UVA DSI Capstone Project
Wikimedia Foundation, Trust & Safety

Automatic Detection of Online Abuse and Prediction of Problematic Users in Wikipedia

Students - Charu Rawat , Arnab Sarkar, Sameer Singh
Faculty Advisor – Rafael Alvarado

Clients – Patrick Earley , Sydney Poore
Advisor – Lane Rasberry
Trust and Safety Team, Wikimedia Foundation

@DFRLab in DFRLab
Dec 21, 2017 · 9 min re

October 4, 2015
How can we stem the tid
Anastasia Powell, RMIT University anc
Women and men are just as likel

PC Magazine
Feb 14, 2017 · 4 min read

Intel Levels Up AI to Battle Toxicity in Online Games

essing text-based chat, but Intel has been working
peech, too, so it can to flag toxic behavior on

Russian Trolls T
Online abuse from

DECEMBER 18, 2018 | TOM SIMONITE

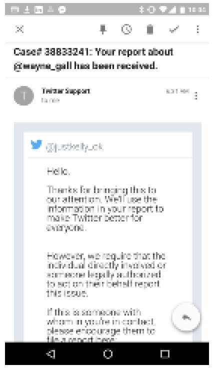
AI Has Started Cleaning Up Facebook, but Can It Finish?

Artificial intelligence has prov
nudity and pornography off of
hate speech and bullying is a r

What a brief history of web harassment can tell us ab
the contrarian crisis in op-ed land.

Kelly Ellis
Sep 2, 2016 · 2 min read

Twitter: We only review abuse when the target reports it



Read more...

19

1 response

n

Raising the bar on content moderation for social and gaming platforms (VB Live)

VB STAFF MARCH 27, 2019 12:46 PM

is week again tweaked the visibility of abuse

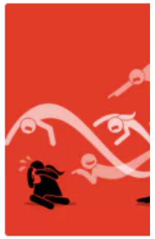
CEOs Can
CEO naiveté
and harassm

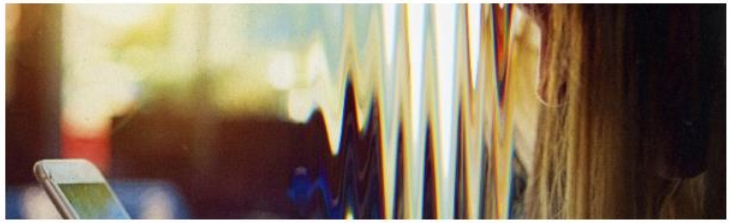
Sep 4, 2017 · 18 min read

Amnesty Global Insights in Am
Nov 20, 2017 · 15 min read

Unsocial Media: The against Women

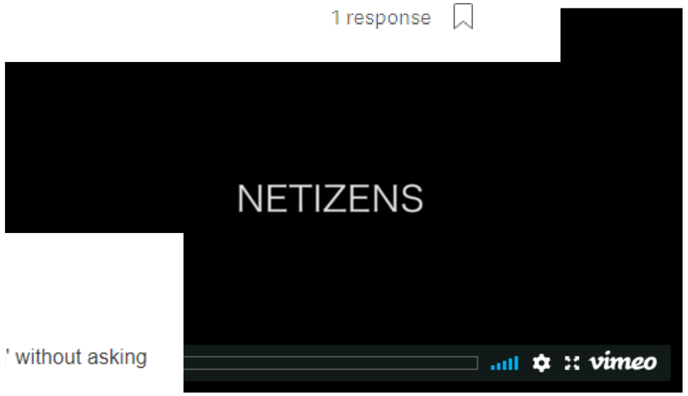
By Azmina Dhrodia, Researcher, Technology and Human Rights





Read more...

