

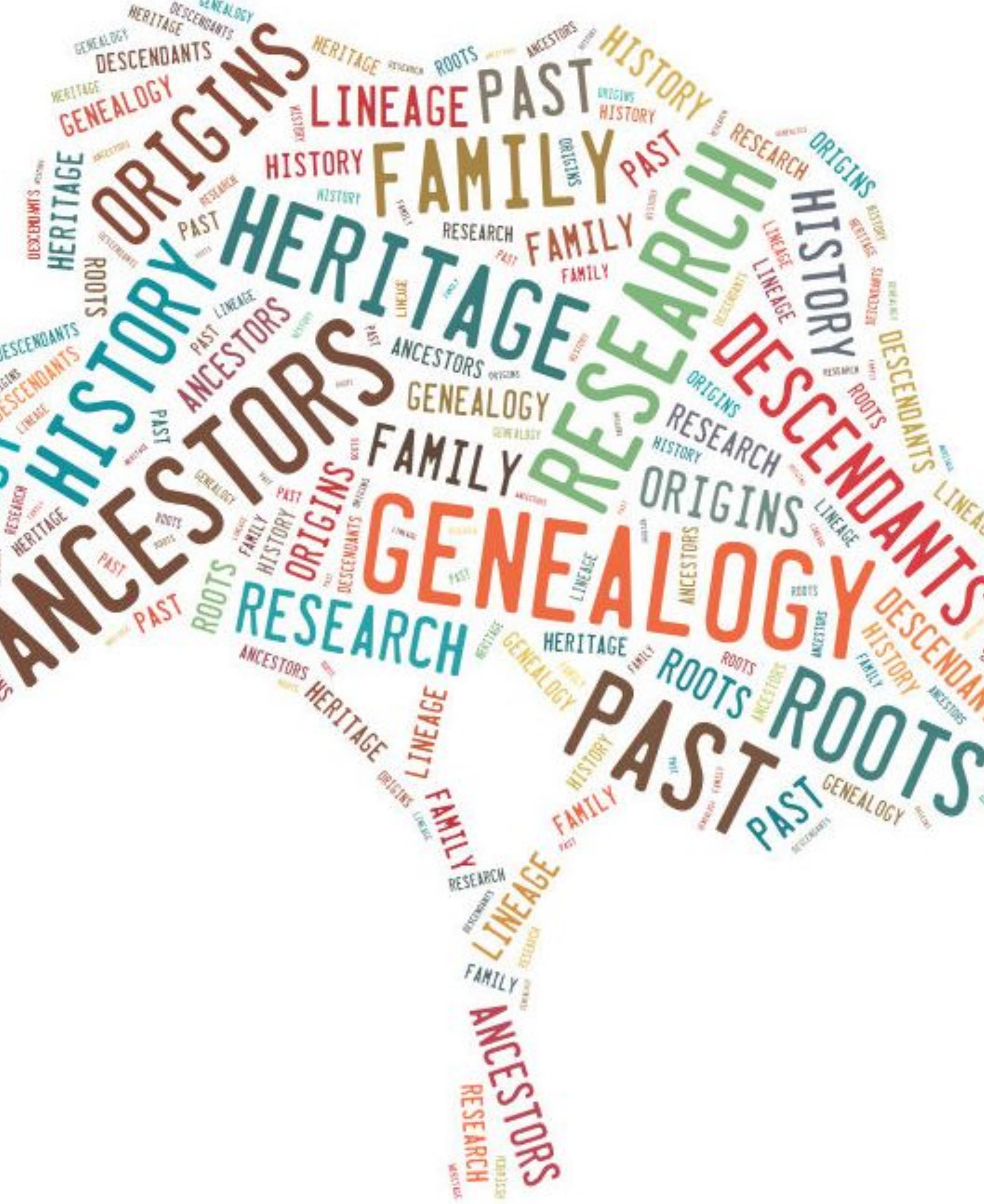
ZEERAK TALAT | 26.02.2020 | ITU
UNIVERSITY OF SHEFFIELD

ON THE MARGINS OF ABUSE DETECTION TECHNOLOGIES



LAW





A GENEALOGY OF ABUSE

- ▶ Dictionary based approaches
- ▶ Linear models
- ▶ Word embeddings
- ▶ Non-linear models



IMAGINATIONS

- ▶ Designer

- ▶ Address marginalization processes as they occur online
- ▶ Address complex and nuanced issues (e.g. stereotyping, vilification)
- ▶ Identify multiple forms of abuse
- ▶ Encode subjectivity into data
- ▶ Encode ideological positions into data

- ▶ Model

- ▶ Operate on the assumption of pre-existing equality
- ▶ Identify correlations between tokens and labels
- ▶ Unable to distinguish forms
- ▶ Objective functions don't reward subjectivity
- ▶ Encode simplified ideological positions into models



MODELING NORMATIVITY

- ▶ Modeling normative values
- ▶ The God trick

DESIGNING NORMATIVITY

- ▶ Data sources
- ▶ Context
- ▶ Label Selection
- ▶ Annotation guidelines
- ▶ Annotation processes

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. “#BanIslam”, “#whoriental”, “#whitegenocide”
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

MODELING NORMATIVITY

- ▶ Linear
- ▶ Non-linear
- ▶ Word Embeddings
- ▶ Contextual embeddings



REMEDIES

- ▶ Deciding on norms
 - ▶ Individual
 - ▶ Community
 - ▶ Should any community be imposed norms?

