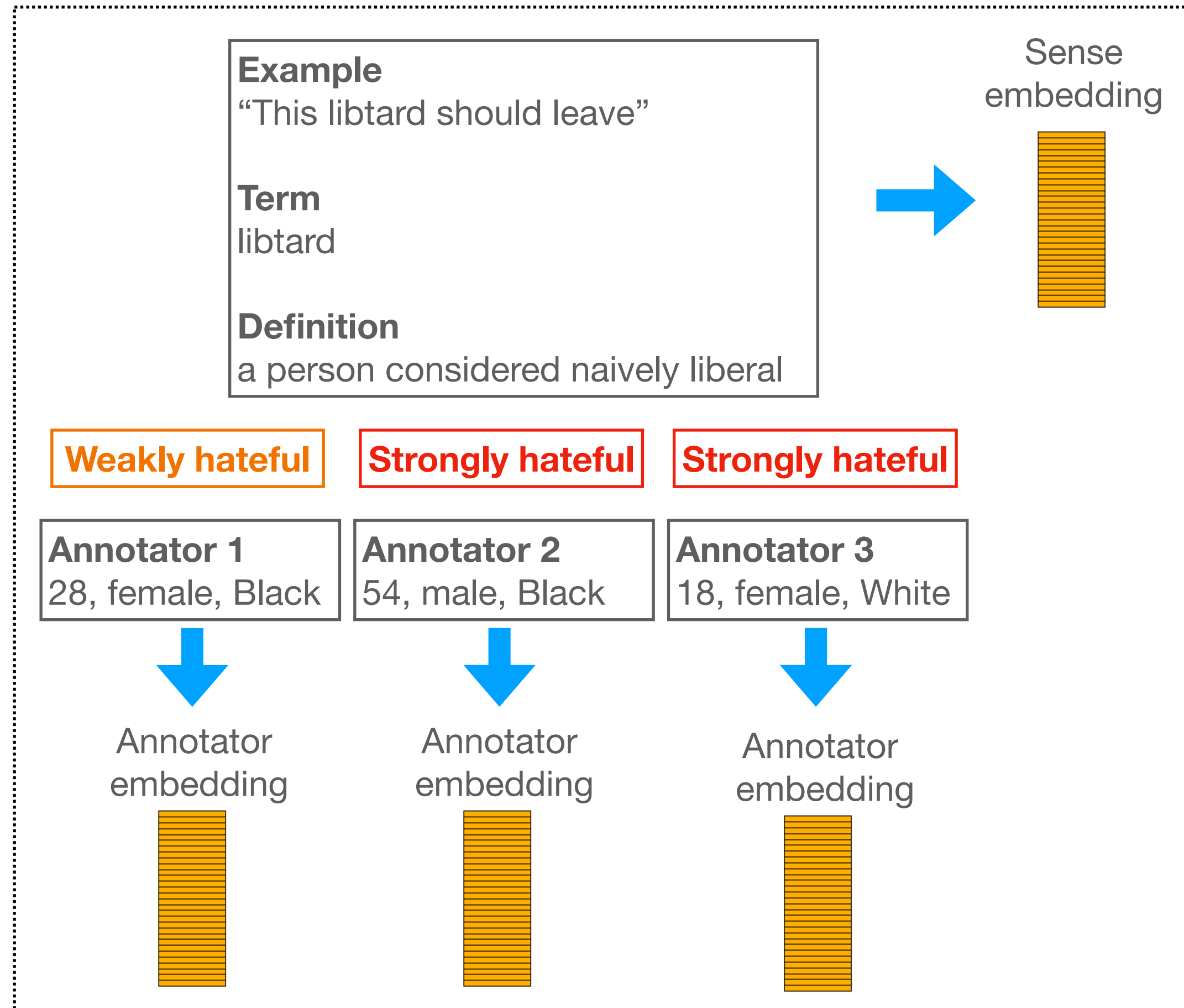
Example	Term	Definition	Annotations	Binary labels	Majority label	Hate-heterogeneous sense	Agreement on binary
(1) “Me having an up to date style even though I’ve turned into a carrot cruncher.”	carrot cruncher	Someone from a rural background.	Nh, Nh, Nh	0, 0, 0	0	True	True
(2) “you’re a friggn’ carrot cruncher and you support the bloody scally’s.”	carrot cruncher	Someone from a rural background.	Sh, Sh, Sh	1, 1, 1	1	True	True
(3) “The bugger’s given me the wrong change.”	bugger	A foolish person or thing.	Wh, Sh, Sh	1, 1, 1	1	False	True
(4) “He’s a silly bugger for losing his keys.”	bugger	A foolish person or thing.	Nh, Wh, Sh	0, 1, 1	1	False	False

Table 1: HateWiC examples with their annotations, illustrating the phenomena of annotator disagreement and hate-heterogeneous word senses (Nh = Not hateful, Wh = Weakly hateful, Sh = Strongly hateful)

- 319 hate-heterogeneous definitions (wrt majority ratings!)  
→ hateful connotation of a word sense is not exclusively determined by its descriptive definition!

