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

Nedjma Ousidhoum
Department of Computer Science and Technology
University of Cambridge

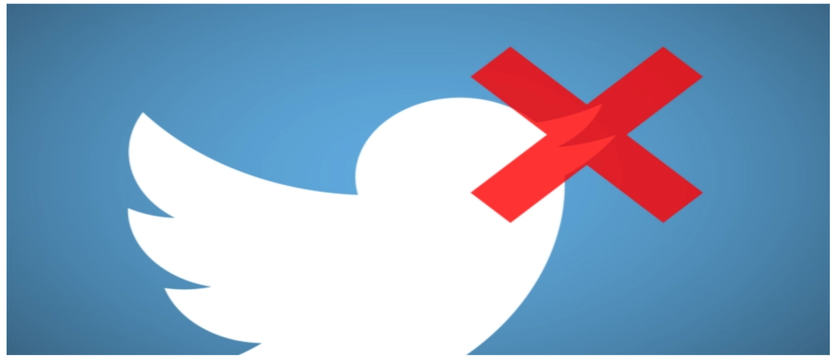
UNIVERSITY OF CAMBRIDGE

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 ...minate everyone around them.

Probably not!

People w...sability] 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.

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

@userلاتكون تكثر من شرب بول البعير فقد اكثرت من الشرب حتى اصبحت حمارا

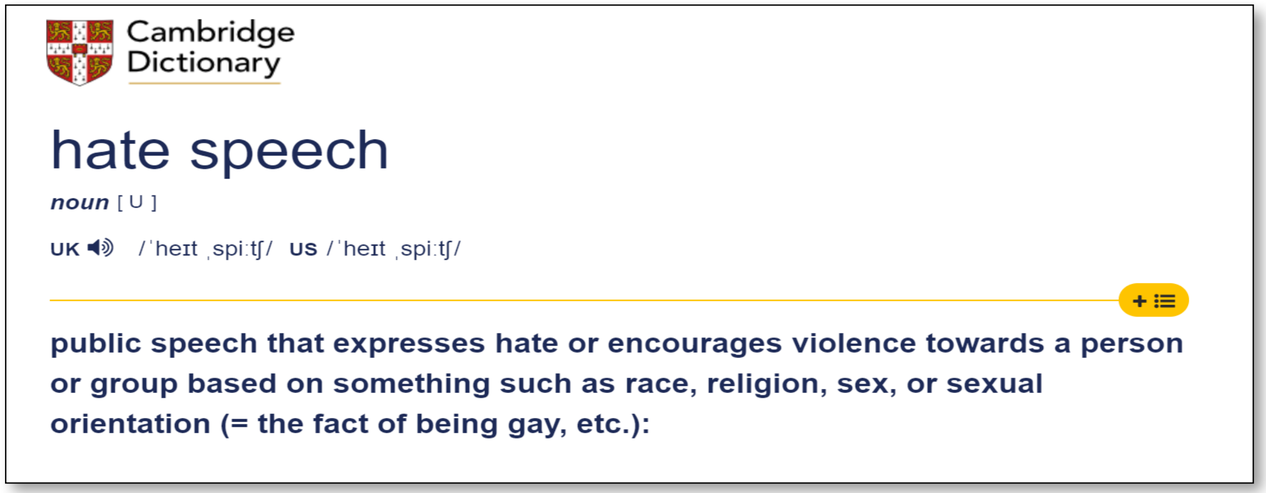
TR *you may have drunk too much camel urine so you became a donkey*

