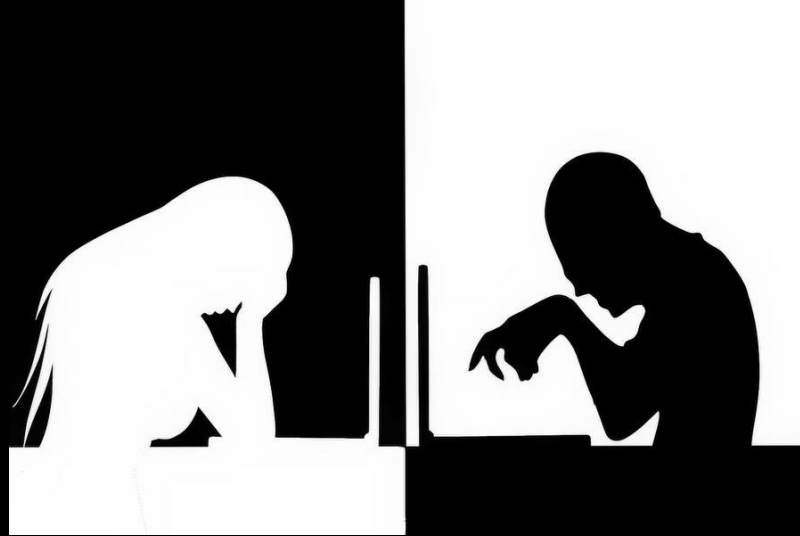




# Improving *Perspective* API Granularity With NBSVM

**Presenter:**  
*Francisco McGee*  
*Data Scientist*

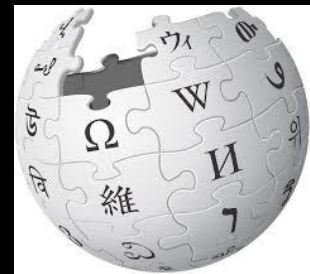
# Has anyone here ever been harassed online?



54% reported decreased participation

Harassment Survey, Wikipedia.org. 2015

# Toxicity affects your bottom line





Jigsaw

**How can technology  
make the world safer?**



# How can data science maximize profit?

- Toxic comments → warning/ban: 17.9%
- 9% of “insult” attacks in 2015 from 34 users
- Mitigate trolling
- Identify trolls by their comments
- “Sentiment analysis” via NBSVM
- Naive Bayes Support Vector Machines

Harassment Survey, Wikipedia.org. 2015

# Dataset:

## *Wikipedia talk page edits*



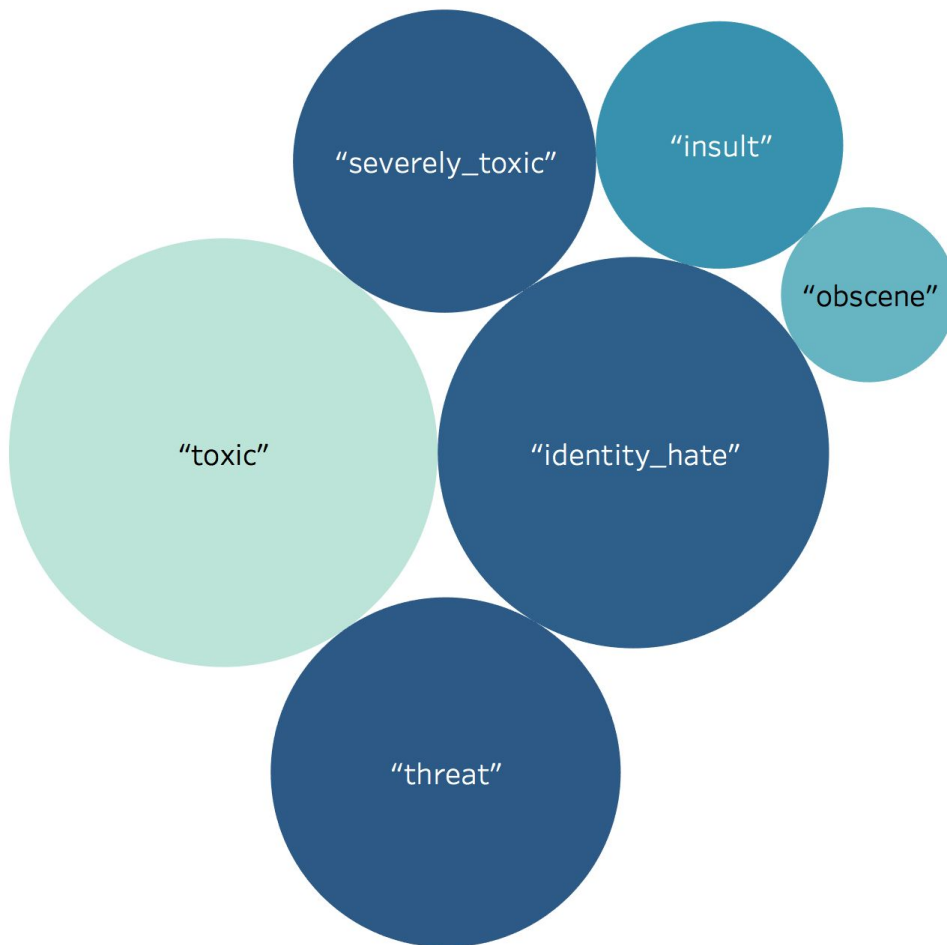
- Online comments generally “short snippets”
- Categories: Toxic, severely toxic, obscene, threat, insult, identity hate
- Crowd-sourced determination

