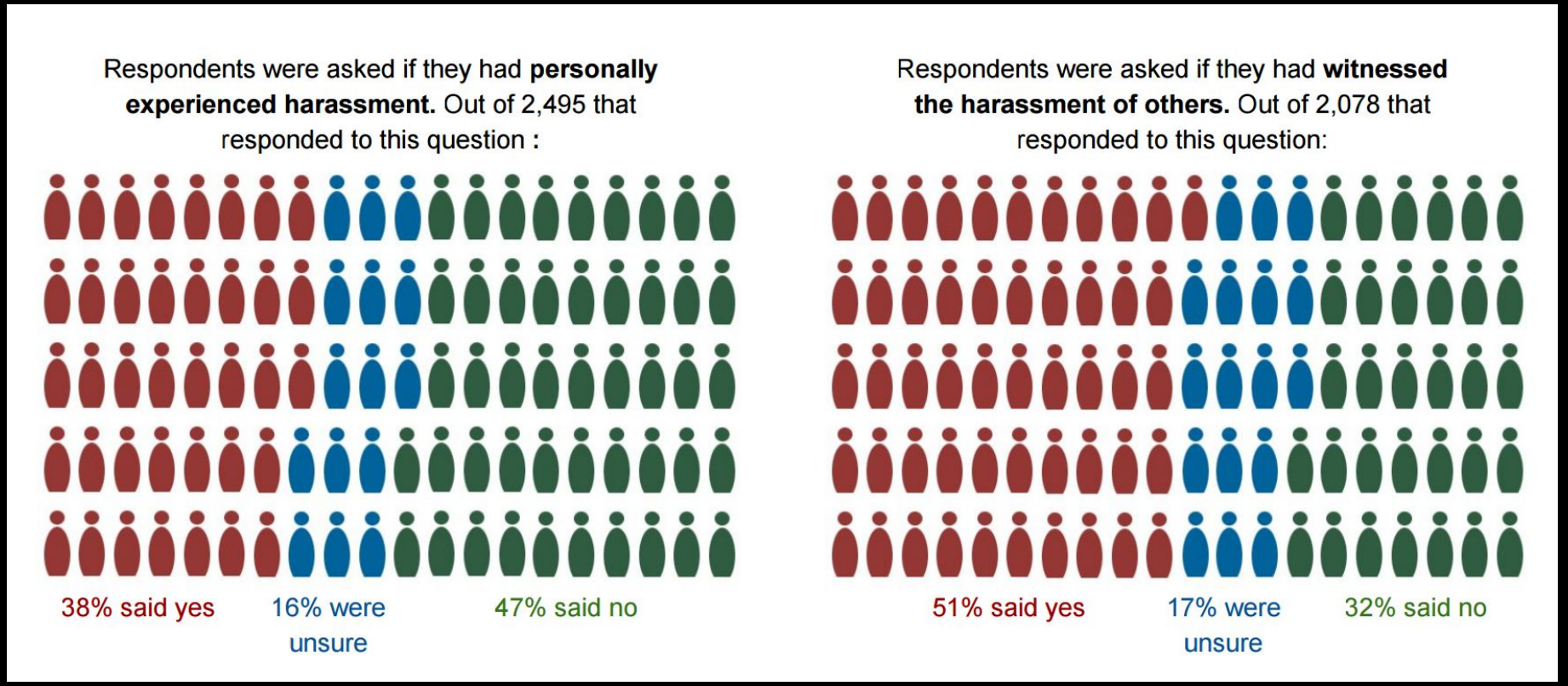
Understanding Personal Attacks on Wikipedia

