

Arabic Toxic Content Detection in NLP (Panel Discussion at IWABigDAI)

Nedjma Ousidhoum
Department of Computer Science and Technology
University of Cambridge

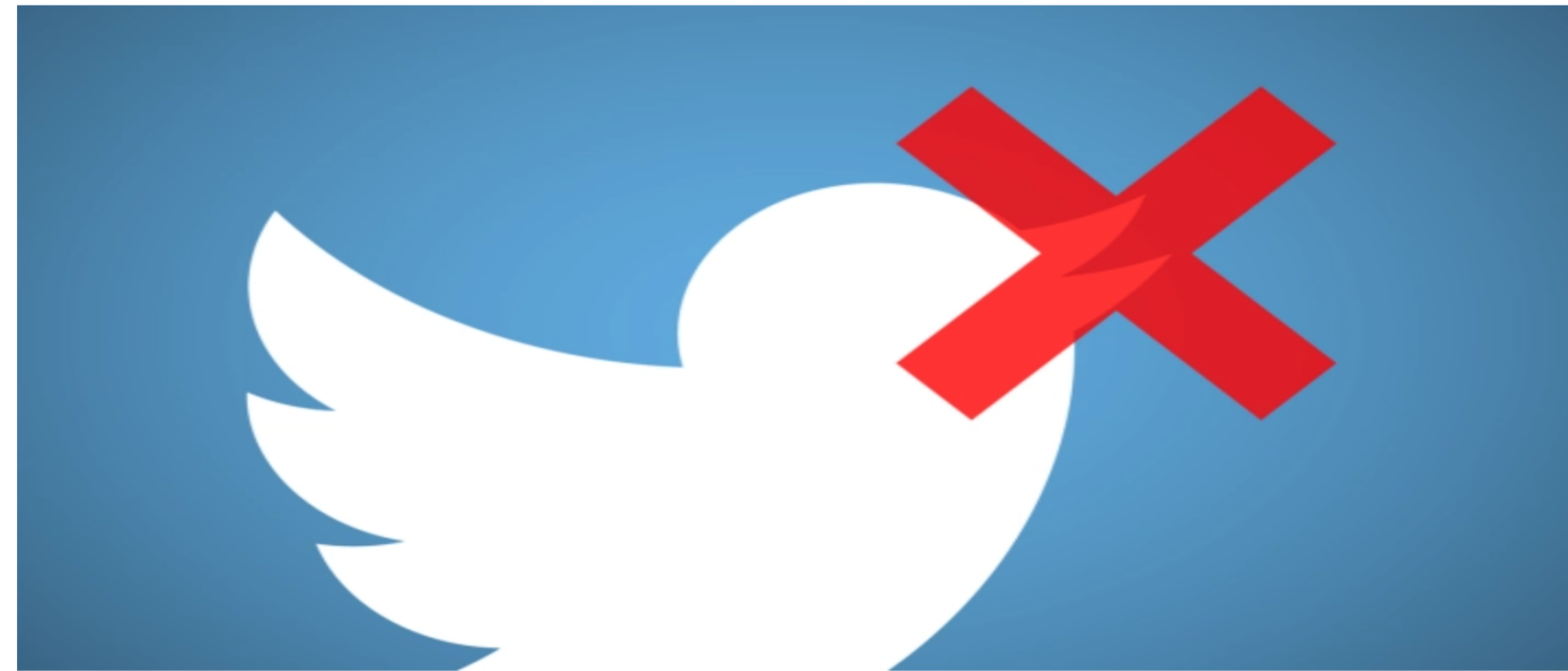
This was part of my PhD work at HKUST



Online Policy Against Hate Speech

Twitter expands hateful conduct rules to ban dehumanizing speech around age, disability and now, disease

Sarah Perez @sarahintampa / 1 year



Is this enough?

All [Age Group] are leeches and don't deserve any support from us.

People with [Disease] are rats that contaminate everyone around them.

People with [Disability] are subhuman and shouldn't be seen in public.

[Religious Group] should be punished. We are not doing enough to rid us of those filthy animals.