# HateWiC Classification

## Overview



# HateWiC Classification

## Sense representations

- Encoder models
  - BERT (*Devlin et al., 2019*)
  - HateBERT (*Caselli et al., 2021*)
  - WSD Biencoder (*Blevins and Zettlemoyer, 2020*)
- Embeddings
  - Word in Context (WiC)
  - Definition (Def)
  - T5-generated definition (T5Def)

### Example

“This libtard should leave”

### Term

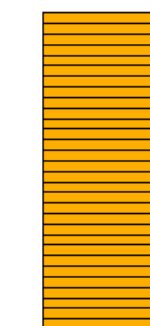
libtard

### Definition

a person considered naively liberal



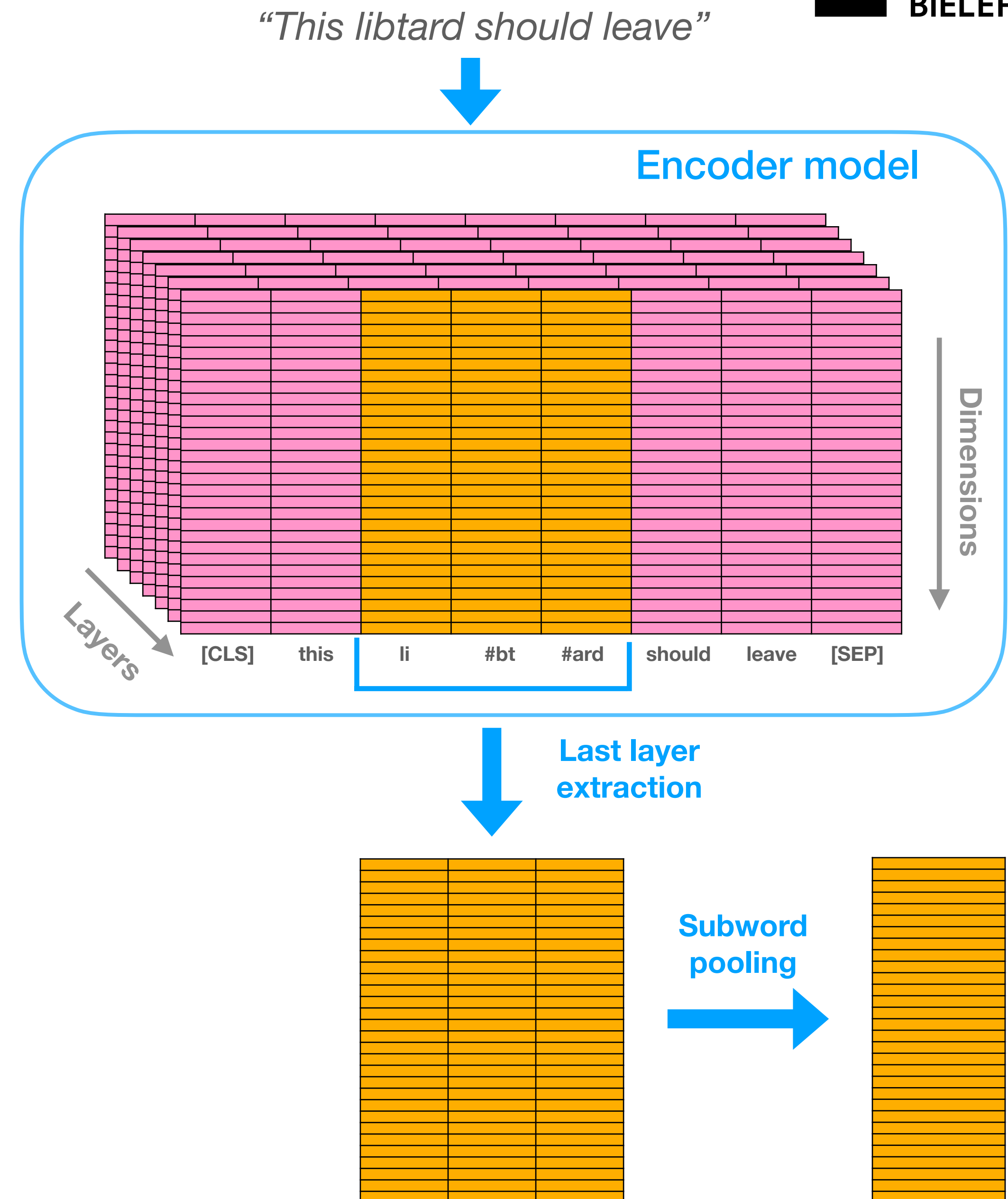
sense  
embedding



# HateWiC Classification

## Sense representations

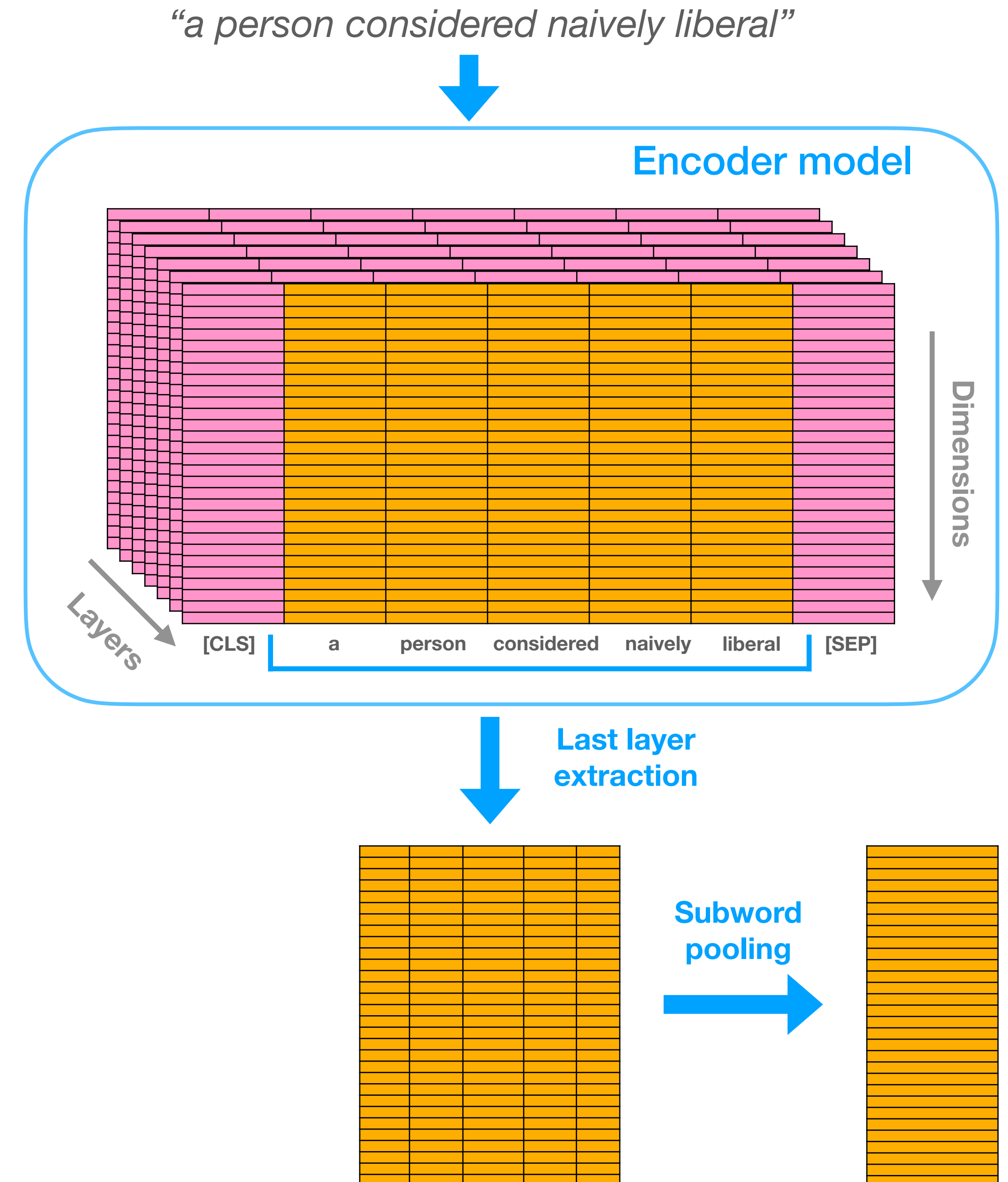
- Encoder models
  - BERT *(Devlin et al., 2019)*
  - HateBERT *(Caselli et al., 2021)*
  - WSD Biencoder *(Blevins and Zettlemoyer, 2020)*
- Embeddings
  - Word in Context (WiC)**
  - Definition (Def)
  - T5-generated definition (T5Def)



# HateWiC Classification

## Sense representations

- Encoder models
  - BERT *(Devlin et al., 2019)*
  - HateBERT *(Caselli et al., 2021)*
  - WSD Biencoder *(Blevins and Zettlemoyer, 2020)*
- Embeddings
  - Word in Context (WiC)
  - **Definition (Def)**
  - T5-generated definition (T5Def)



# HateWiC Classification

## Sense representations

- Encoder models
  - BERT (Devlin et al., 2019)
  - HateBERT (Caselli et al., 2021)
  - WSD Biencoder (Blevins and Zettlemoyer, 2020)
- Embeddings
  - Word in Context (WiC)
  - Definition (Def)
  - **T5-generated definition (T5Def)**

*“This libtard should leave.  
What is the definition of libtard?”*



**FLAN-T5 Base**  
(Giulianelli et al., 2023)

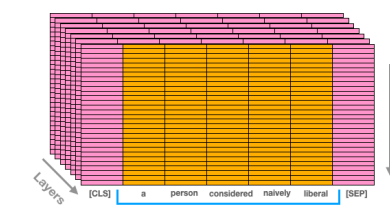
[finetuned on English definitions and usage examples]