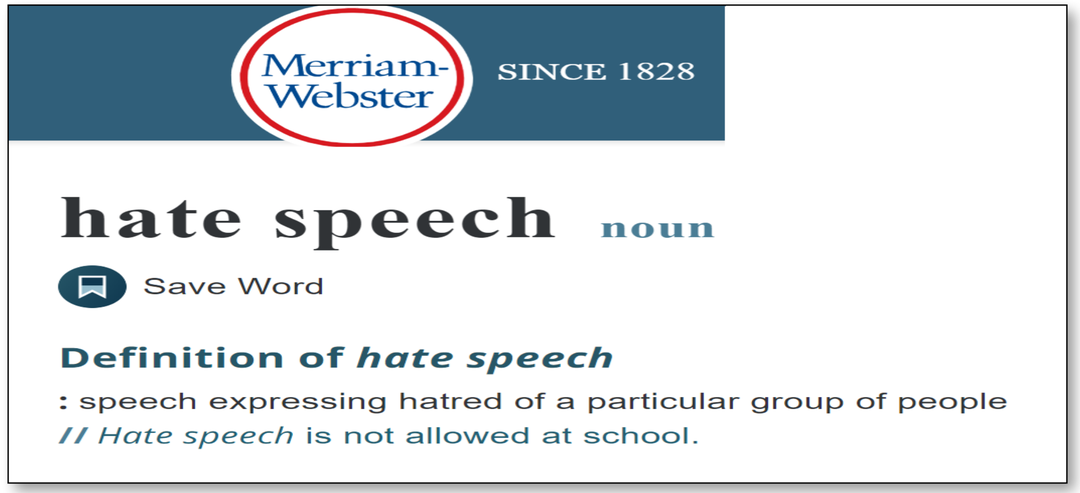
- Different targets!
- Topics are spoken about differently (compared to English fo example)!

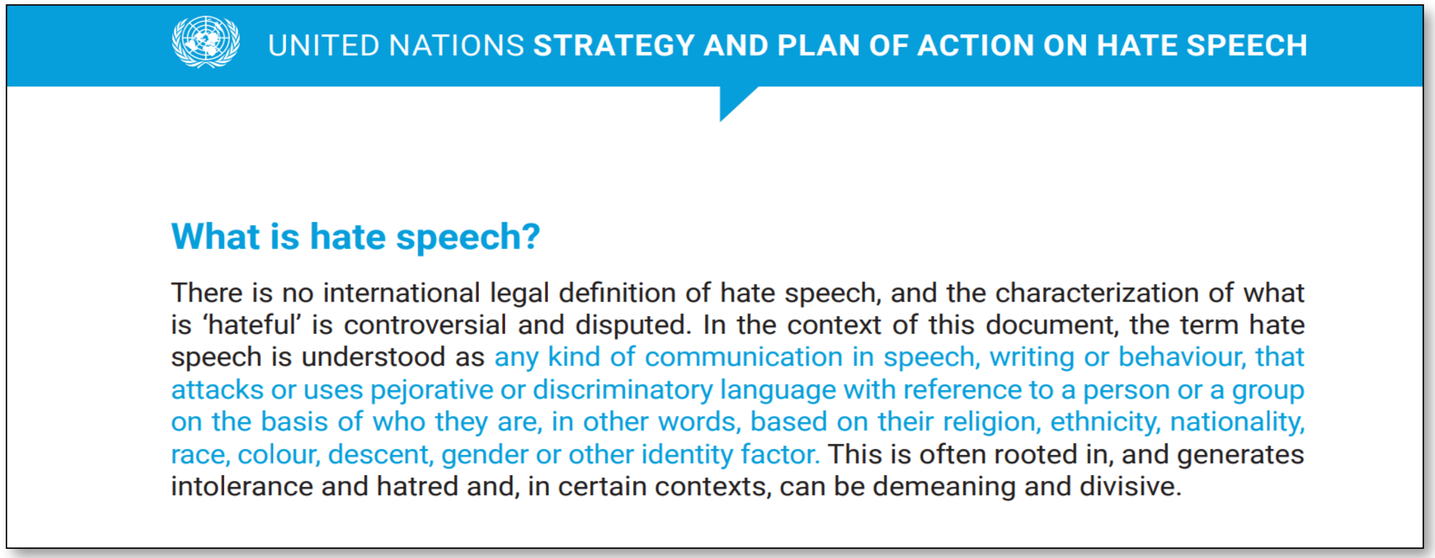
التزامن مع محاولات تحريض المواطن على وطنه بدء المشردون "اللاجئون" في كندا وبريطانيا تحريض السفهاء على الهجرة

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?







What is Hate Speech?

