A Survey on Hate Speech Detection using Natural Language Processing

detekce nenávistných promluv

Sémantika a počítače

Marie Novotná

9. 5. 2017

Motivace

- neustálý růst obsahu sociálních médií
- svoboda slova
- online projevy nenávisti
- detekce sprostých slov, nevhodných obrázků, …

Hate speech

- jakákoliv komunikace, která znevažuje osobu nebo skupinu lidí na základě určitých vlastností:
  - rasa,
  - barva,
  - etnická příslušnost,
  - pohlaví,
  - sexuální orientace,
  - národnost,
  - náboženství,
  - …

* abusive messages, hostile messages, flames, cyberbullying

Hate speech

- jakákoliv komunikace, která znevažuje osobu nebo skupinu lidí na základě určitých vlastností:
  - rasa,
  - barva,
  - etnická příslušnost,
  - pohlaví,
  - sexuální orientace,
  - národnost,
  - náboženství,
  - …

*„Go fuc*ing kill yourself and die already useless ugly pile of shit scumbag!“*

Problém

- automatická detekce těchto projevů
- ruční anotace korpusů
- základní slovní filtry nefungují

Funkce pro rozpoznávání nenávistných promluv

- jednoduché povrchové funkce
  - (znakové) n-gramové modely *ki11 yrslef a$$hole*
- zobecňování slov
  - Brown clustering
- analýza sentimentu
  - detekce negativní polarity (SentiStrength)
- lexikální zdroje
- jazykové aspekty
- znalostní báze *Nasaďte paruku a rtěnku a buďte tím, čím jste opravdu.*
- metainformace (muži, recidivisti)
- multimodální informace

Děkuji za pozornost

- http://aclweb.org/anthology/W/W17/W17-1101.pdf
- http://www.iurium.cz/2016/09/28/nenavistne-projevy-facebooku/