*“a person who is libertarian”*



**Encoder model**



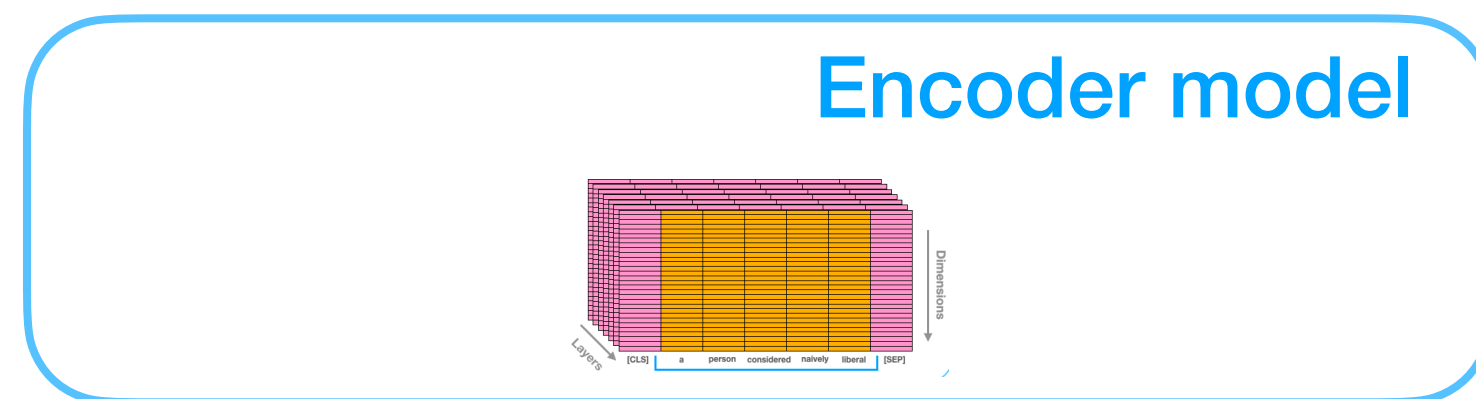


# HateWiC Classification

## Annotator information

- Annotator description embeddings (Ann)

*“Reader is 28, female and Black”*



# HateWiC Classification

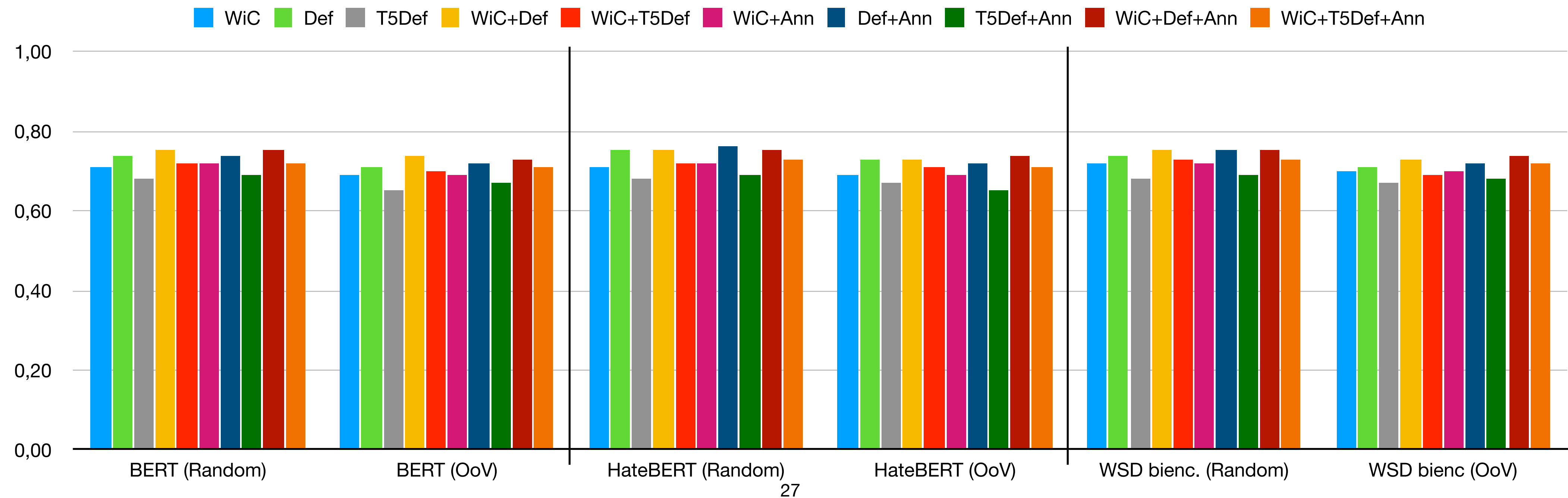
## Evaluation

- Evaluating individual label prediction (i.e. 12442 instances)
- Ten-fold cross-validation with two variants of data split for each fold:
  1. **Random:** based on example sentences
  2. **Out-of-Vocabulary (OoV):** based on terms  
→ testing zero-shot capabilities

