Refined Cyberbullying Representation for Machine Learning Classification

Caleb Ziems
Emory University

Fred Morstatter
USC Information Sciences Institute

1. Problem Statement

Automatic cyberbullying detection methods are unfit for real-world applications [3]. This is largely due to:

- **Unreliable data:** inconsistent criteria, [2,3,4] context-blind annotations, [3] class imbalance [2,3]
- **Coarse features:** bag-of-words (BoW) methods lack nuance and cannot adapt to language change

Goal 1: Produce a reliable dataset of labeled cyberbullying cases within Twitter threads
Goal 2: Train a cyberbullying classifier from a refined set of social features

2. Data Collection

- **Scrape:** 1.3 million tweets from Stream API
- **Filter:** English, @ mentions, non RTs, visible threads, hate speech / offensive language [1]
  - 6,897 message threads
- **Collect user data:** account information (friends, following) and 6 months of each timeline

3. Annotation Task

MTurk study: 3 annotations per message thread

- Label *author* & *target* @handles for each tweet
- Given the full message thread and up to 15 recent mentions, provide labels for 5 criteria
  1) **Aggressive language:** confrontational, derogatory, insulting, threatening, hostile, violent, hateful, or sexually abusive language directed towards individual or group [2,3,5]
  2) **Repetition:** 2+ aggressive messages [2,3,4]
  3) **Harmful intent:** author intends to tear down or disadvantage the target user [3,4,5]
  4) **Visibility among peers:** one other user has liked, quoted, retweeted or responded to the author [3]
  5) **Power Imbalance:** does the author or target have greater social advantage / perceived authority? [2,4]

Criterion	Class Balance	Inter-annotator Agreement	Cyberbullying Correlation
aggression	74.8%	0.23	0.68
repetition	6.6%	0.18	0.27
harmful intent	16.1%	0.42	0.22
visibility among peers	30.1%	0.51	0.07
target power	78.9%	0.37	0.11
author power	3.1%	0.10	-0.02
equal power	59.7%	0.22	-0.09
cyberbullying	0.7%	0.18	-

- **Advantages:** clear criteria, flexible cyberbullying definition, context-aware annotations, more balanced class distributions

4. Feature Engineering

Baseline Features

- **Text:** N-Grams, LIWC, VADER, Flesch-Kinkaid Reading Ease [1,5]
- **User:** Friend/following counts, verified status, number of posts

Thread Features

- **Visibility**
  - Message count, reply message count, reply user count, max author favorites, max author RTs
- **Aggression**
  - Aggressive message count, aggressive author message count, aggressive user count [1]

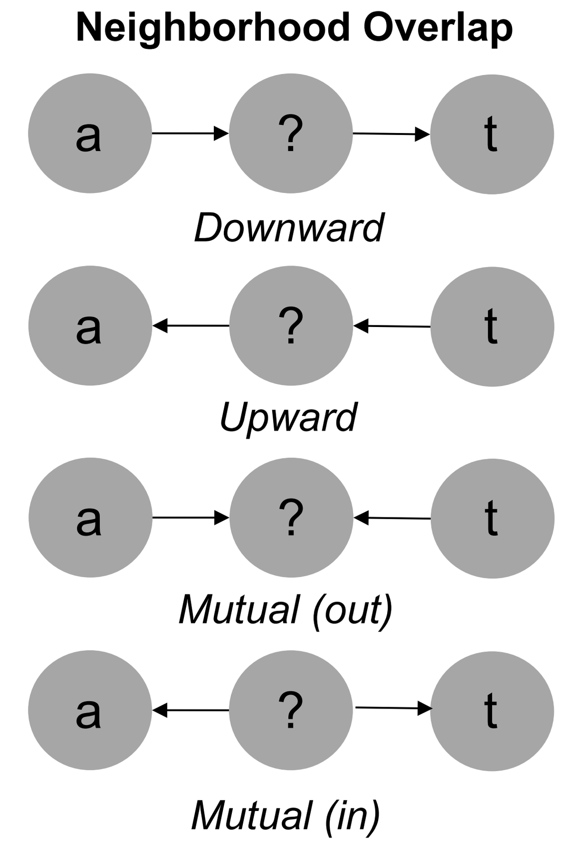
Timeline Features

- **Message Behavior**
  - Directed message counts
  - Mentions overlap (Jaccard)
- **Language Models**
  - New-words ratio
  - Cross-entropy
    - $H(m) = -\frac{1}{N}\sum_i \log P(b_i)$
      for message m with bigrams $b_1, b_2, \ldots, b_N$
- **Timeline similarity**
  - $\cos\theta = \frac{\vec{A}\cdot\vec{T}}{\|\vec{A}\|\|\vec{T}\|}$
    for author timeline $\vec{A}$ and target timeline $\vec{T}$