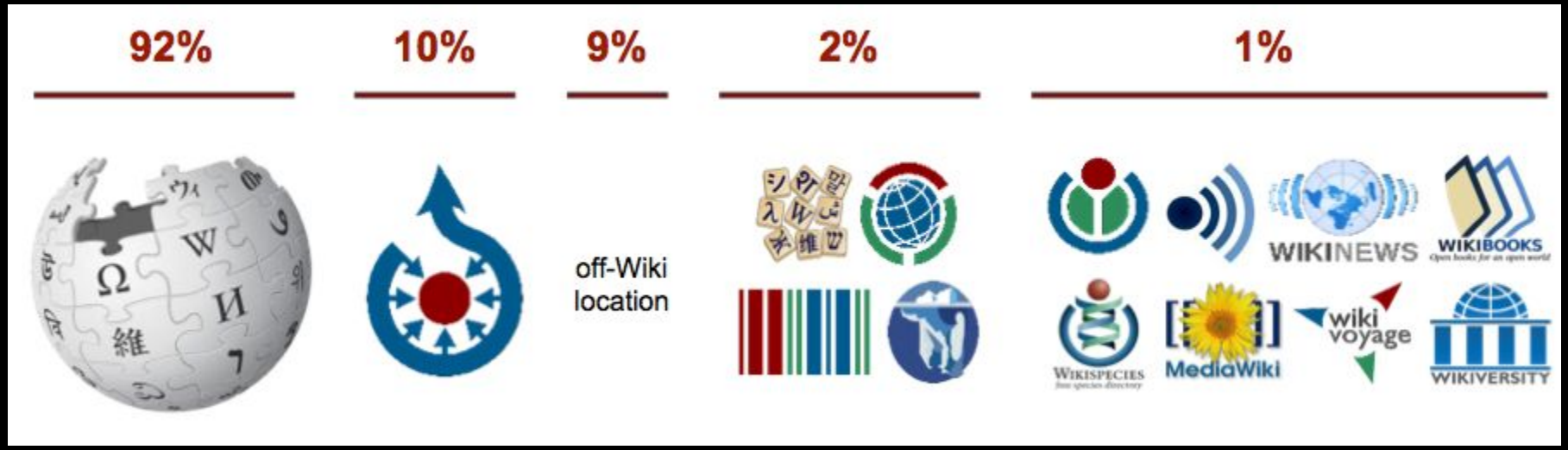
Ellery Wulczyn and Nithum Thain

Harassment is prevalent on wikimedia projects



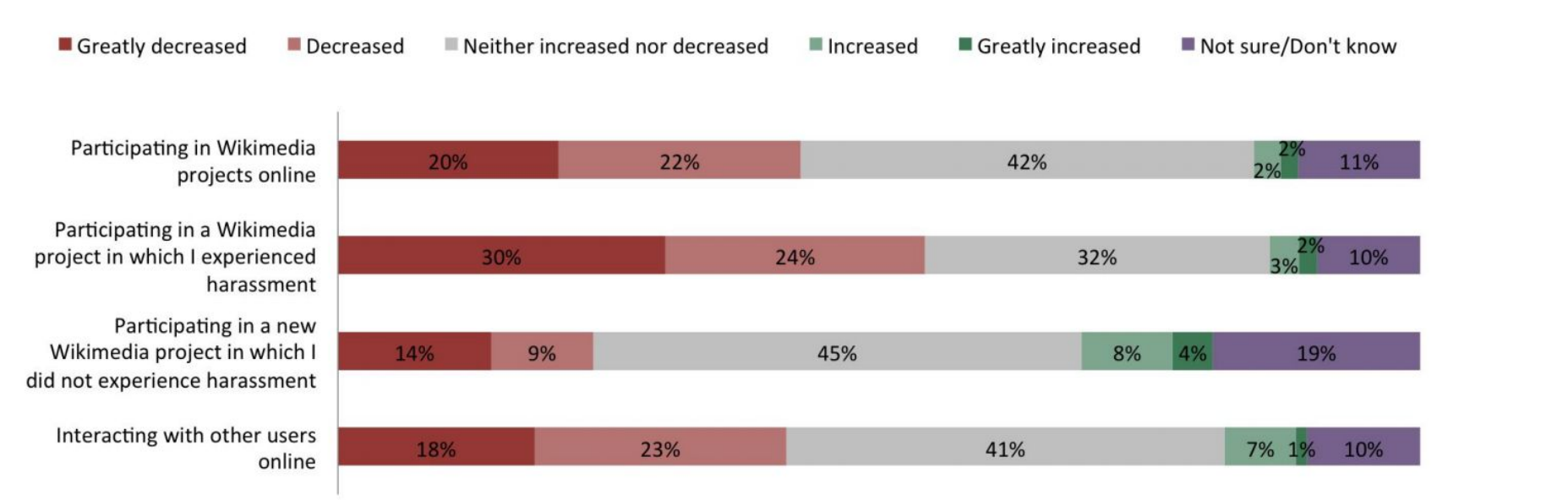
Source: The Harassment Survey 2015

Most harassment occurs on Wikipedia



Source: The Harassment Survey 2015

Victims of harassment are less likely to contribute to Wikimedia projects



Source: The Harassment Survey 2015

Goals

1. Develop an algorithmic approach to detect personal attacks on Wikipedia

2. Use these algorithms to extend the analysis of personal attacks on Wikipedia

Outline

1. Data Pipeline
2. Model Building
3. Analysis

Outline

1. Data Pipeline
2. Model Building
3. Analysis

Data Pipeline

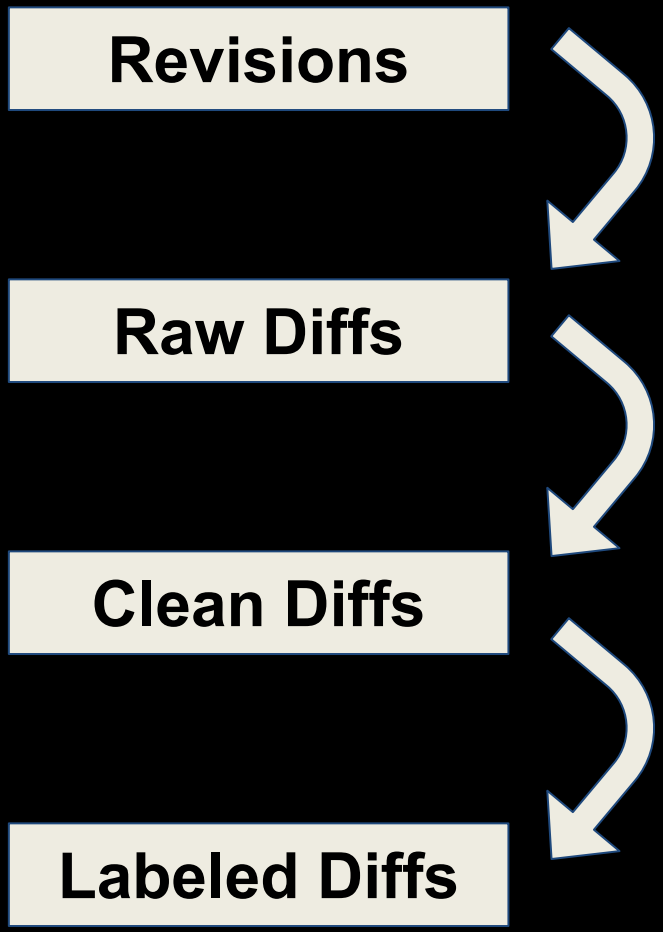
Goal:

Set of labeled talk page comments

Input:

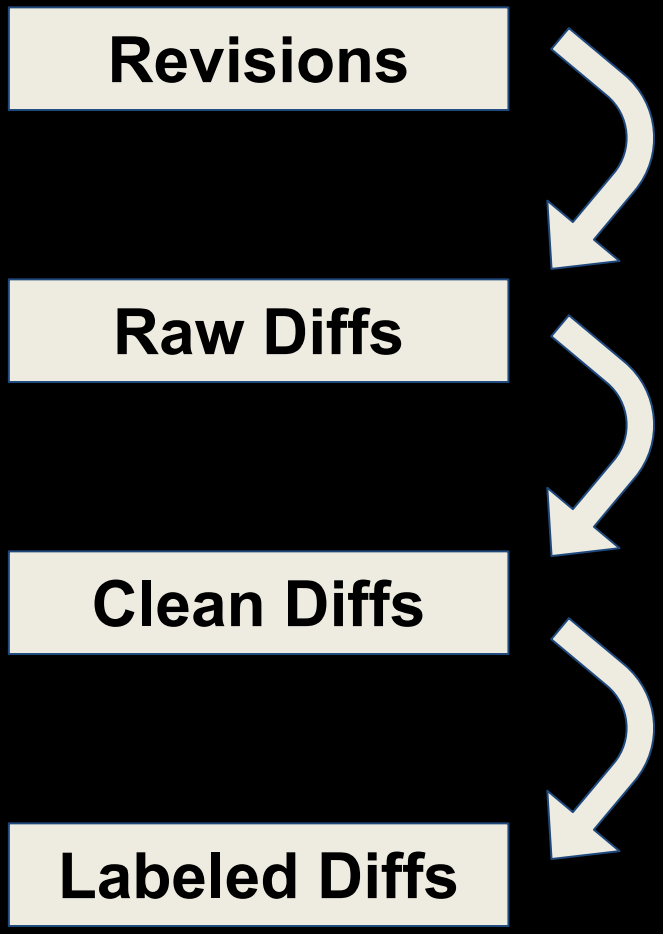
English Wikipedia revision history

Data Pipeline



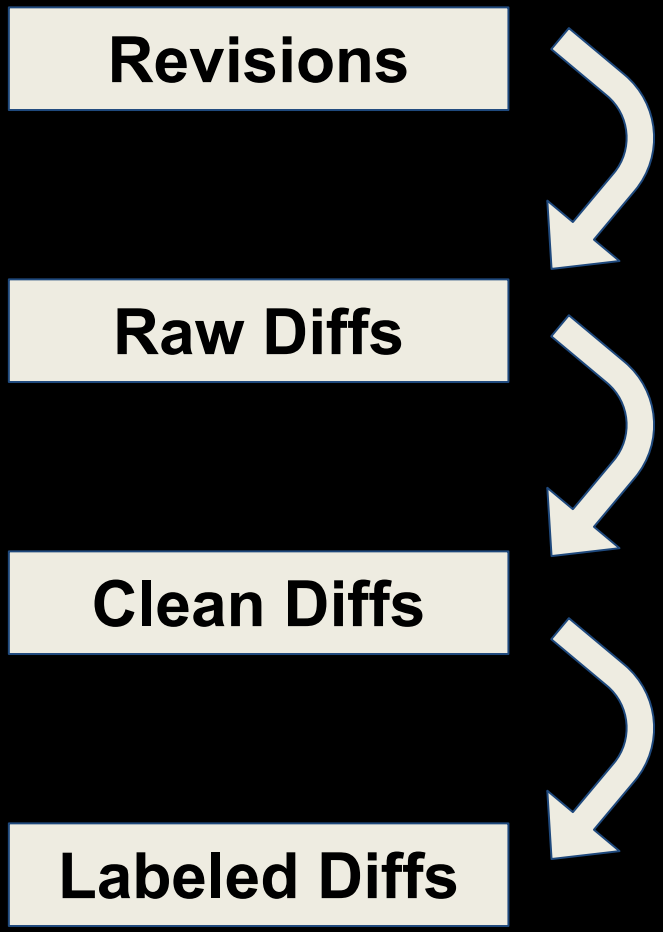
mwdiffs python library

Data Pipeline



- extract content added
- remove mw markup, etc
- filter out administrative messages

Data Pipeline



crowdsourced labeling via **CrowdFlower**

Labelled Training Data

Random Data

A representative sample of revisions from article and user talk pages

- Correct prior distribution
- Important for validation
- Few examples of attack

Blocked Data

A sample of revisions written by a user near a “block event” for personal attacks

- High proportion of attacking comments
- Speeds up training

Choosing a Question

Does the comment contain a personal attack or harassment? Please mark all that apply:

- *Targeted at the recipient of the message*
- *Targeted at a third party*
- *Being reported or quoted*
- *Another kind of harassment*
- *This is not an attack or harassment*

Crowdsourced Annotation

- Crowdflower platform
- 20,000 random revisions
- 50,000 blocked revisions
- Each rated 10x
- Quality control via test questions

Crowdflower Challenges

- Annotators working quickly
- May have imperfect knowledge of English
- Subjective nature of task

Outline

1. Data Pipeline
2. Model Building
3. Analysis

Model Building

Goal:

build classifier that takes in a talk page comment and outputs the probability that the comment contains a personal attack

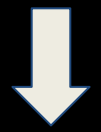
Input:

70k comments, each annotated 10x

Model Building: ML Overview

collection of comments + annotations

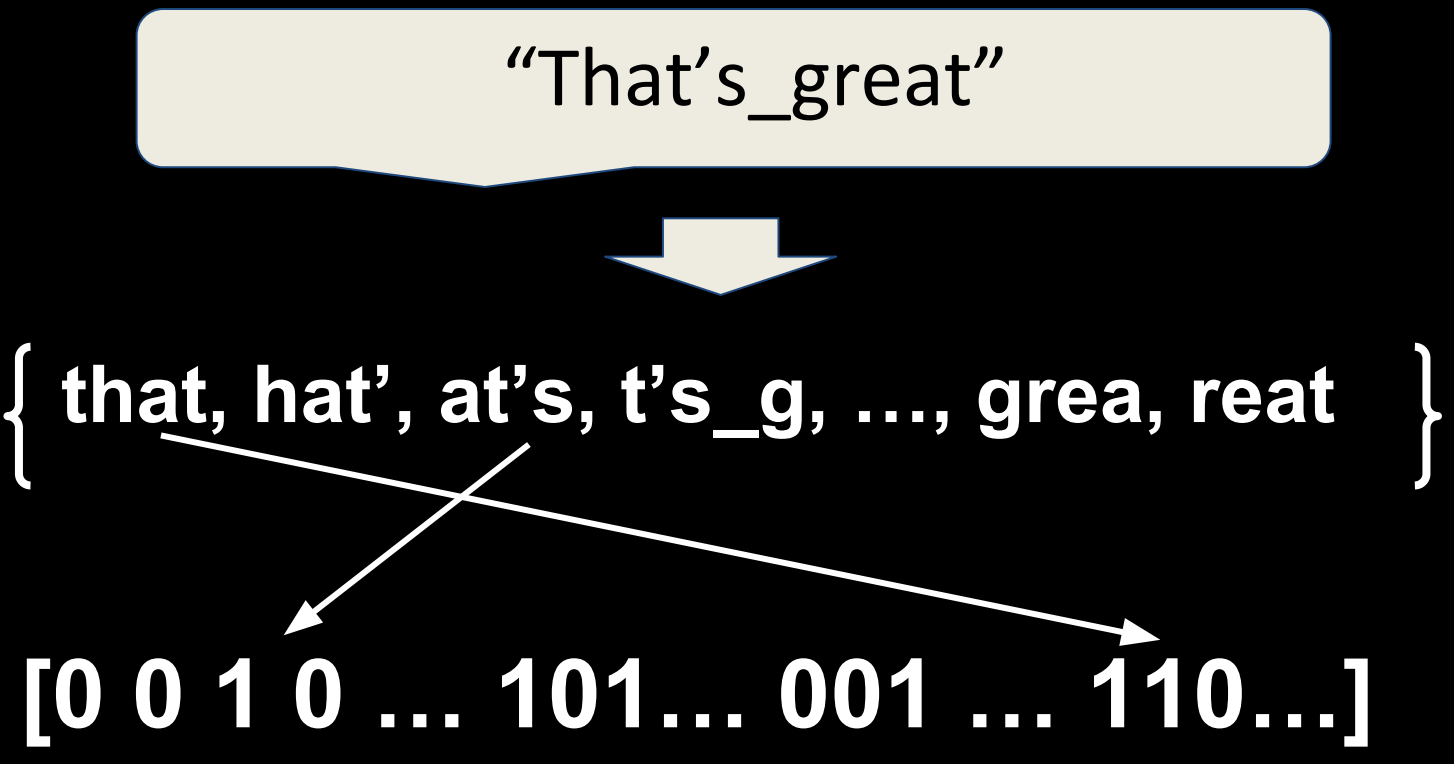
collection of features + labels



learning algorithm

classifier

Model Building:
From Comments to Features



Model Building: From Annotations to Labels

Fraction of annotators who thought the comment **is** a personal attack

0.7

Model Building: Learning Algorithms

Final Choice: Logistic Regression

Experimented with: MPLs, RNNs, CNNs:

added complexity, little performance gain

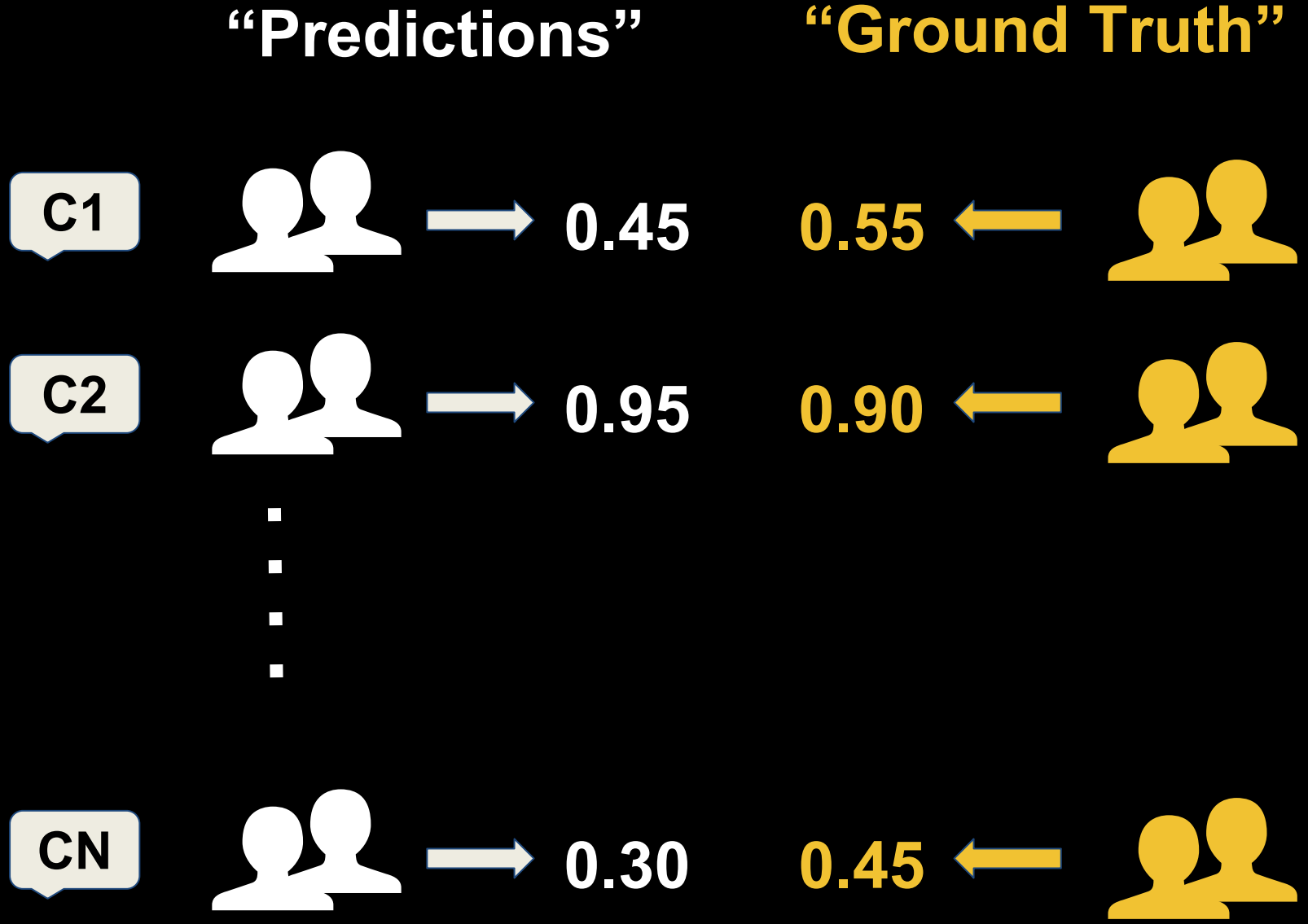
Model Building: Evaluation

Question:

How good is our classifier/model?

Idea:

Use one group of people to predict what another group of people thinks about a comment. Compare our model's predictive power, to the predictive power of a group of people.