87



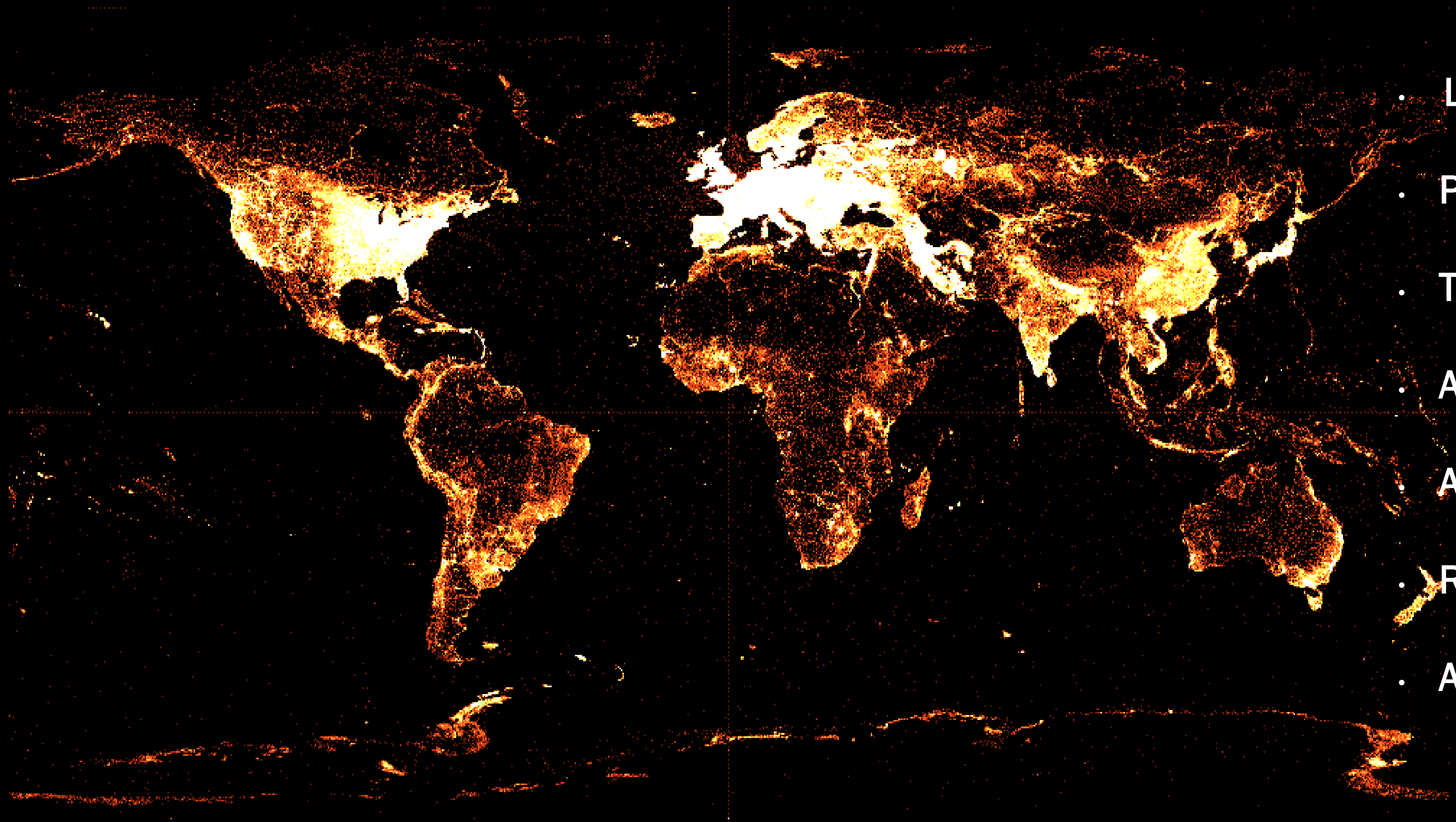
' without asking

Unsocial Media: Tracking Twitter Abuse agai Women MPs

Read more...

How can we as Data Scientists do what we do best, and make these communities safer from onlir abuse

The English Wikipedia over the last few years…



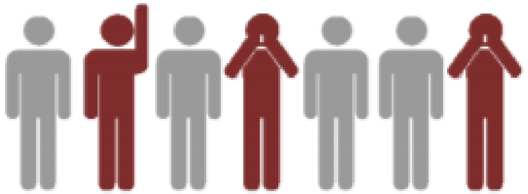
- Languages: 250+
- Popularity: Alexa Rank #5
- Total Edits: 868k
- All Pages: 46 mill
- Articles 5 mill
- Registered Users: 35 mill
- Administrators: ~1,193

https://commons.wikimedia.org/wiki/File:Wikidata_Items_Map_2014_%E2%80%93_2017_gif.gif

Online Harassment at Wikipedia

Pew Research Centre Survey (America centric)