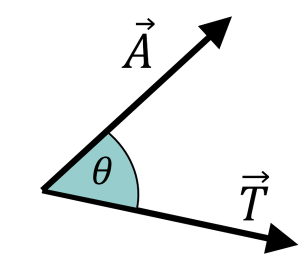
Network Features

- **Neighborhood Overlap**
  - $\mathrm{JC} = \frac{|N(a) \cap N(t)|}{|N(a) \cup N(t)|}$
    for author a and target t,
    $N(u)$ is the neighborhood set of user u

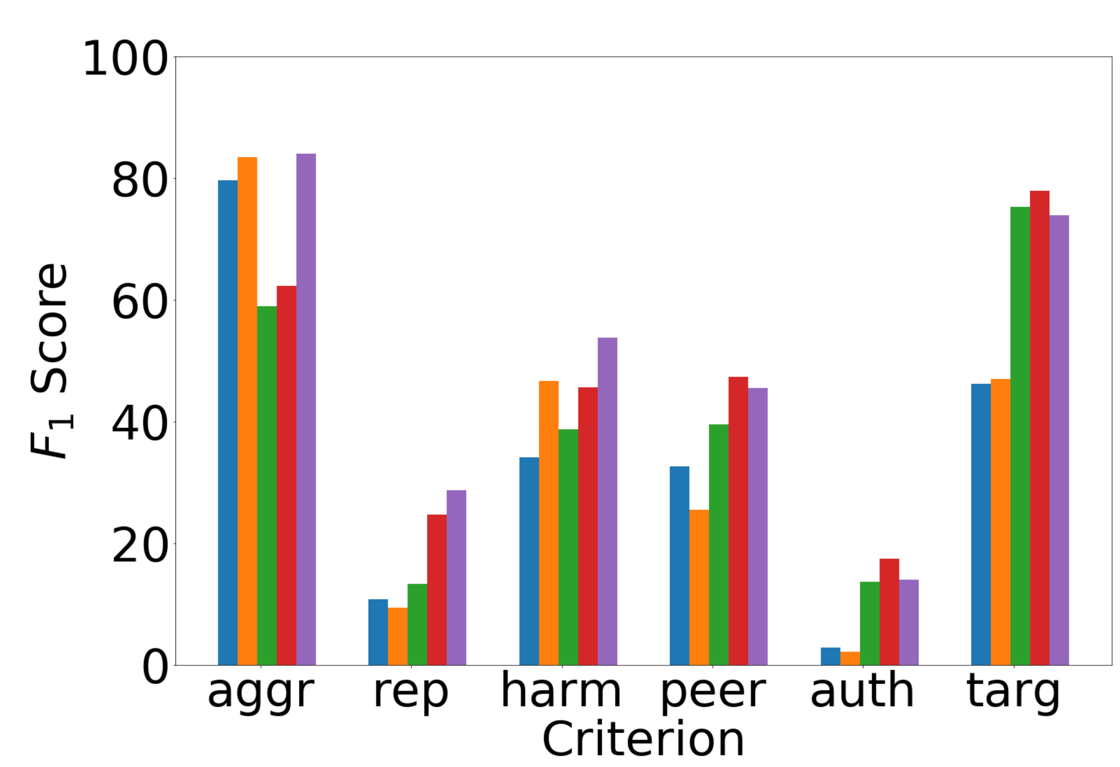
Neighborhood Overlap



Timeline Similarity



5. Model Evaluation



BoW: unigrams, bigrams, trigrams
Text: n-grams, LIWC, VADER, Flesch-Kincaid
User: social network features, timeline features
Proposed: social network features, timeline features, thread features
Combined: all features

6. Conclusions

- **Text-based** methods can reliably detect *aggressive language*
- **Social** features are better suited for detecting *repetition, visibility among peers*, and *power imbalance*
- **Classifiers** are not yet ready for the real world [3,4]
- **Future Work:** increase performance, build new features, detect social roles, measure efficiency (run time, number of API calls, etc.)

References

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

[2] Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

[3] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.

[4] Salawu, S., He, Y., & Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*.

[5] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794.

If interested contact **Caleb Ziems** <*cziems@emory.edu*>
Project web page: *reu.isi.edu*

Work performed under REU Site program supported by NSF grant #1659886