# Results

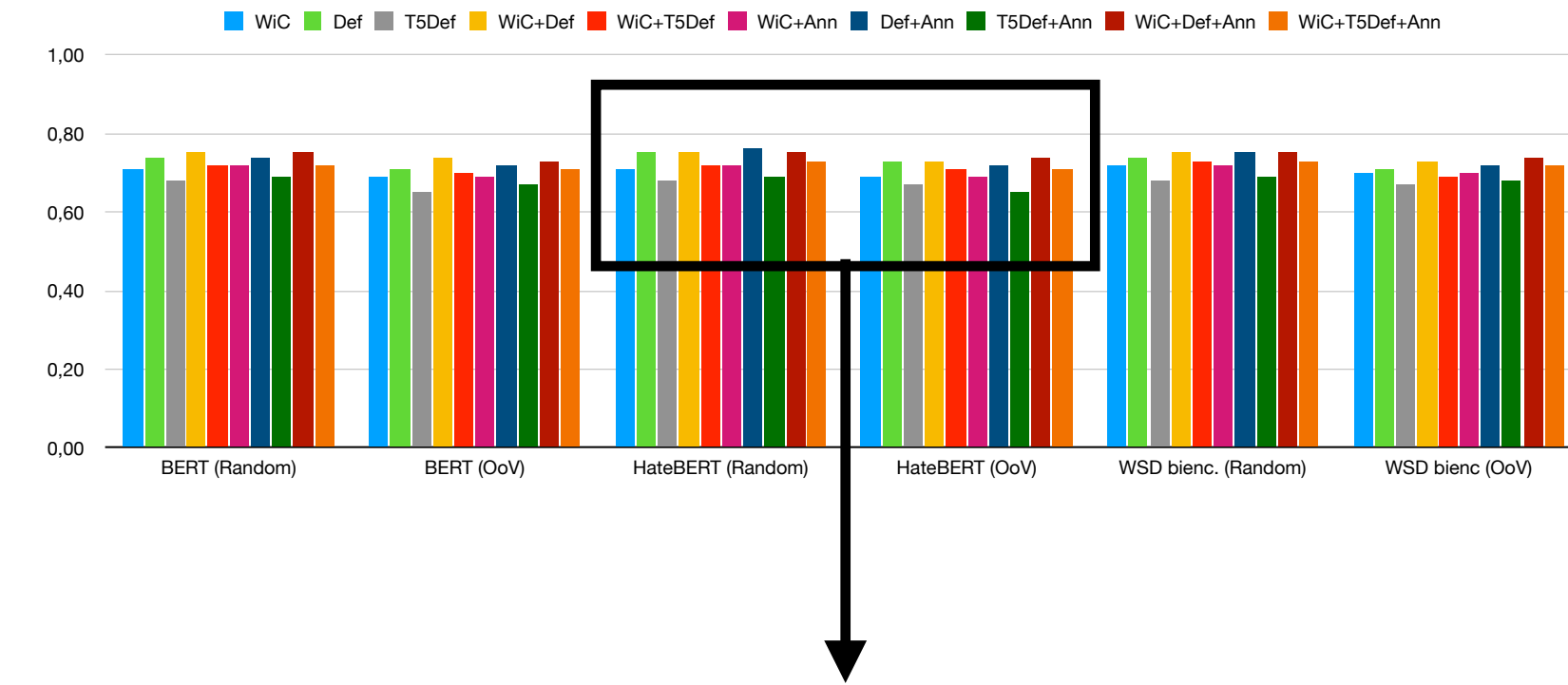
## Overall

- Effectiveness of all methods
- Only slight drop for OoV-terms
- Negligible differences between encoders