Probably not!

Online Policy Against Hate Speech

- Common discriminative attribute
 - ethnicity, religion, gender, ...
- Presence of slurs
- Intention of the post
- Humour has a specific status



Challenges: Topic Coverage is very often language-specific



@user لا تكون تكثر من شرب بول البعير فقد اكثرت من الشرب حتى اصبحت حمارا

TR you may have drunk too much camel urine so you became a donkey

- Different targets!
- Topics are spoken about differently (compared to English for example)!



التزامن مع محاولات تحريض المواطنين على وطنه بدء المشردون "اللاجئون" في كندا وبريطانيا تحريض السفهاء على الهجرة

*TR attempts to make citizens hate their country come in time with the **vagabond “refugees”** in Canada and the UK encouraging fools to **immigrate***

What is Hate Speech?



hate speech

noun [U]

UK /'heɪt ,spi:tʃ/ US /'heɪt ,spi:tʃ/



public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation (= the fact of being gay, etc.):



SINCE 1828

hate speech *noun*



Save Word

Definition of *hate speech*

: speech expressing hatred of a particular group of people

// *Hate speech* is not allowed at school.



UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH

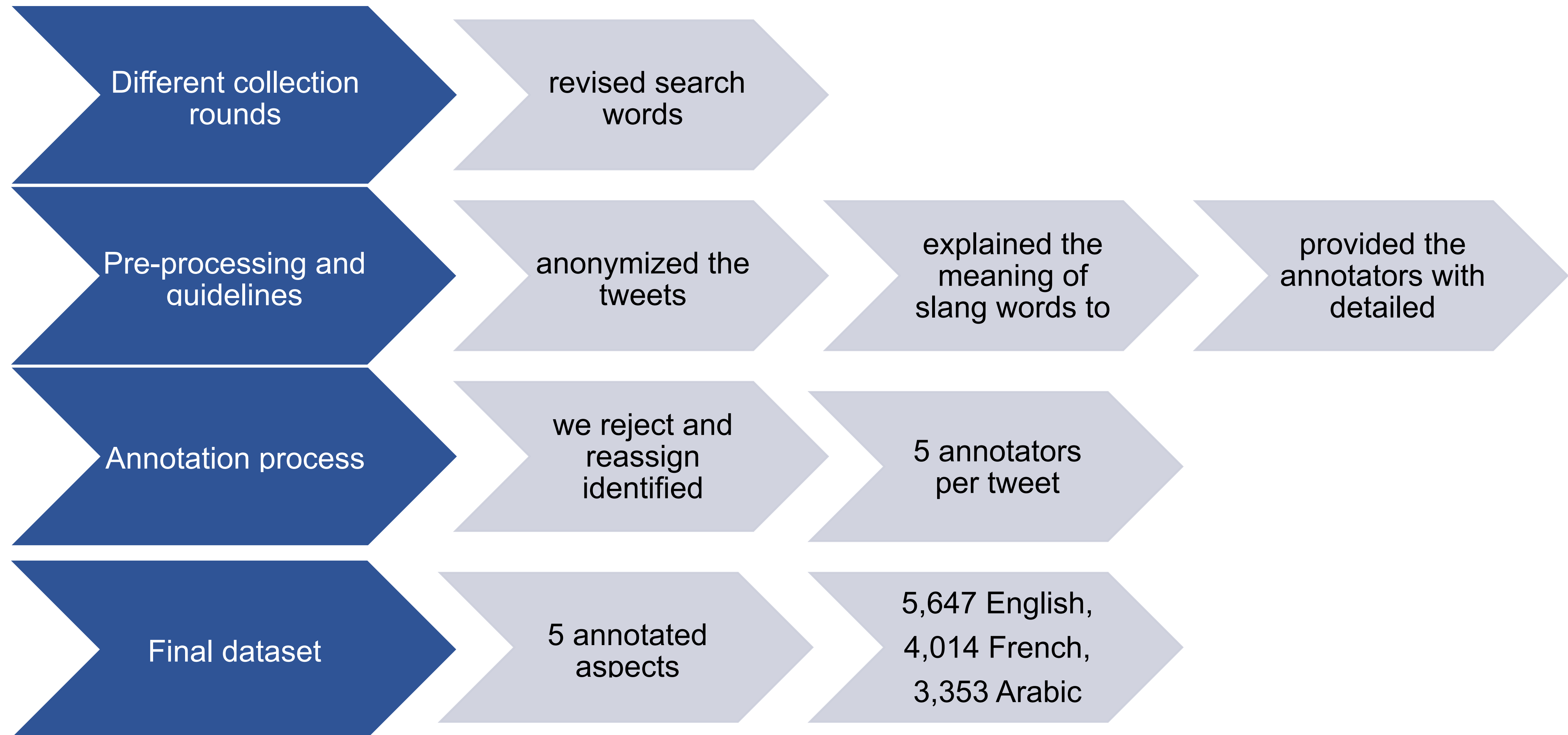
What is hate speech?