41% have experienced personal attacks



66% observed attacks directed towards others

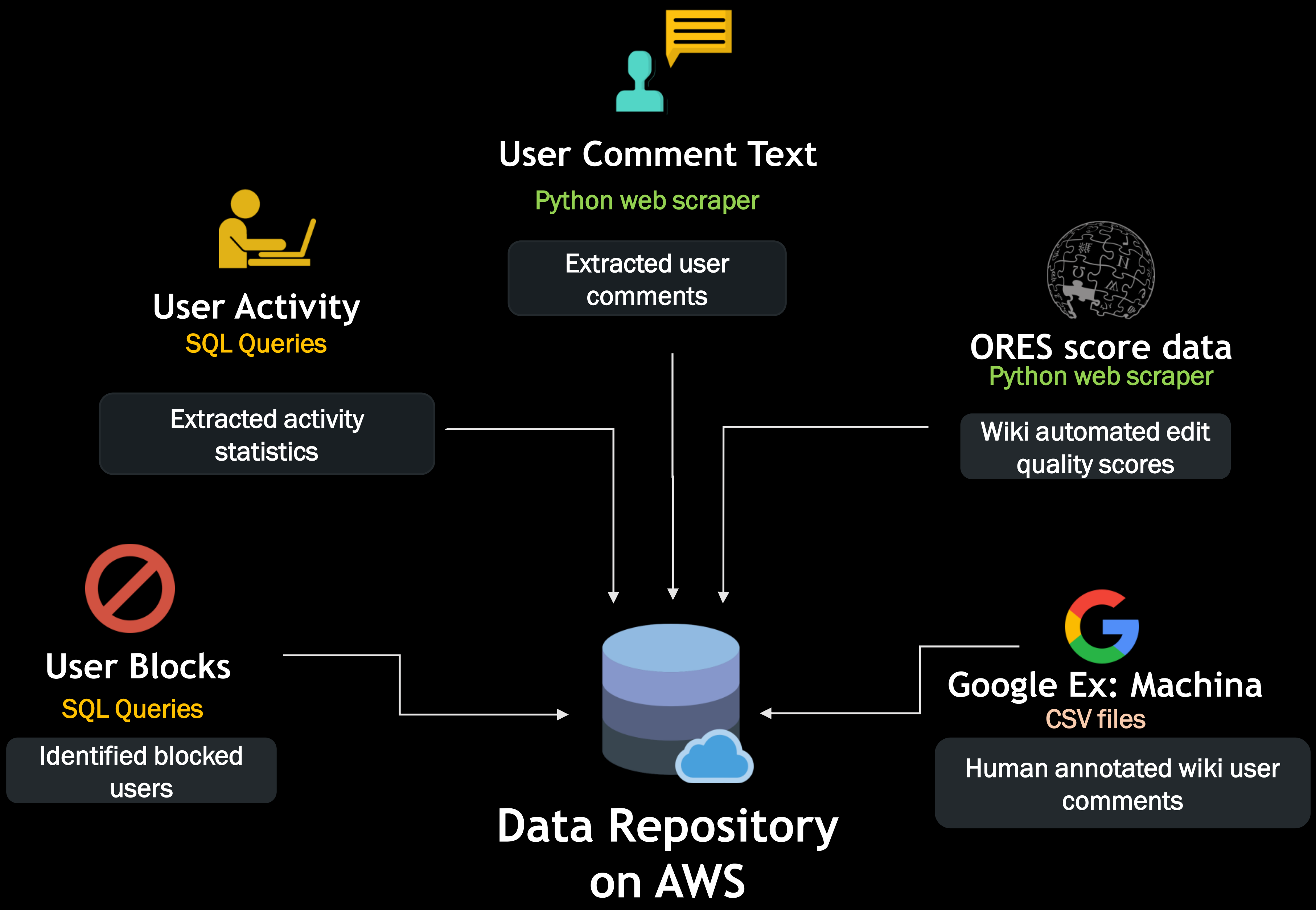
47% reported a decrease in contribution and engagement levels

Problem

There is a need for a more robust process to combat harassment, one that can scale well as the Wikipedia community continues to grow in its size and diversity.