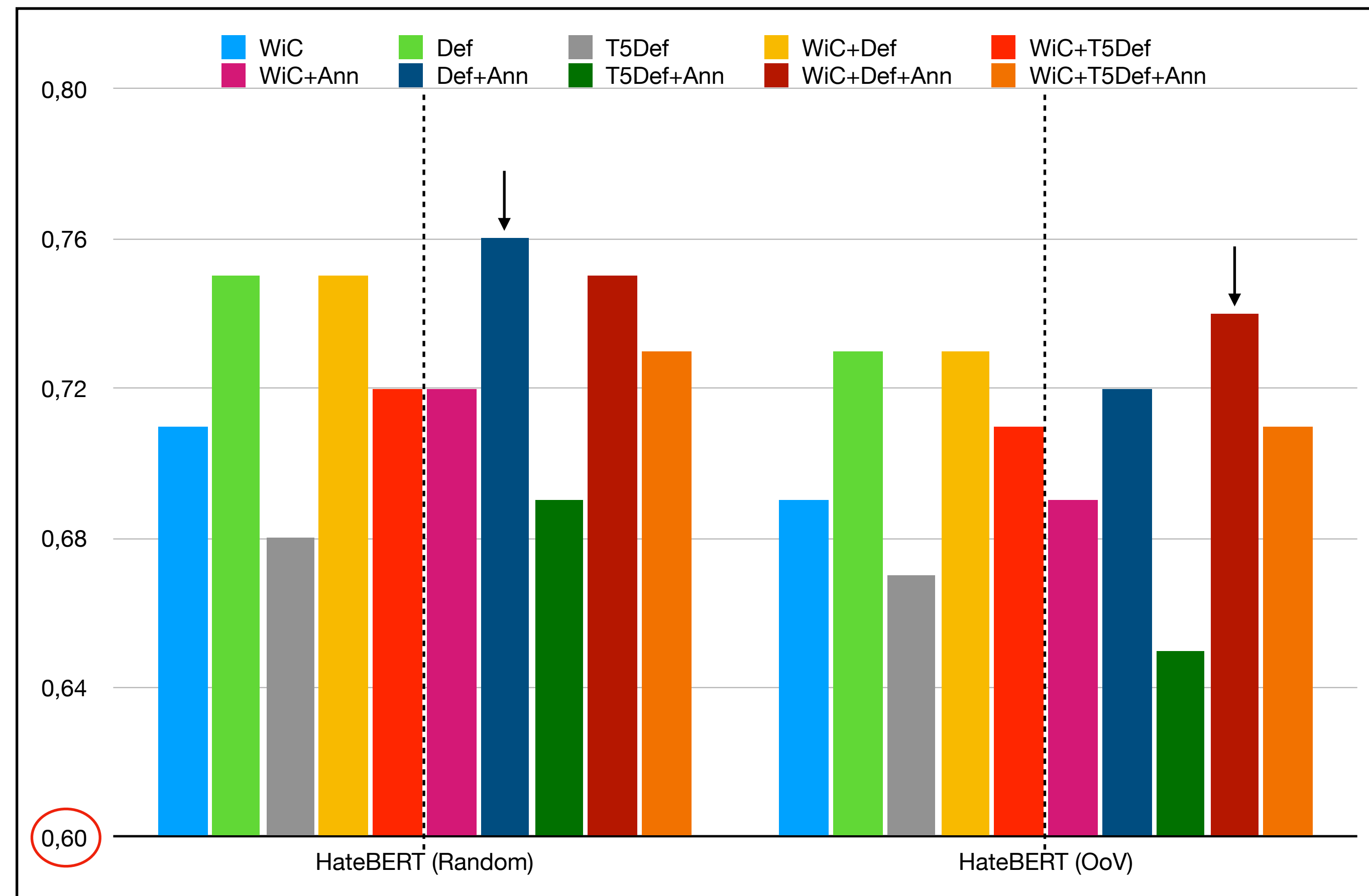


# Results

## Embeddings



- Def and WiC+Def > WiC
- T5Def performs worst
- +Ann: minimal improving effect
- Def+Ann best for Random
- WiC+Def+Ann best for OoV terms