"Predictions"
"Ground Truth"
C1
0.45
0.55
C2
0.95
0.90
CN
0.30
0.45

Model Building: Evaluation

Fix “Ground Truth group size at size 10

Prediction Group Size	ROC AUC
1	0.854
2	0.911
4	0.941
6	0.950
8	0.961
10	0.963

Model Building: Evaluation

Fix “Ground Truth group size at size 10

Prediction Group Size	ROC AUC
1	0.854
2	0.911
4	0.941
6	0.950
8	0.961
10	0.963

Model: **0.951**

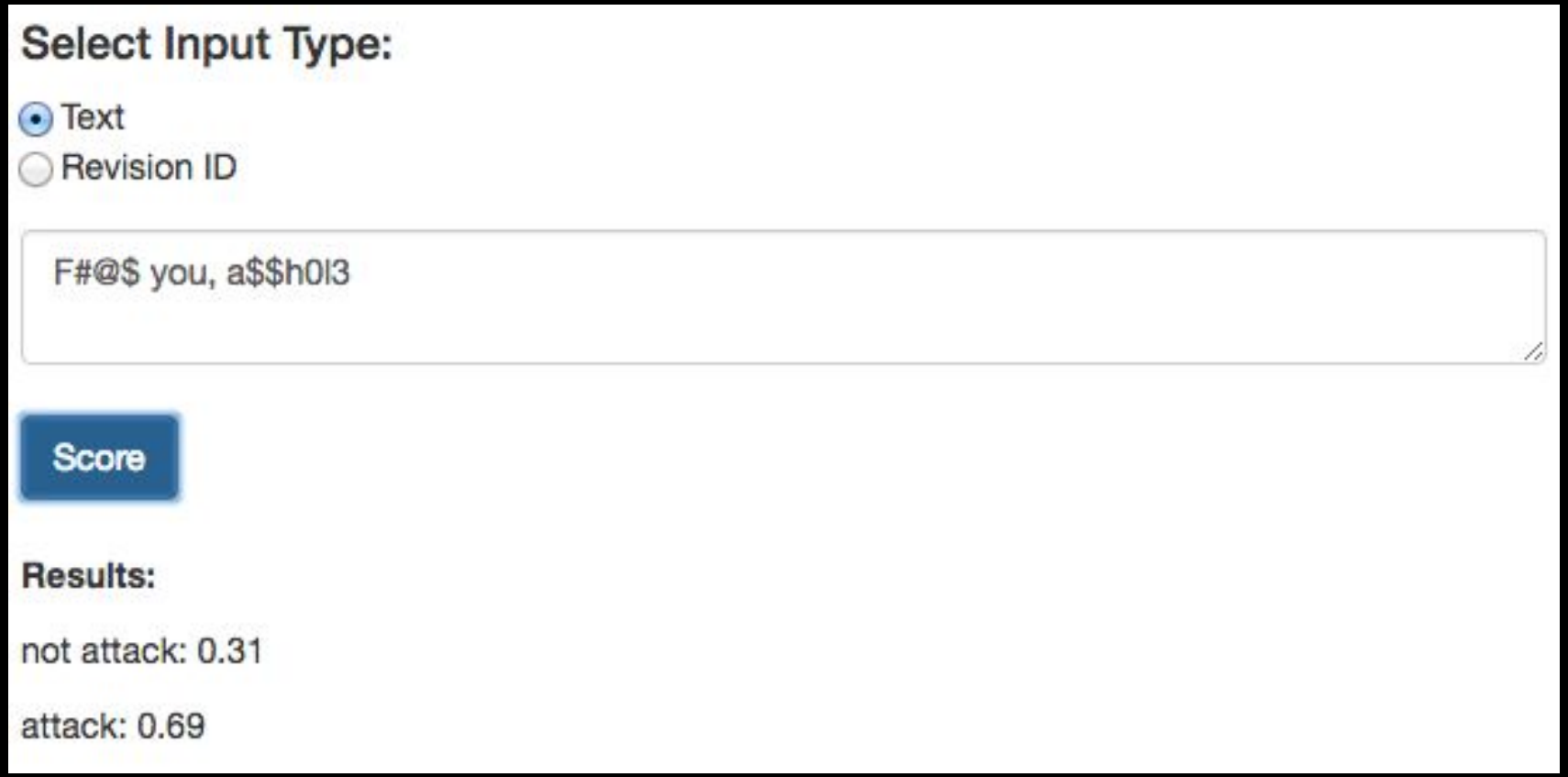
Demo

Available at: *wikidetox.appspot.com*

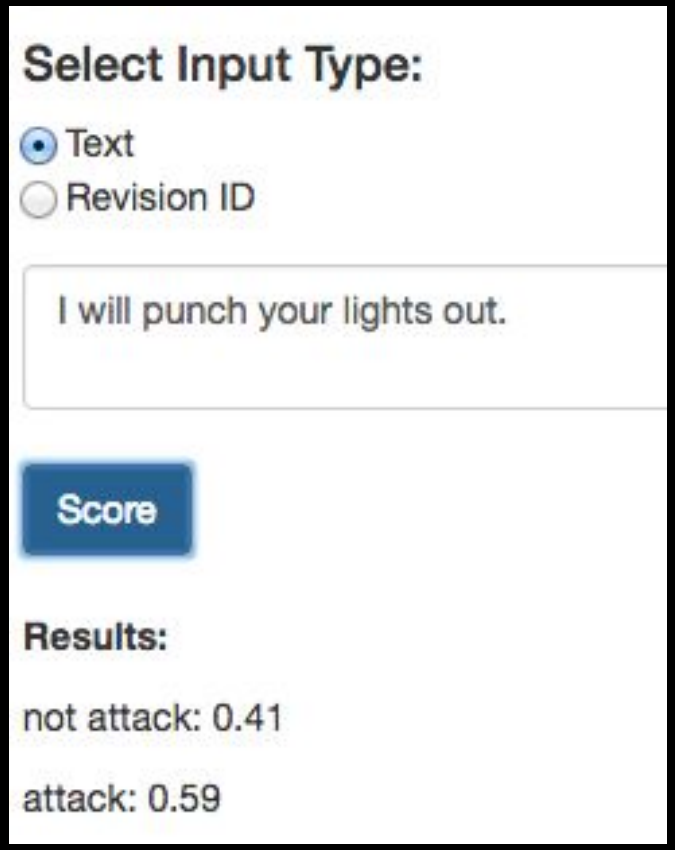
Demo


Select Input Type:
Text
Revision ID
Congratulations. I don't know whether you are aware of this fact or not, but you have shown your qualified stupidity.
Score
Results:
not attack: 0.18
attack: 0.82

Demo


Select Input Type:
Text
Revision ID
F#@$ you, a$$h0l3
Score
Results:
not attack: 0.31
attack: 0.69

