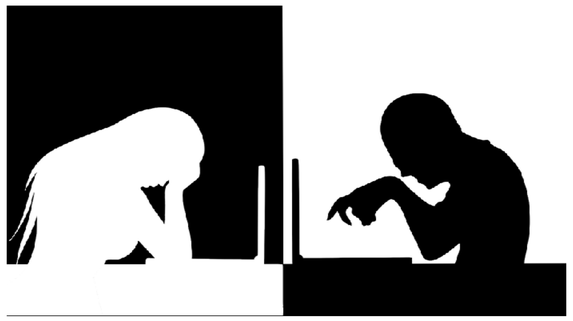
Detecting Nastiness in Social Media

Niloofar Safi Samghabadi[1], Suraj Maharjan[1], Alan Sprague[2], Raquel Diaz-Sprague[2], and Thamar Solorio[1]

[1]Department of Computer Science
University of Houston
{nsafisamghabadi, smaharjan2}@uh.edu solorio@cs.uh.edu

[1]Department of Computer and Information Science
University of Alabama at Birmingham
sprague@cis.uab.edu diazspra@uab.edu

Motivation



Cyberbullying:

- Cyberbullying is bullying that takes place using electronic technology. It affects mostly teens.
- 26.3 % high school and middle school students have been cyberbullied – 16% have cyberbullied.

Methodology

Data:

- Our goal is to detect highly negative posts.
- We focus on teens by using the platform that is popular among them.
- We use data contain profanity to increase the chance of finding highly negative posts.

Features:

- Lexical (word n-gram, character n-gram, k-skip n-gram)
- POS Colored n-gram
- Sentiment (SentiWordNet)
- Domain (Question-Answer)
- Emoticon
- LIWC
- Writing Density
- Hand-crafted (Patterns)
- Embedding (W2V, D2V)
- Topic Modeling (LDA)

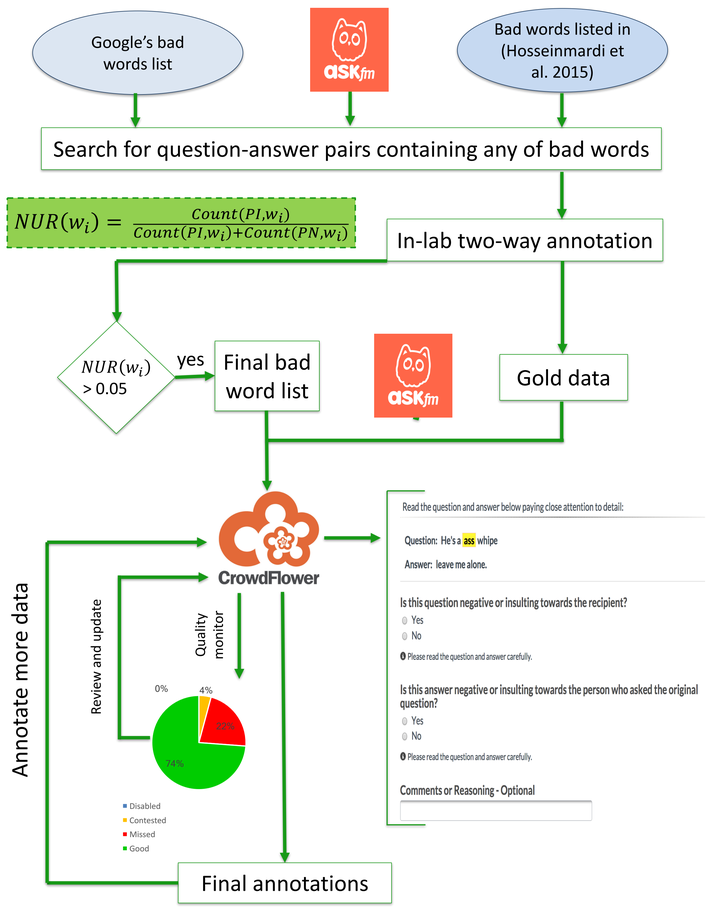
Classification:

- We follow a machine learning approach using a LinearSVM classifier.
- We tune classifier C parameter with a grid search over {1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100, 1000, 10000}.