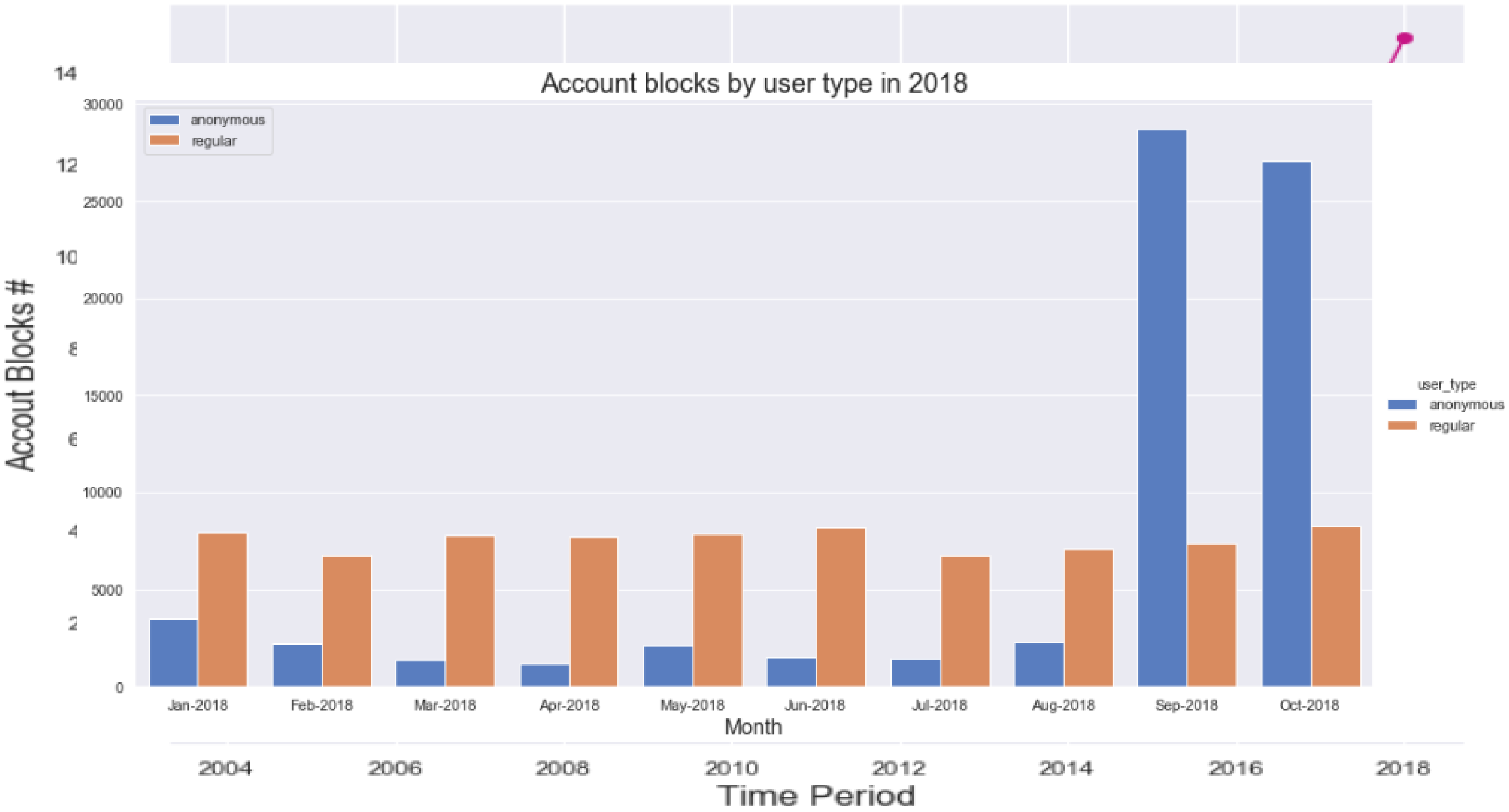
Our goal was to use machine learning to develop a model that can detect abusive content as well as predict problematic users in the community.
To that end, we leveraged a variety of data to analyse the prevalence and nature of online harassment at scale.