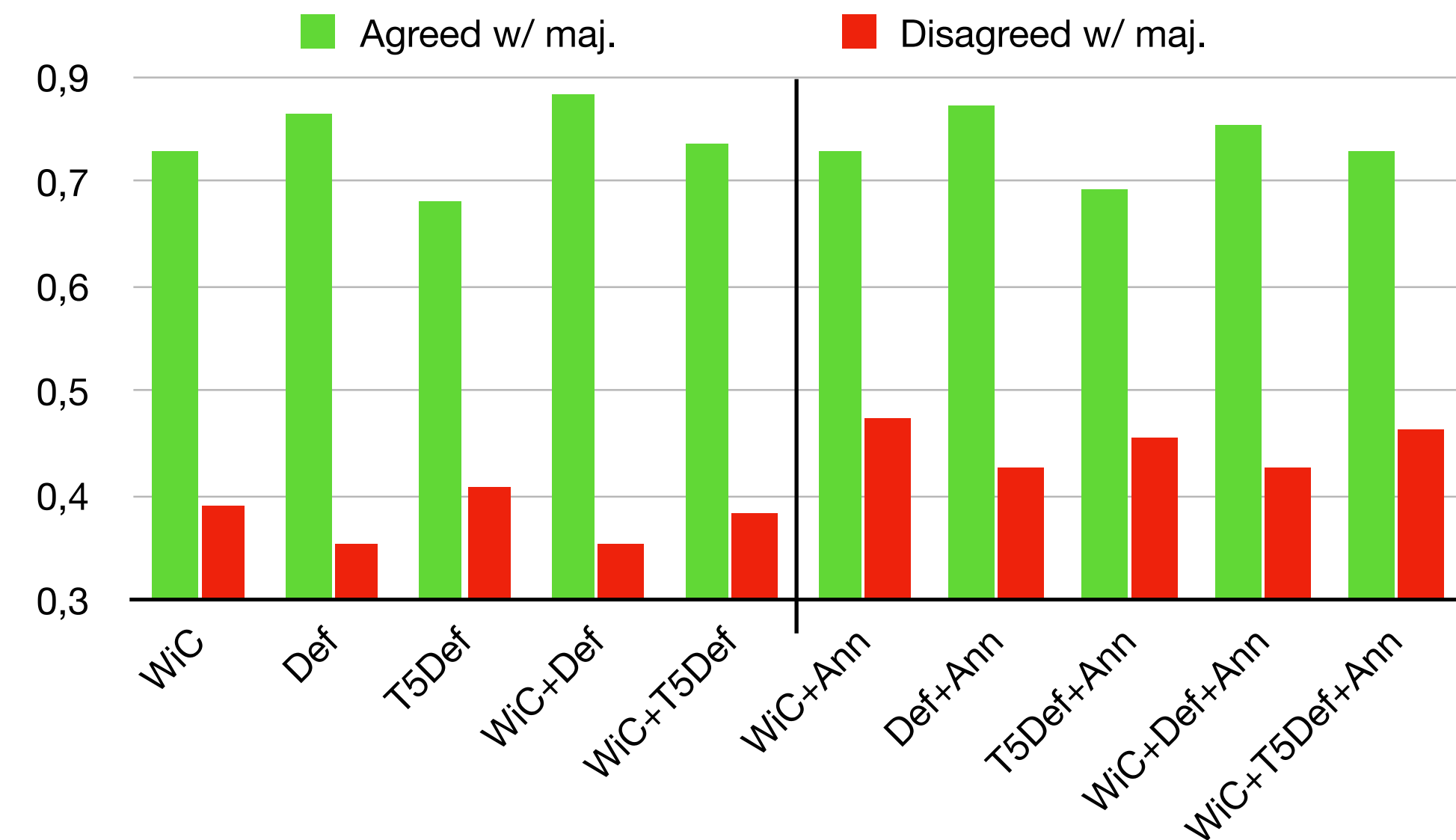


# Results

## in highly subjective scenarios

- Scenario 1: Hate-heterogeneous sense definition
- Scenario 2: Annotator disagrees with majority label
- In both, performance of all embeddings drops significantly!

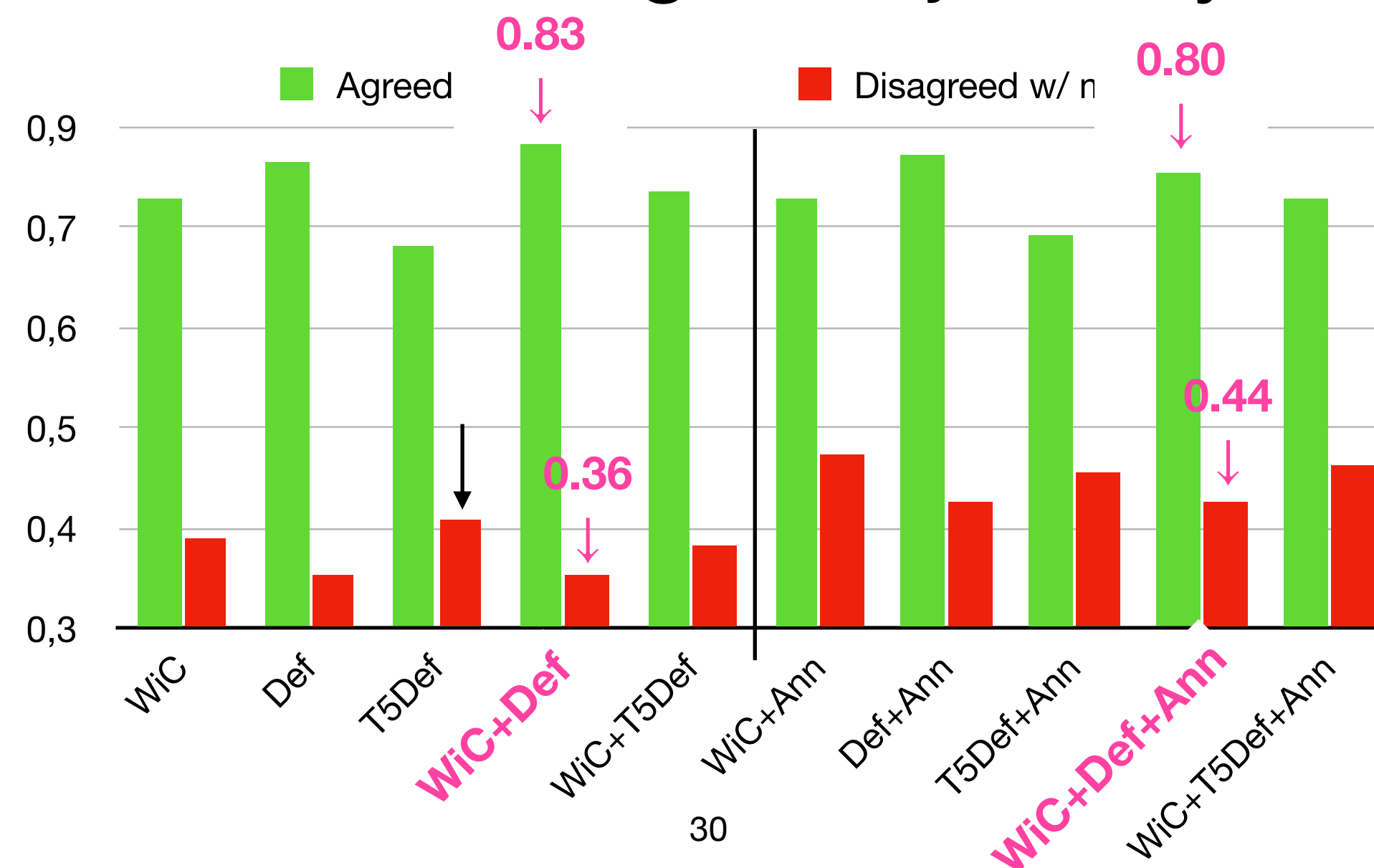
*(results for HateBERT with Random test split)*