Data Collection and Annotation

Original Data:

- 586K question-answer pairs from 1,954 random users of Ask.fm from 28^{th} January - 14^{th} February, 2015.



- Inter-annotation agreement kappa score is 0.453.

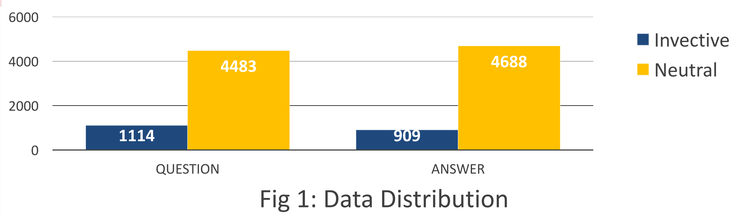


Fig 1: Data Distribution

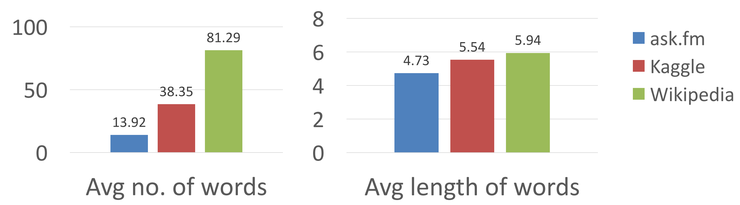


Fig 2: Average length of posts and words

Results

	Ask.fm		Kaggle		Wikipedia	
Feature	**AUC**	**F-score**	**AUC**	**F-score**	**AUC**	**F-score**
Baseline	0.567	0.27	0.597	0.36	0.610	0.28
Unigram (U)	0.768	0.57	**0.813**	0.71	0.882	0.72
Char 4gram (C4)	0.748	0.56	0.812	0.72	0.879	0.73
CT + C4 + C5	0.734	0.55	0.811	**0.73**	0.866	**0.75**
SentiWordNet (SWN)	0.602	0.35	0.575	0.39	0.632	0.30
LIWC	0.662	0.42	0.715	0.57	0.787	0.53
Writing Density (WR)	0.564	0.30	0.566	0.42	0.682	0.31
Word2vec (W2V)	0.745	0.51	0.759	0.63	0.854	0.61
Doc2vec (D2V)	0.750	0.52	0.792	0.66	0.886	0.60
LDA	0.626	0.37	0.559	0.40	0.577	0.26
LIWC + E + SWN + W2V + D2V	0.780	0.56	0.799	0.68	**0.889**	0.65
U + C4 + QA + LIWC + E + SWN + W2V + D2V	**0.785**	0.57	N/A	N/A	N/A	N/A
C4 + U + QA + E	0.766	**0.59**	N/A	N/A	N/A	N/A
All Features	0.756	0.56	0.798	0.71	0.882	**0.75**

Table1: Classification results for invective class

- The most challenging instances are:
  - Single profane word answers
  - Question and answer pairs in which users joke around with use of foul words
  - Posts with mixture of politeness and profanity
  - Post with bad words that are offered as compliments

Negativity of Words

For post with single profane word: $NUR(w_i) = \frac{Count(PI,w_i)}{Count(PI,w_i)+Count(PN,w_i)}$

bad word	negativity
as**ole	51.16%
kill	12.47%
f*ck	33.05%
n**ger	13.30%
sh*t	15.23%
cut	4.85%

bad word	Negativity
b*tch	41.65%
a*s	24.77%
die	7.41%
s*ck	26.88%
h*e	36.58%
stfu	51.55%

Table 2: Degree of negativity for bad words

Conclusion

- Our model can be successfully applied to other datasets.
- It seems it is much harder to detect nastiness in shorter texts.
- Analyzing the degree of negativity for bad words reflects a sexualized teen culture.