Data Pipeline



UNIVERSITY OF VIRGINIA
DATA SCIENCE
INSTITUTE

Block User Trends in Wikipedia

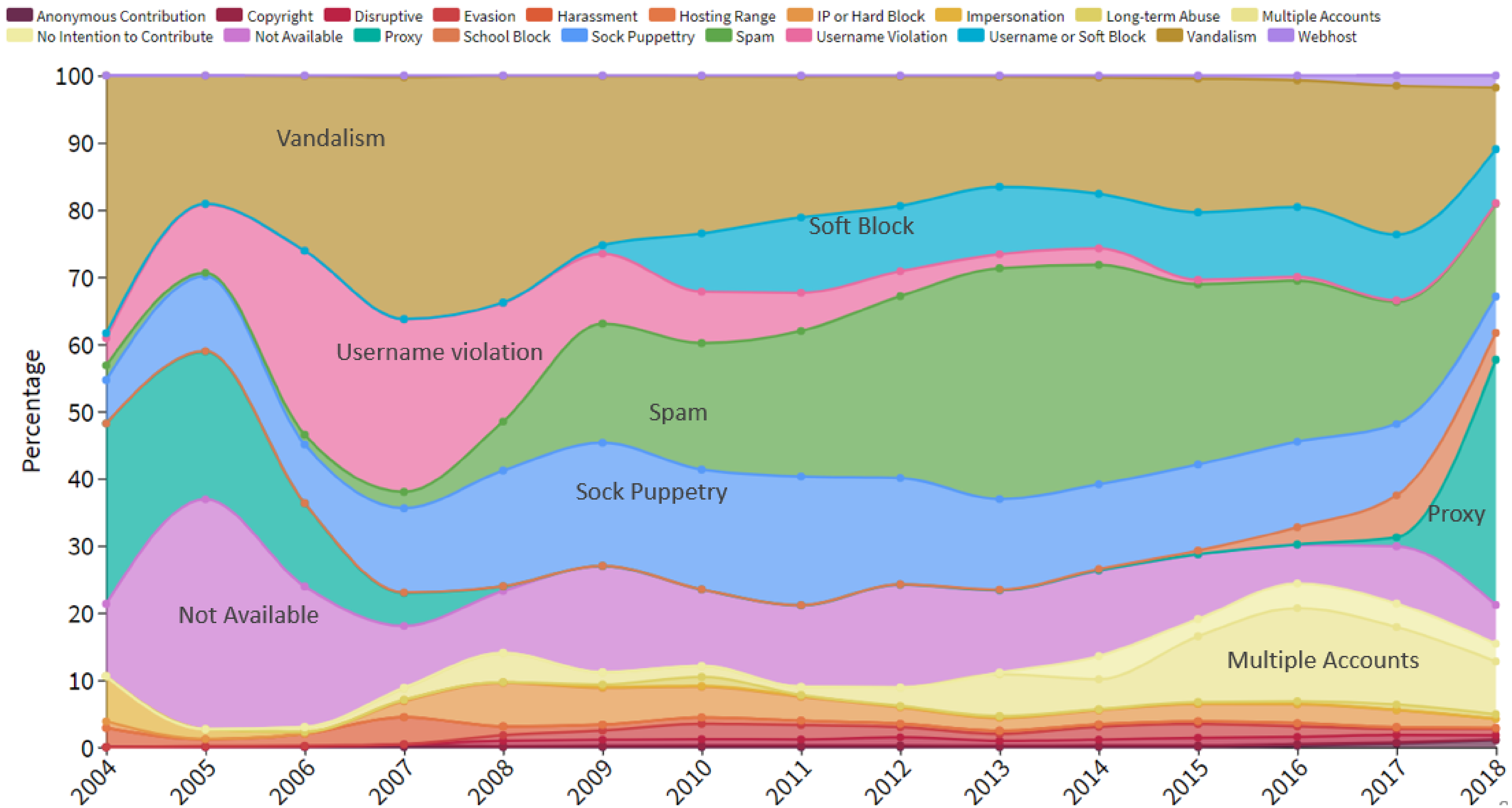


- **~1.02** million unique users have been blocked in Wikipedia till Oct 2018

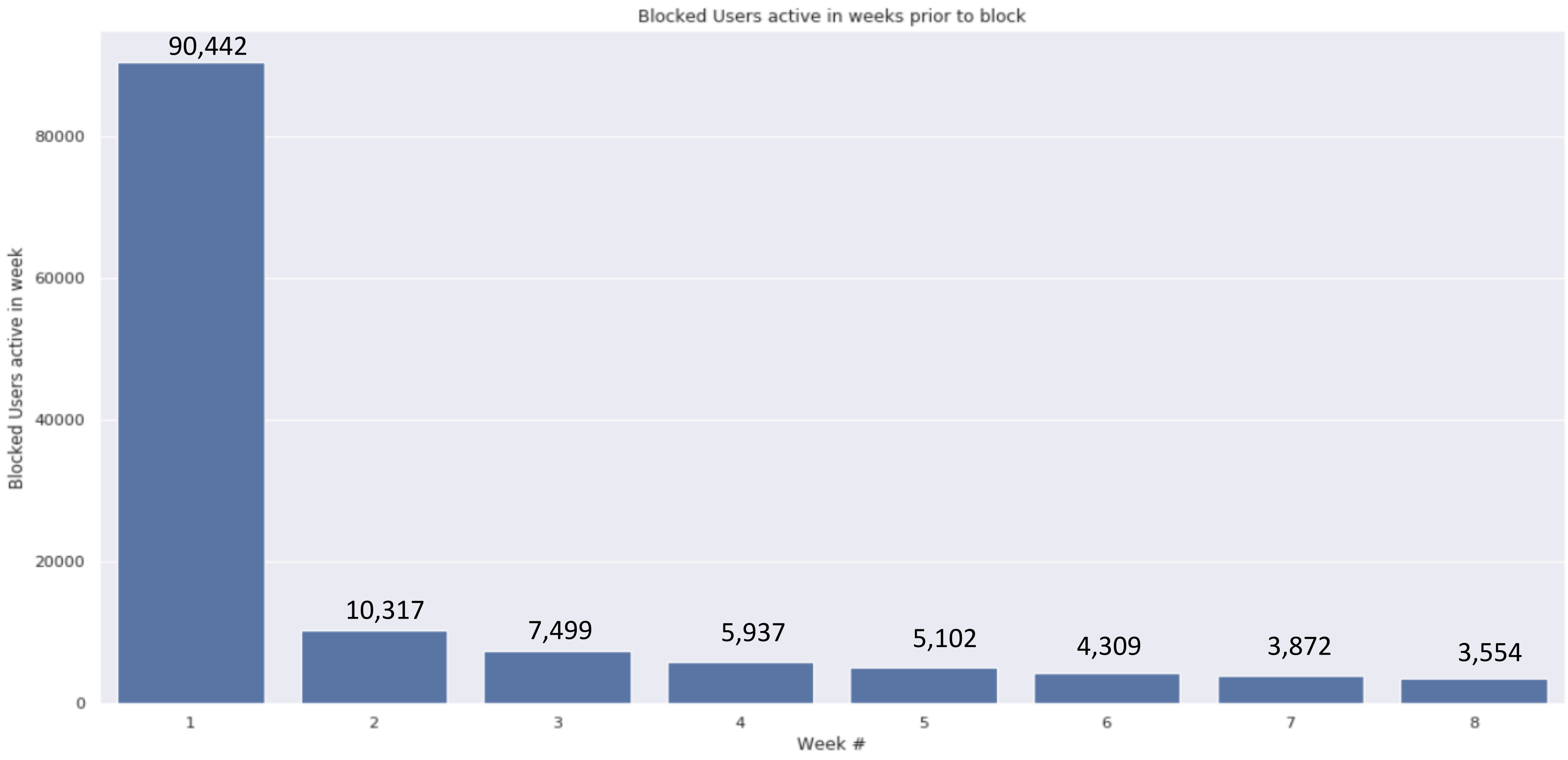
- **91.5%** registered users
- **8.5%** are anonymous users