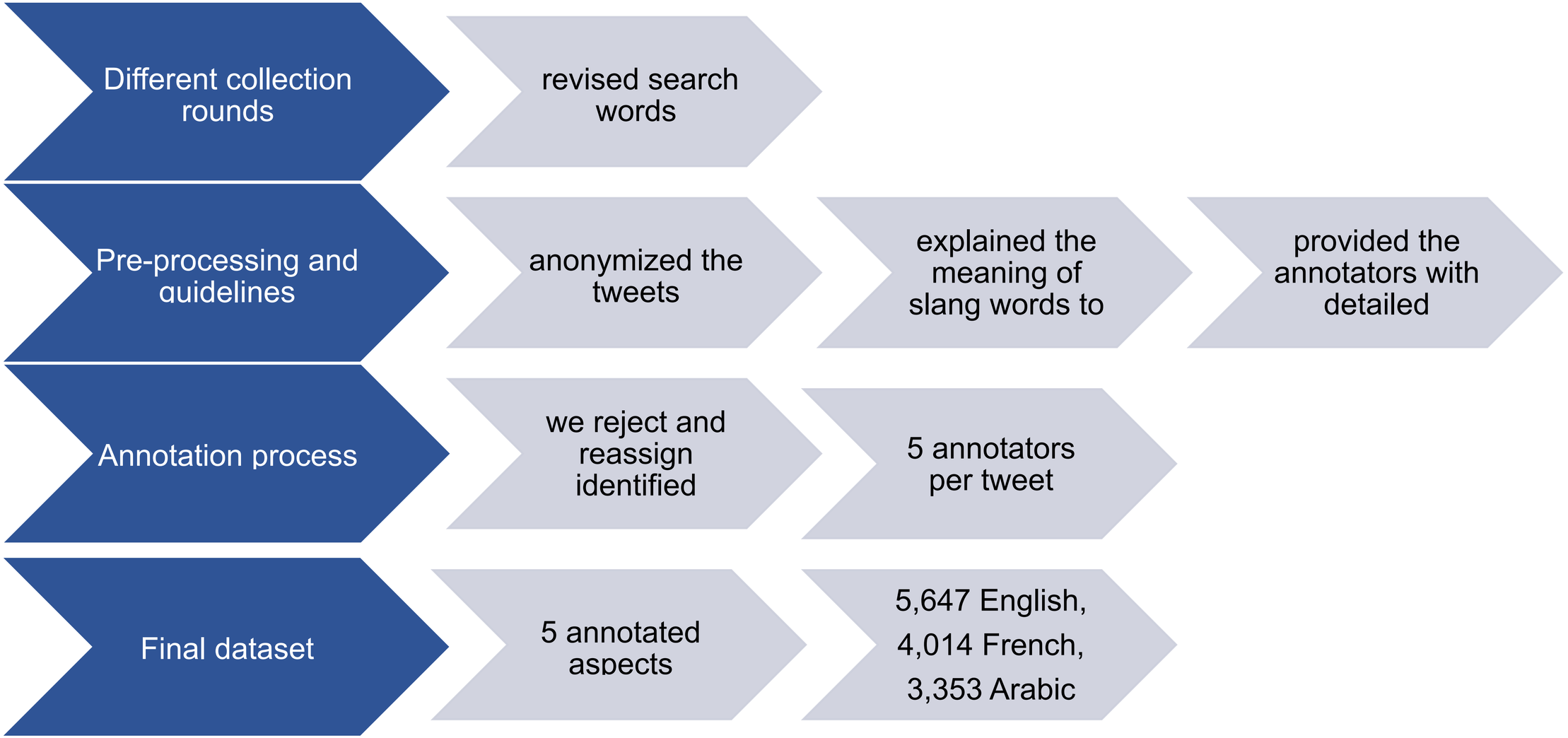
- *Seeking to silence and criticize a minority without an [illegible] unded argument, and requires the statement (tweet) to [illegible] ensive screen name or use a slur, and promote xen[illegible]* and Hovy 2016)

- Language that is used [illegible] towards a targeted group or is intended to be [illegible] miliate, or to insult the members of the group. In [illegible] may also be language that threatens or incites violence [illegible] quired to contain slurs (Davidson et al. 2017)

- We chose to use the term **hate speech** to refer to toxic language with respect to different nuances

NO UNIFIED DEFINITION!