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.

What is Hate Speech?

- Seeking to *silence and criticize a minority without an **und**ounded **argument***, and requires the statement (tweet) to *be **onsive** screen name or use a slur, and **promote xen*** (Lewin and Hovy 2016)
- Language that is used *towards a targeted group or is **intended to be** **affiliate**, or to insult the members of the group. In **some cases** may also be **language that threatens or incites violence*** (Davidson et al. 2017)
- We chose to use the term **hate speech** to refer to toxic language with respect to different nuances

Dataset



Dataset Annotation

- Our annotations indicate the tweet's
 - directness
 - the text is direct or indirect
 - hostility type (multi-label)
 - degree of hostility: offensive, disrespectful, hateful, fearful out of ignorance, abusive, or normal
 - target attribute
 - attribute based on which it discriminates against an individual or a group of people: origin, religious affiliation, gender, sexual orientation, disability, other
 - annotator's sentiment (multi-label)
 - how the annotators feel about its content: shock, fear, disgust, anger, sadness, indifference
 - targeted group / individual
 - 16 groups

Multilingual Hate Speech Detection

Multitask Learning

- Consider each annotated aspect as a classification task
- 5 tasks per dataset
 - 3 datasets -> 1 per language
- Test multi-task learning on the different datasets and tasks

Expectations vs. Reality

- We can reach a unified representation that takes nuances into account
 - Yes but the disagreement is high
- Transfer and multitask learning can improve the detection
 - Not really!
 - Additional experiments including data augmentation using machine translation did not help. (Results aligned with Nozza 2021's)

Why didn't the multilingual settings boost the performance that much?



Cultural and cross-lingual differences



Why Should We Look into Cultural Differences?

- Languages evolve
- Linguistic nuances are culture-dependent
- Lack of context in the data
- Social structures are language and culture-dependent

Frequent Words in Toxic Posts

Arabic Datasets



Mulki et al. (Levantine dataset)



Albadi et al. (Sectarian dataset)