- Increase in user bocks in 2017, 2018

For what reasons are users getting blocked over the years?



Active status of blocked users



90% of all blocked users are most active in the week that they get blocked within their 8-week rolling window

Abuse Detection - Problem at hand

- Goal
  - Prediction of a toxicity score for each user comment

- Inputs
  - User Blocks, User Comment Text, User Activity, and Google Ex: Machina Corpus

- Challenges
  - Data Cleaning and Pre-processing
  - Ground truth labeling
  - Class imbalance

Model Building Methodology

- Corpus Aggregation + Annotation

- Feature Extraction
  - Natural Language Processing
  - Orthographical Features

- Implementing Machine Learning Algorithms
  - Classification Techniques
  - Train – Test split
  - K-fold cross validation

- Model comparison
  - Using Google Ex Machina dataset to compare best performing model

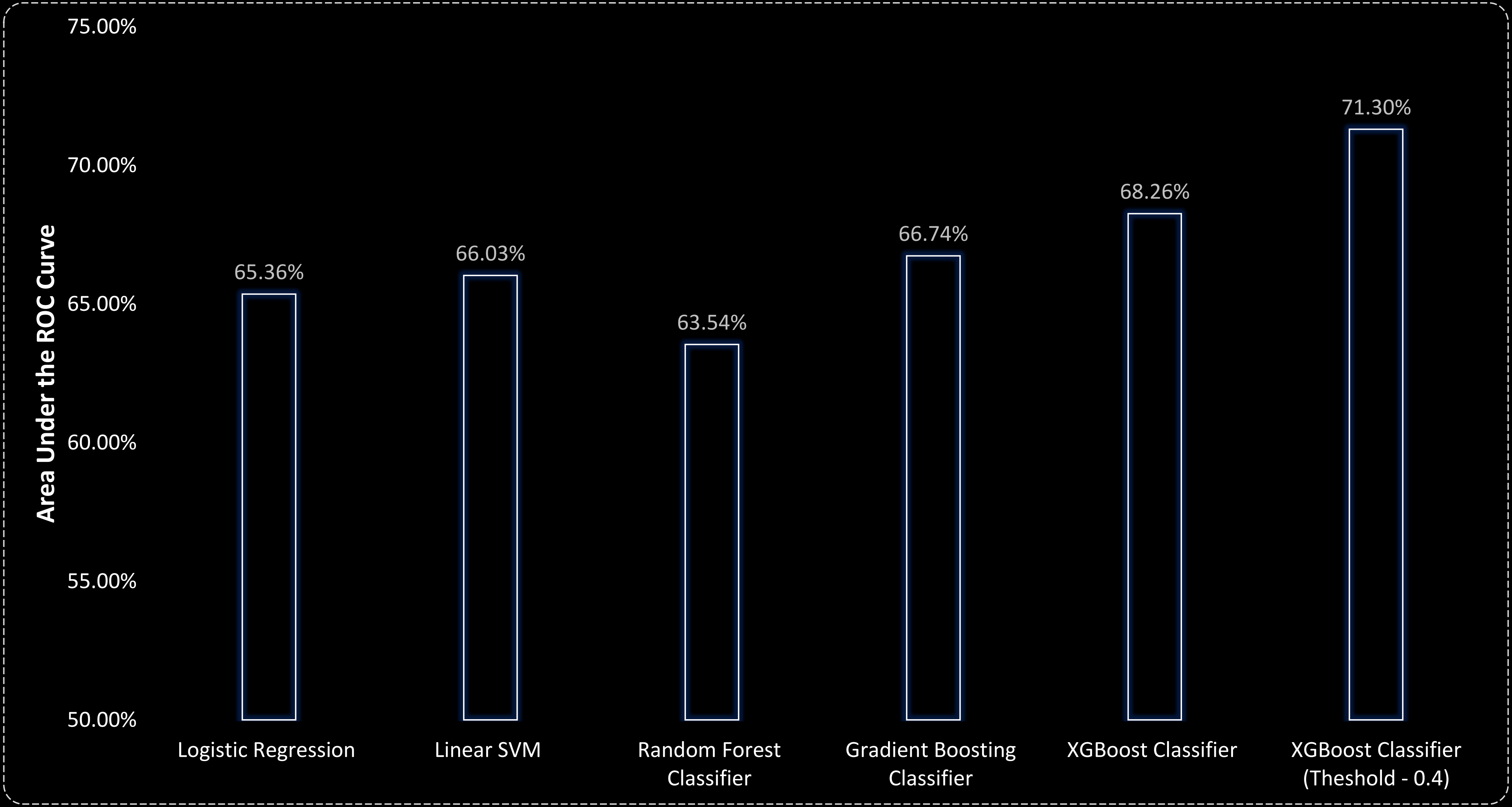
- Model Tweaking

Modeling Deep Dive - Feature Extraction

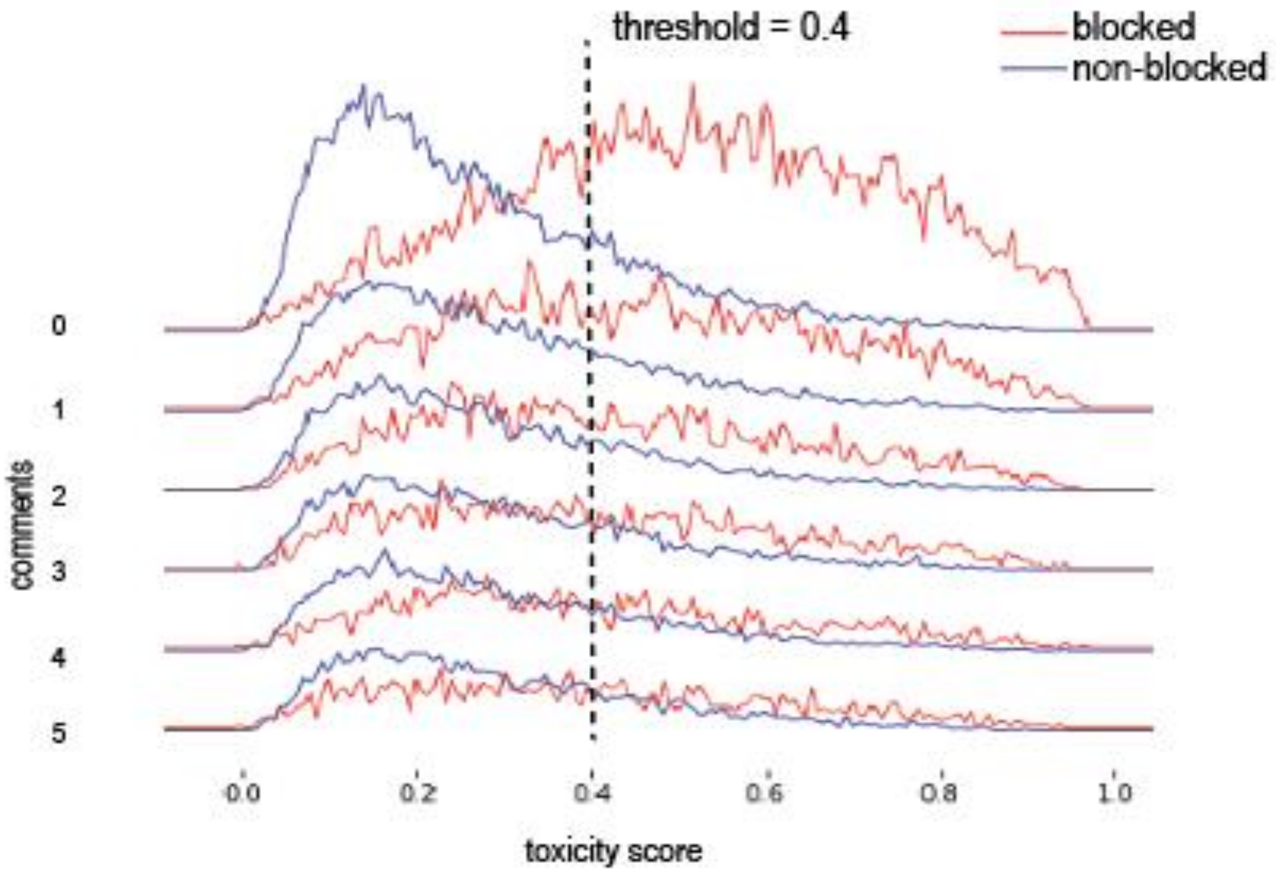
- **Natural Language Processing**
  - Char n-gram
  - Word n-gram
  - NLTK Sentiment Analyzer
  - Latent Dirichlet Allocation
  - Word Embedding - GloVe
  - Word Embedding - fastText

