Ugh... ML

Anna, Annie, Jake, J  sé

Review of Task Goals and Spec

- Goal: Annotate hate speech in Twitter posts, with the goal of training a machine learning algorithm to detect it independently
- Difficulties: Slurs, Seemingly innocuous/sounding discriminatory…
- Annotation Spec and tags:
- 9 attributes for the Group tag
- Target
- Slur
- Sentiment and Stereotype; Refs

Characteristics of Dataset

- Collected from a Twitter feed
- @hatespeechMLA account
- Twython Python wrapper for the Twitter API
- Collecting live tweets based on key terms and scraping timelines
- 400 tweets

Difficulties Collecting Data

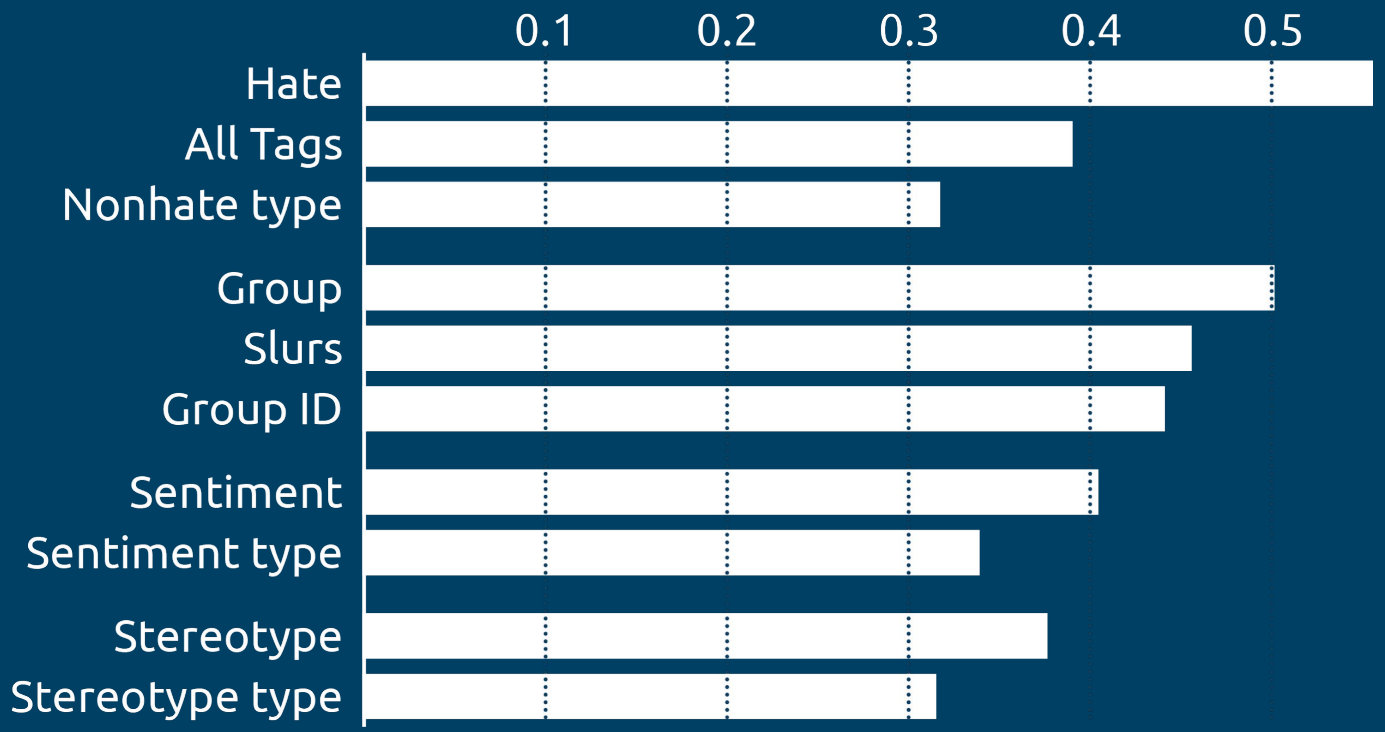
- Week 1
  - Unfinished (Avg 33 of 100 tweets)
  - Reading guidelines & familiarizing oneself with the task
- Week 2
  - MAE lag (attempting to annotate full 100)
  - 1, 8, & 10 files
  - Required manual transferring to 10 file format
- Weeks 3 & 4
  - More smoothly
  - 10 files each of 10 tweets
  - Most bugs were fixable with simple scripts

Postprocessing

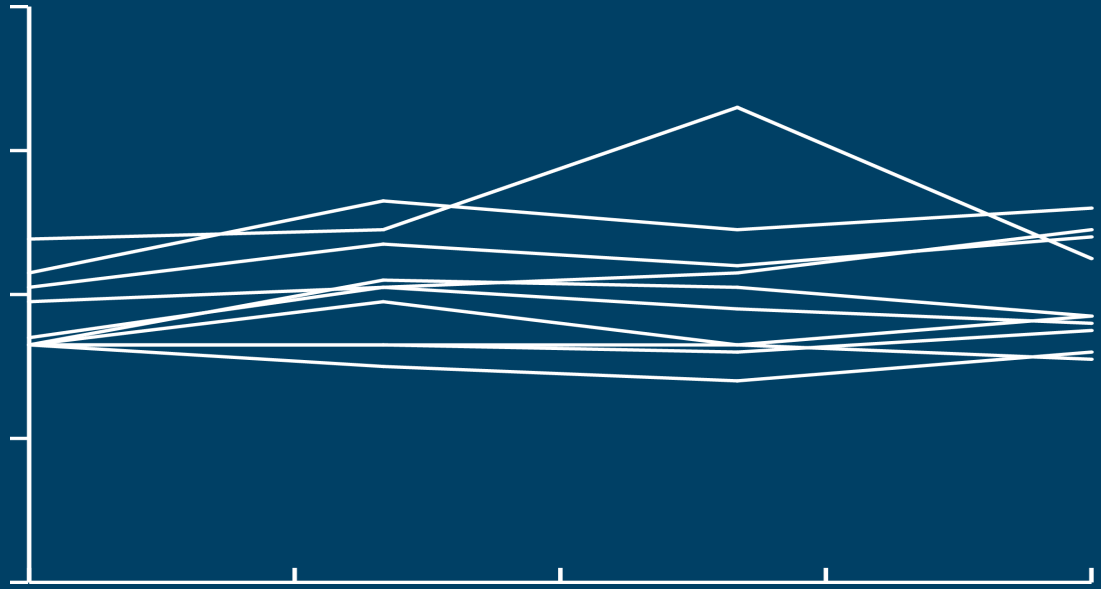
- A few minor edits
  - spans issues such as inclusion of spaces or punctuation
- Link tag direction
  - Guidelines: `from` Stereotype/Sentiment `to` Group
  - Many link tags went in the other direction
    - `from` “play loud shitty music” `to` “Mexicans”
    - `from` “Mexicans” `to` “play loud shitty music”
  - Showed up in MAE as not agreeing during adjudication
  - Simple script to reverse the two if the `fromID` was a `Group` tag

Inter-Annotator Agreement

- Fleiss's κ
- Overall
  0.56



IAA Week by Week



Machine Learning Experiment

- Our machine learning experiment is still in progress
- Our baseline will be a Naïve Bayesian classifier with the individual words as features
- We intend to test two algorithms that use our annotations: a Bayesian algorithm and a Maxent algorithm