Dataset



Ousidhoum et al.: Multilingual and Multi-aspect Hate Speech Analysis.(EMNLP 2019)

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. (Levantive dataset)

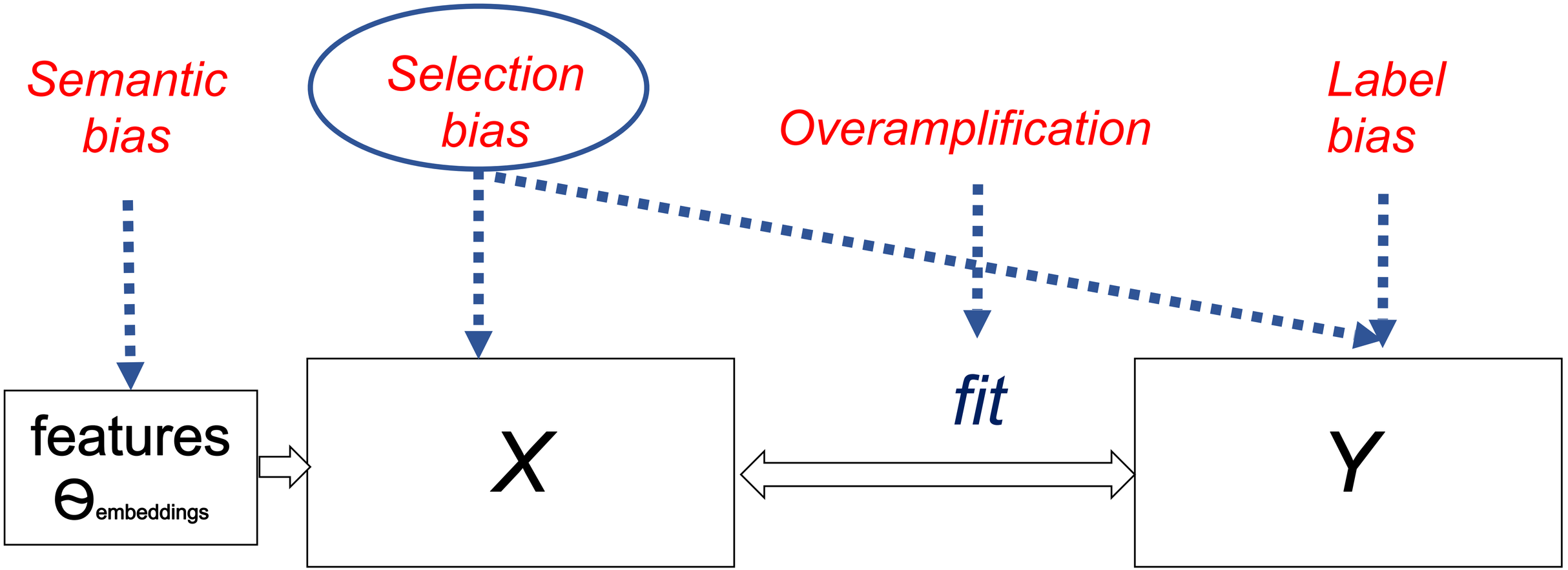
Albadi et al. (Sectarian dataset)

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

B_1	B_2
determines the average stability of topics given keywords	measures how regularly keywords appear in topics
measures the relatedness of keywords and each topic word then each topic and computes the average over the number of topics	verifies whether each topic word is similar or identical to a keyword and computes the average over the number of topics
$\mathbf{Sim}_1(\text{topic, keywords}) = \frac{1}{\#words}\frac{1}{\#keywords}\sum_{1}^{\#words}\sum_{1}^{\#keywords}\mathbf{Sim}(word, keyword)$ $B_1 = \frac{1}{\#topics}\sum_{1}^{\#topics}\mathrm{Sim}_1(topic, keywords)$	$\mathbf{Sim}_2(\text{topic, keywords}) = \frac{1}{\#\,words}\sum_{1}^{\#words}\max\mathbf{Sim}\,(word, keywords)$ $B_2 = \frac{1}{\#topics}\sum_{1}^{\#topics}\mathrm{Sim}_2(topic, keywords)$

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 generalisations 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