# Results

## in highly subjective scenarios

- Highest drop for Def embeddings (up to 47%), less so for T5-generated → aligning with more context-specific nature of T5Def-embeddings
- Incorporating annotator information mitigates drop up to 11% → thus, contributes to cases with high-subjectivity



# Final remarks

## Insights into hate speech detection through the lens of lexical semantics!

- To define or not define?  
→ potential usefulness of generating context-specific definitions for subjective lexical semantic tasks.
- To individualize anyway?  
→ yes, value of personalizing models to account for subjectivity in annotations.

# Final remarks & next steps

## Insights into hate speech detection through the lens of lexical semantics!

- To define or not define?  
→ potential usefulness of generating context-specific definitions for subjective lexical semantic tasks.
- **Next steps:** more advanced and task-tailored definition generation methods?
- To individualize anyway?  
→ yes, value of personalizing models to account for subjectivity in annotations.
- **Next steps:** exploring the effectivity of different annotator embeddings?  
→ *going beyond annotator demographics?*



# Next steps

## My questions...



Utrecht  
University



Dealing with Meaning Variation in NLP

- Can we systematically identify dimensions to **profile** hateful word meanings in order to explain their variation?
  - i) Lexical semantic dimensions: what semantic features (e.g. referential transparency), relations (e.g. metaphor) and literal domains (e.g. animals, food, diseases) can we observe?
  - ii) Pragmatic dimensions: what contextual features can we observe (e.g. speaker intention and identity, time, place)?
- Can we **model** meaning variation of hateful words better, incorporating this structured information?

# Next steps

Some more concrete (but preliminary)  
example thoughts...



Utrecht  
University



Dealing with Meaning Variation in NLP

- Can we systematically identify dimensions to **profile** hateful word meanings in order to explain their variation?
  - i) Lexical semantic dimensions: what semantic features (e.g. **referential transparency**), relations (e.g. metaphor) and literal domains (e.g. animals, food, diseases) can we observe?
    - **Referential transparency**: a *bastard* versus *cheesehead* issue?
    - E.g., do word meanings with more descriptive content carry a higher degree of derogatory autonomy?

# Next steps

Some more concrete (but preliminary)  
example thoughts...



Utrecht  
University



Dealing with Meaning Variation in NLP

- Can we systematically identify dimensions to **profile** hateful word meanings in order to explain their variation?
  - i) Lexical semantic dimensions: what semantic features (e.g. referential transparency), relations (e.g. metaphor) and **literal domains** (e.g. animals, food, diseases) can we observe?
    - **Literal domain**: a *pig* versus *potato* issue?
    - E.g., are metaphorical mappings (onto a target group) from animals more sensitive to reinforce a subjective hateful meaning than from food?

# Next steps

## My questions...



Utrecht  
University



Dealing with Meaning Variation in NLP

- Can we systematically identify dimensions to **profile** hateful word meanings in order to explain their variation?
  - i) Lexical semantic dimensions: what semantic features (e.g. referential transparency), relations (e.g. metaphor) and literal domains (e.g. animals, food, diseases) can we observe?
  - ii) Pragmatic dimensions: what contextual features can we observe (e.g. speaker intention and identity, time, place)?
- Can we **model** meaning variation of hateful words better, incorporating this structured information?

**Next steps**  
**Your questions?**



**Thank you for listening!**



# References

- Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006-1017, Online. Association for Computational Linguistics.
- Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17-25, Online. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAI Conference on Web and Social Media*, 11(1):512–515.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aldo Frigerio and Maria Paolo Tenchini. 2019. Pejoratives: a classification of the connoted terms. *Riviera Italian di Filosofia del Linguaggio*, 13(1).
- Mario Giulianelli, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130-3148, Toronto, Canada. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAI Conference on Artificial Intelligence*, 35(17):14867–14875.
- Zeera Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.