# Examples

- Toxic: “Nonsense? kiss off, geek. what I said is true. I'll have your account terminated.”
- Toxic/Obscene/Insult: Ban one side of an argument by a bullshit nazi admin and you get no discussion because the islamist editors feel they ""won""."

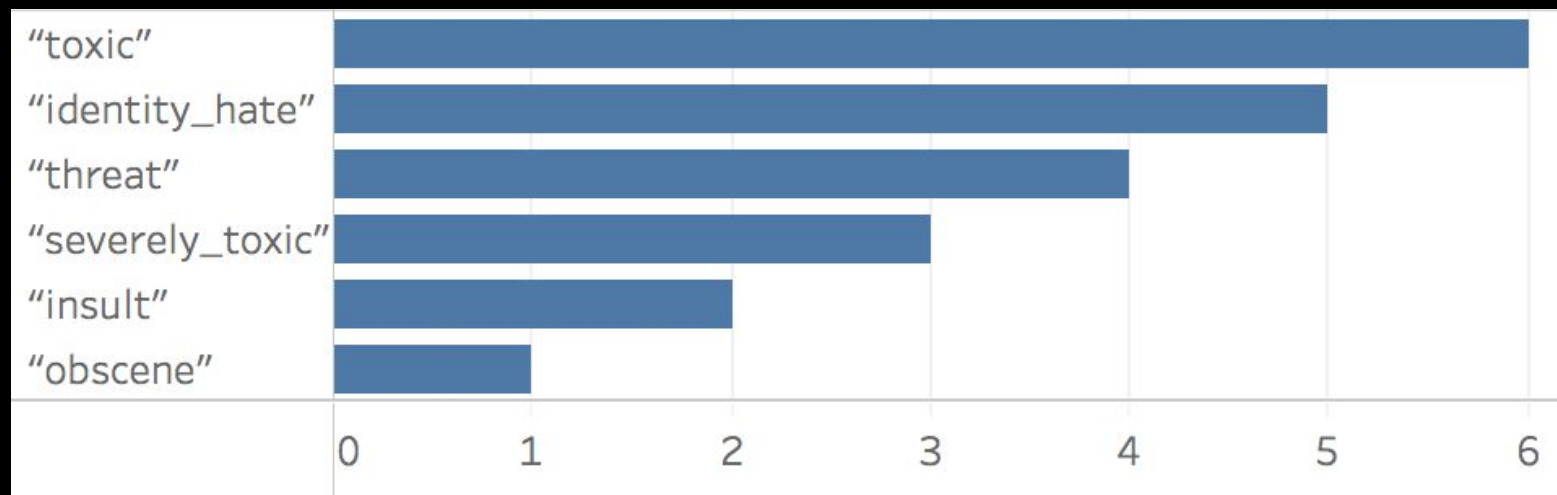
# Preliminary Results

NBSVM Accuracy By Toxicity Category





# Preliminary Results



# Challenges/Limitations

- Scalable code
  - Had to halve dataset
- Validation cutoff: 0.9
  - Floored/ceiled to binarize toxicity calls
  - Excludes much of the ambiguous data

# Next Steps

- Scale code
- Explore categorical dependence
- Tune validation cut-offs
- Benchmark other kernels

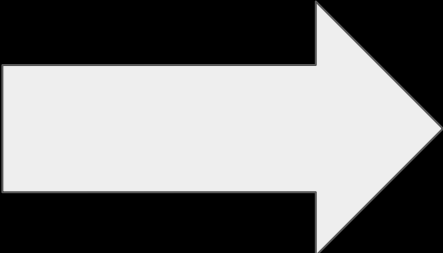
# NBSVM aka “bag of words”

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.



```
[  
  "John",  
  "likes",  
  "to",  
  "watch",  
  "movies",  
  "Mary",  
  "too",  
  "also",  
  "football",  
  "games"  
]
```



```
(1) [1, 2, 1, 1, 2, 1, 1, 0, 0, 0]  
(2) [1, 1, 1, 1, 0, 0, 0, 1, 1, 1]
```