Demo

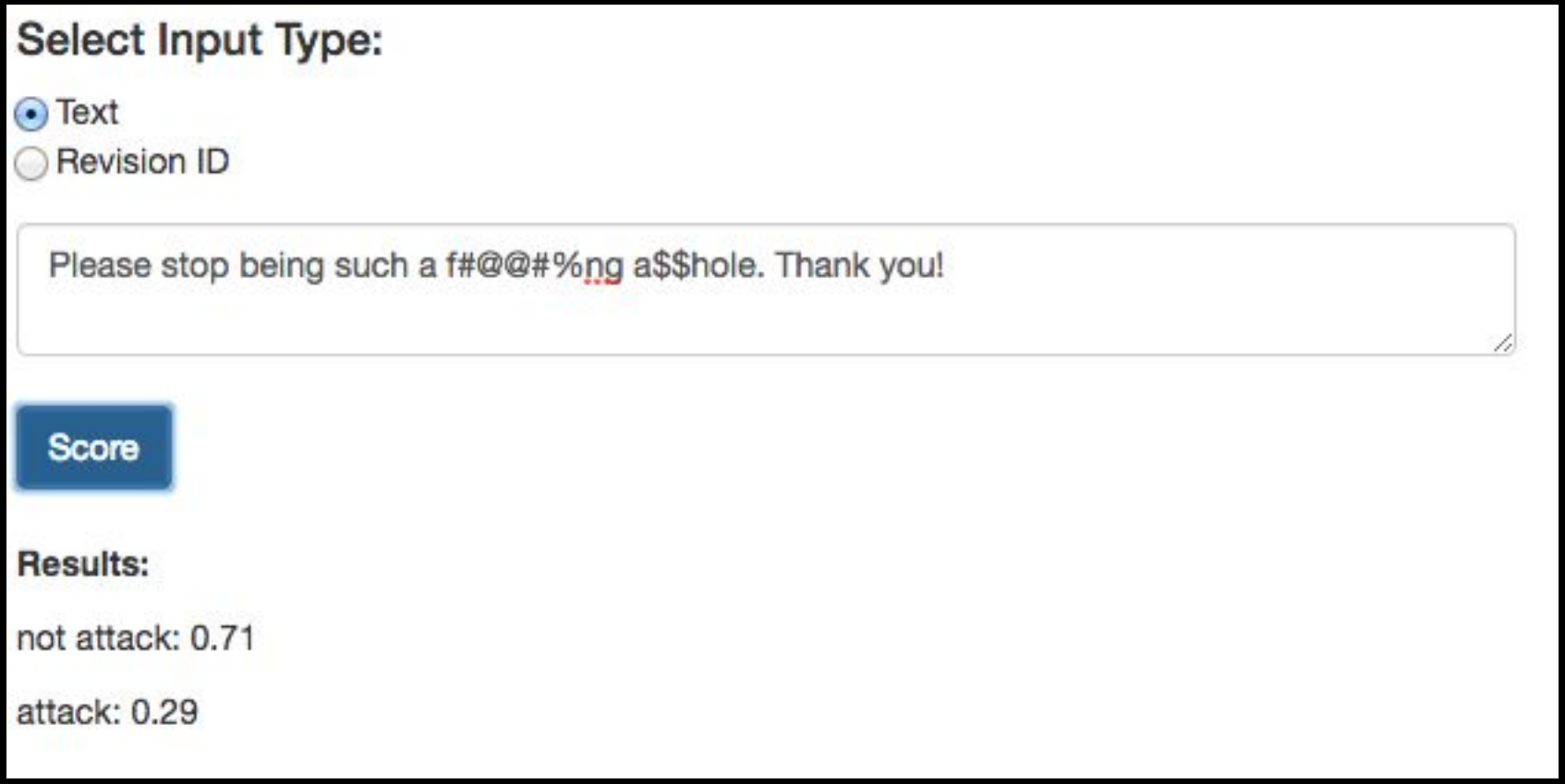

Select Input Type:
Text
Revision ID
I will punch your lights out.
Score
Results:
not attack: 0.41
attack: 0.59


Select Input Type:
Text
Revision ID
Let's drink punch.
Score
Results:
not attack: 0.83
attack: 0.17

Demo


Select Input Type:
Text
Revision ID
Your intellect is lacking
Score
Results:
not attack: 0.90
attack: 0.10

Demo


Select Input Type:
Text
Revision ID
Please stop being such a f#@@#%ng a$$hole. Thank you!
Score
Results:
not attack: 0.71
attack: 0.29

Demo


Select Input Type:
Text
Revision ID
p i s s off!
Score
Results:
not attack: 0.78
attack: 0.22

Outline

1. Data Pipeline
2. Model Building
3. Analysis

Analysis

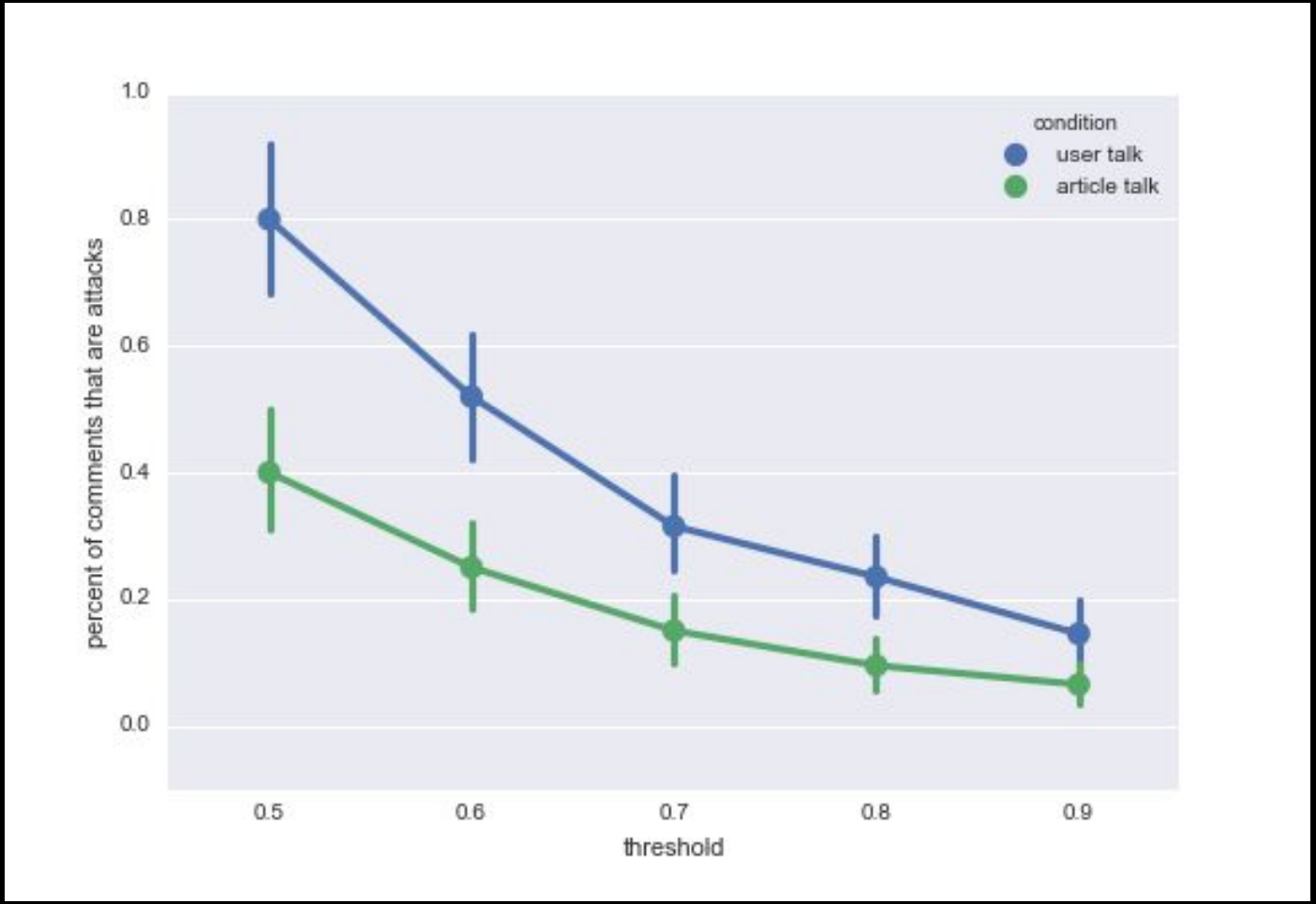
Goal:

Explore prevalence, dynamics and impact of personal attacks on English Wikipedia

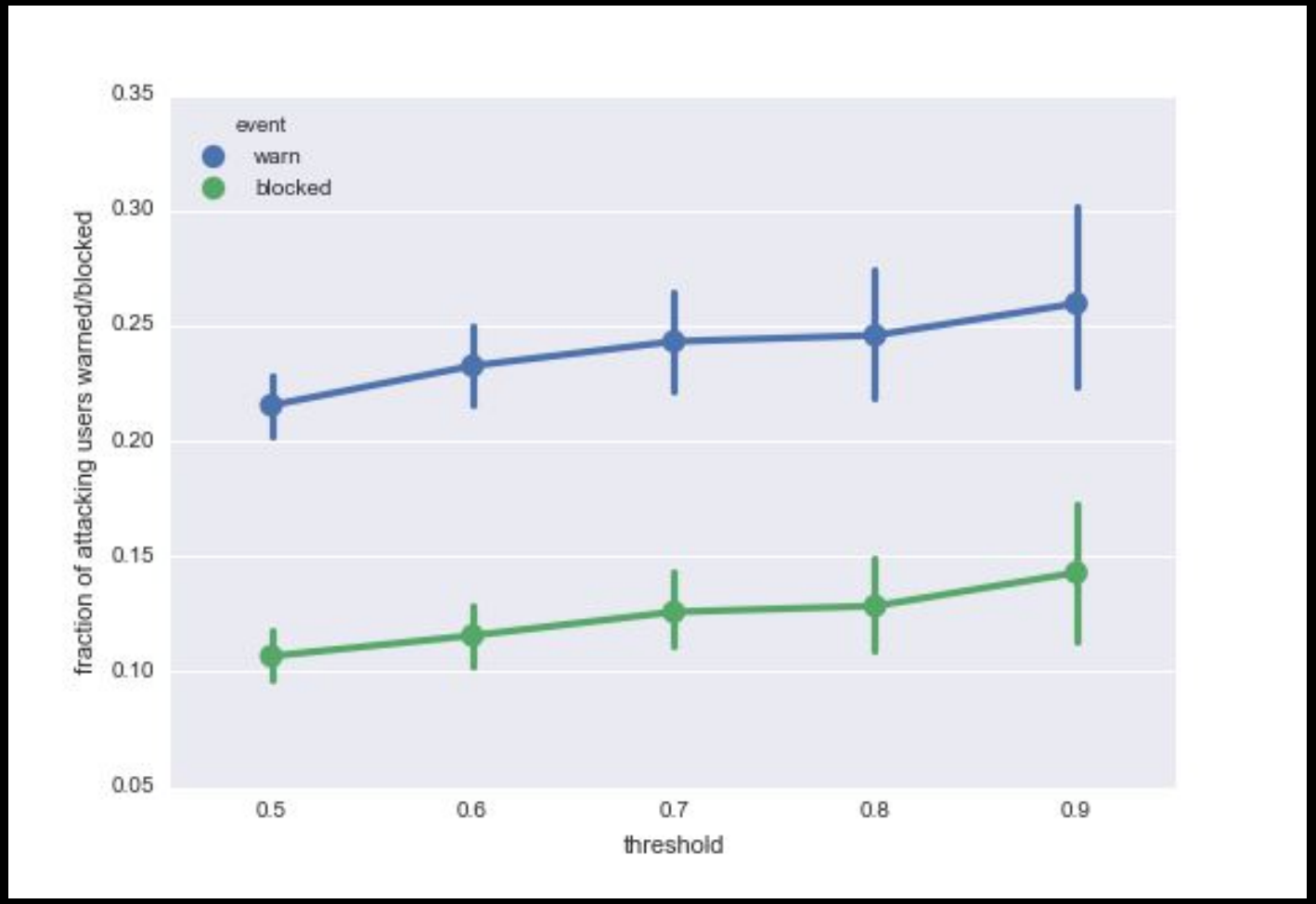
Input:

Complete historical data set of talk page comments + classifier scores

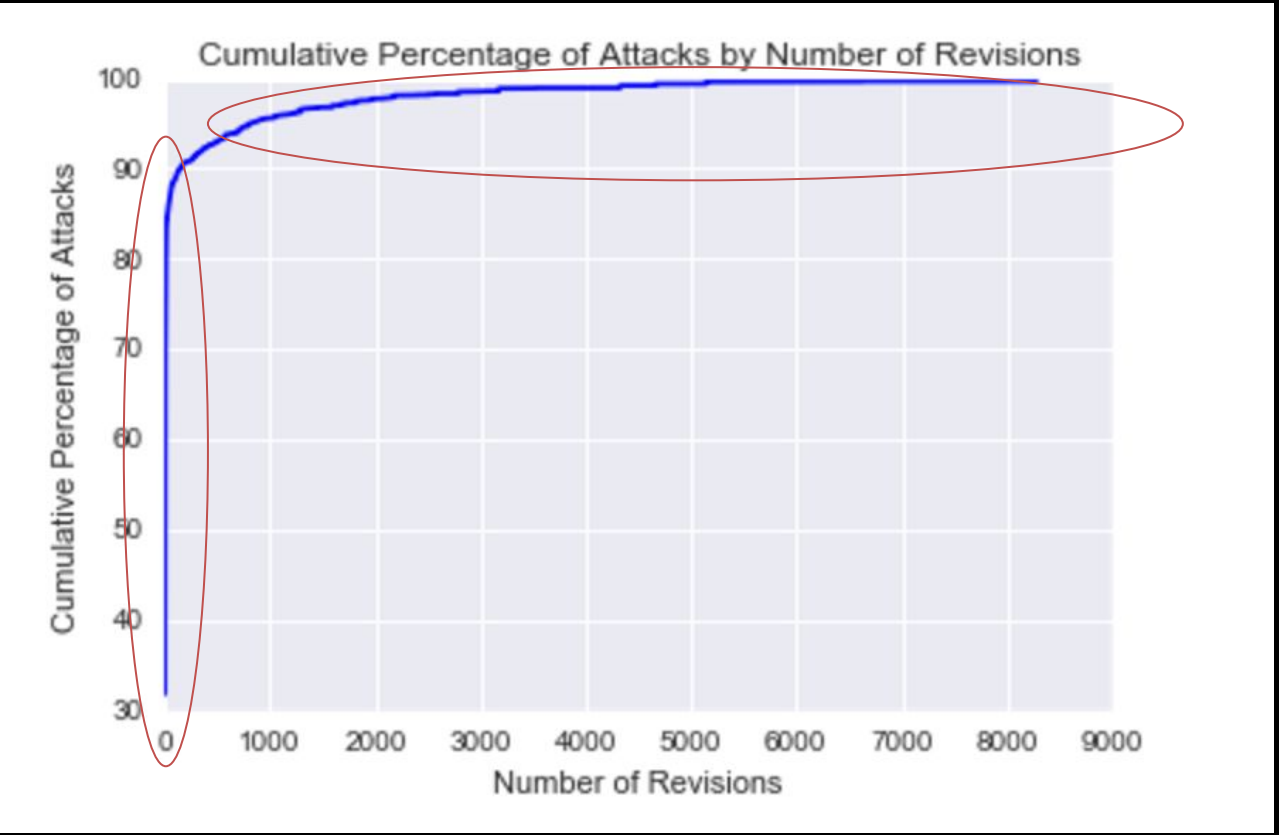
How many comments are personal attacks?