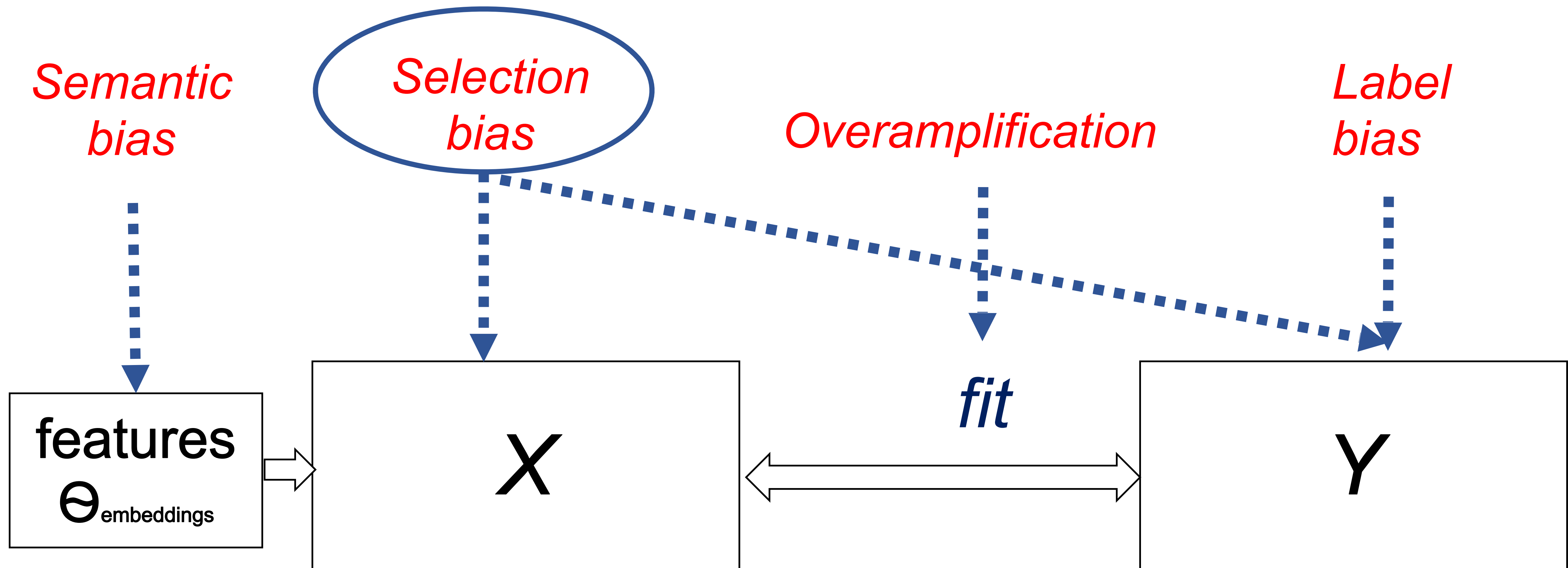
Our dataset

We observe recurring topics, could this be due to the same collection strategy based on keywords?

Let's examine selection bias



Bias in Toxic Language Detection Models



Evaluating Selection Bias

- Run topic models
- Compare topics and keywords
 - based on a semantic similarity measure
 - on average via a first metric **B1**
 - by looking at the maximal similarity between topic words and the selection keywords on average via a second metric **B2**

Bias Evaluation Metrics

B1

determines the **average stability of topics** given keywords

measures the **relatedness of keywords and each topic word then each topic** and computes the average over the number of topics

Sim₁(topic, keywords) =

$$\frac{1}{\#words} \frac{1}{\#keywords} \sum_1^{\#words} \sum_1^{\#keywords} \mathbf{Sim}(word, keyword)$$

$$\mathbf{B}_1 = \frac{1}{\#topics} \sum_1^{\#topics} \mathbf{Sim}_1(topic, keywords)$$

B2

measures **how regularly keywords** appear in topics

verifies whether **each topic word is similar or identical to a keyword** and computes the average over the number of topics

Sim₂(topic, keywords) =

$$\frac{1}{\# words} \sum_1^{\#words} \max \mathbf{Sim} (word, keywords)$$

$$\mathbf{B}_2 = \frac{1}{\#topics} \sum_1^{\#topics} \mathbf{Sim}_2(topic, keywords)$$

Sim can be based on (1) Babylon multilingual embeddings, (2) WordNet , (3) other embeddings

Lessons learned

- There is no unified representation of toxic language for good reasons
 - no unanimous definition of hate speech vs. toxic/abusive language
 - annotating data given the lack of context is hard
 - cultural differences
 - set constraints before normalizing an annotation scheme
- Transfer learning can help with the detection under specific conditions
- Cultural differences and selection bias ought to be addressed early

Points to be considered in Arabic Toxic Content Detection/Data Collection

- Diglossia especially given the keyword/hashtag-based collection process.
- False generalisation and cognates within different dialects
 - E.g. North African dialects are different and should not be included in one “Maghrebi” family of dialects for such a subjective task (see <https://tinyurl.com/yejra7zu>)
- Topics’ coverage and social structures differ across different Arabic speaking countries
- The annotators should be native speakers and ideally survivors/potential victims