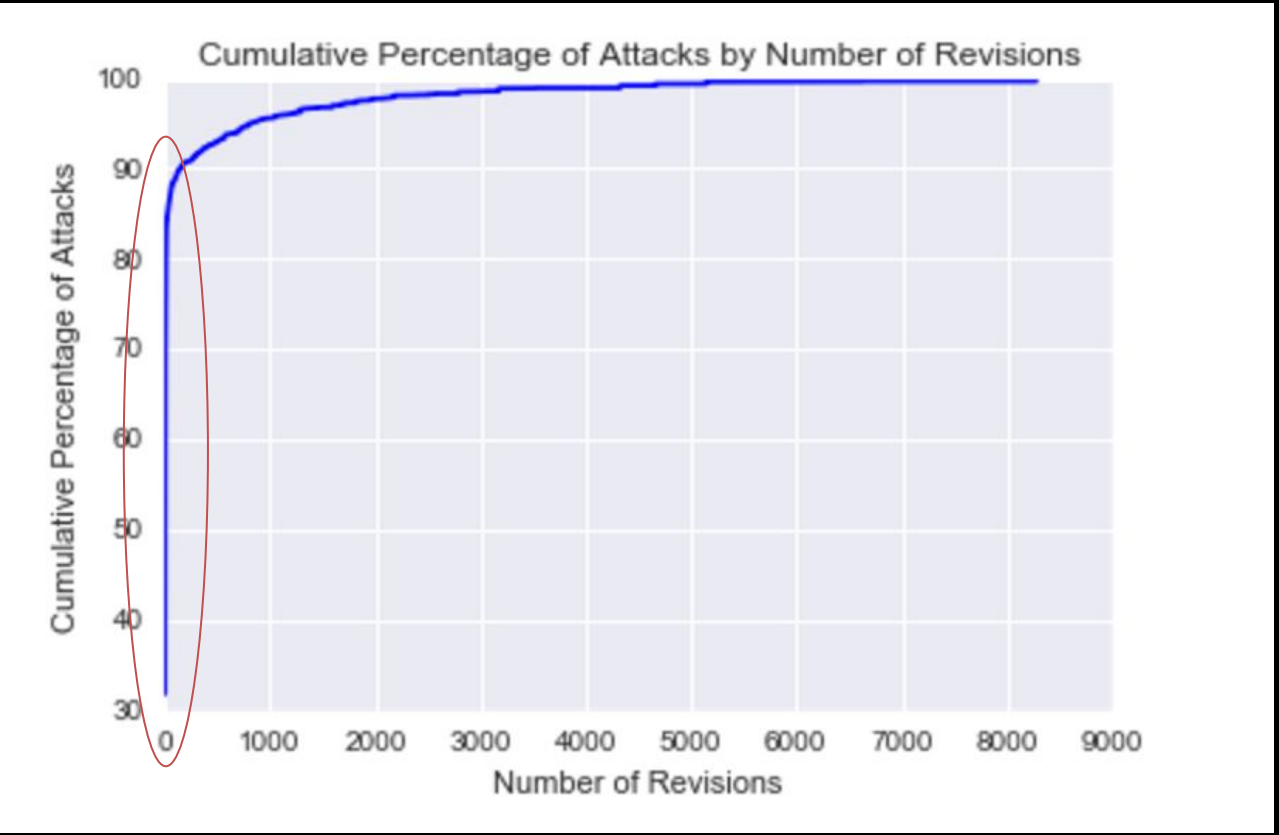
How many attackers have been warned/blocked?



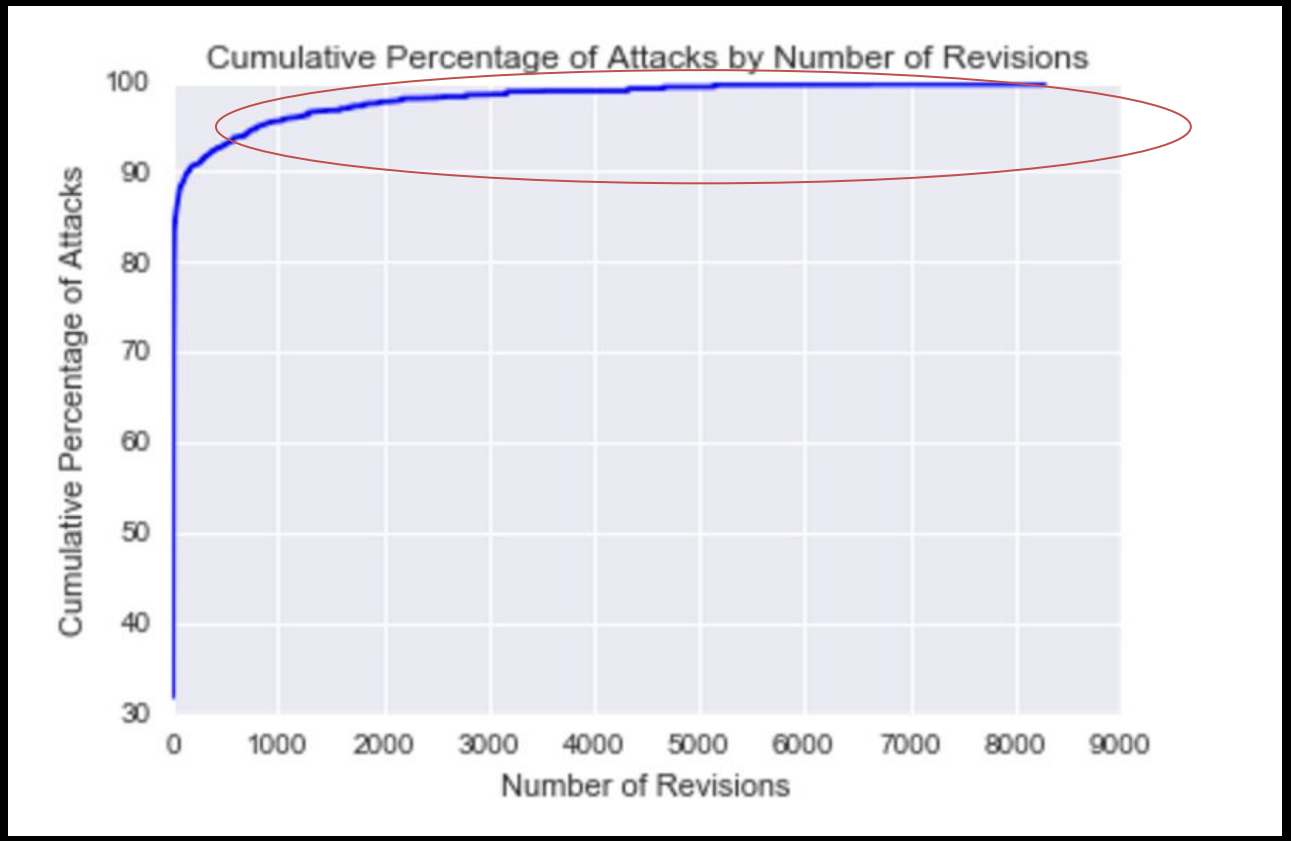
Two major types of attackers



75.7% of attacks come from users that have made fewer than 10 total revisions



9.3% of attacks come from users with over 200 total revisions



Next Steps

- Improve Modeling
- Extend Analysis
- Release of Annotated Datasets
- Integration with ORES

Questions?