- **Orthographical Features**
  - Numeric Digits
  - Capital Characters
  - Special Characters
  - Average length of characters in word

Modeling Deep Dive - ML Algorithms



Toxicity Score Evolution



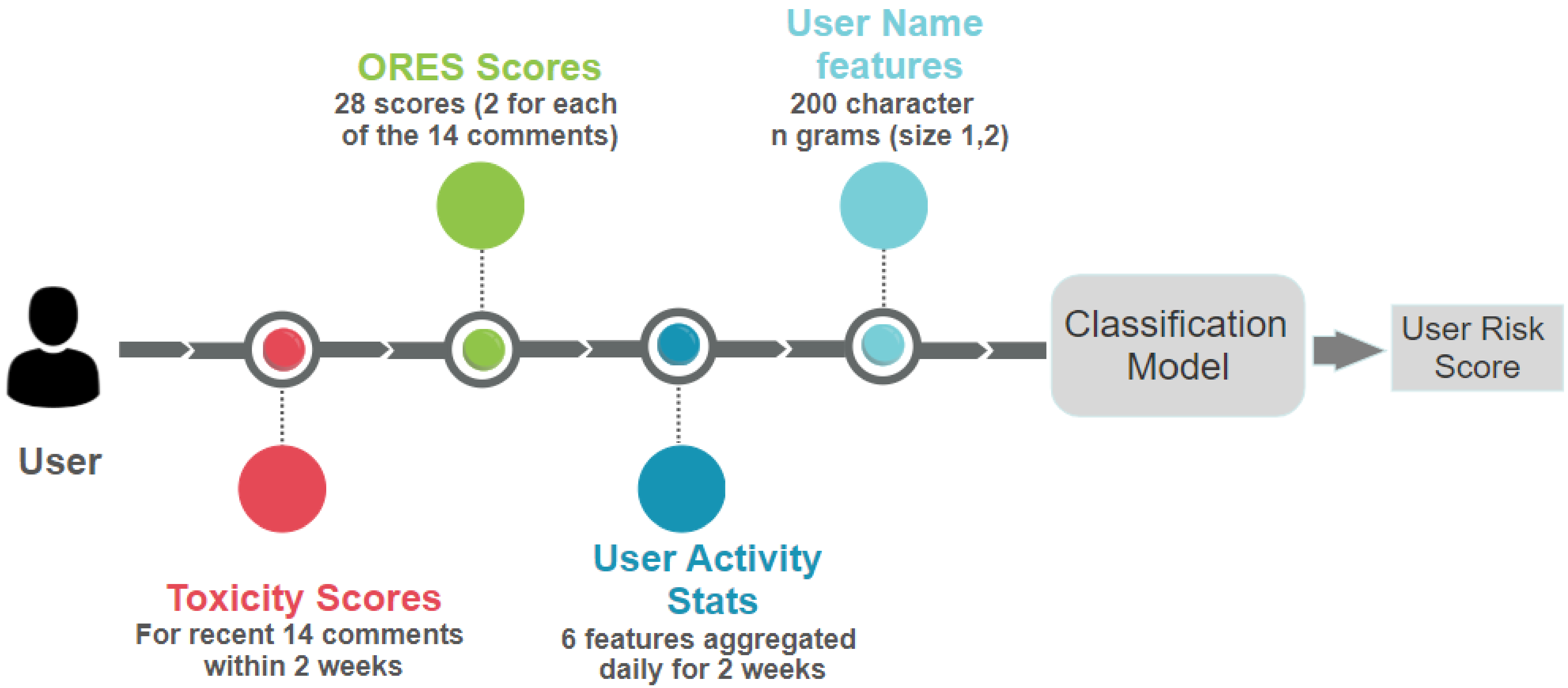
User Risk Model – Early Detection of Problematic Users

- **Goal**
  - ➢ Prediction of propensity of user to be blocked in the future

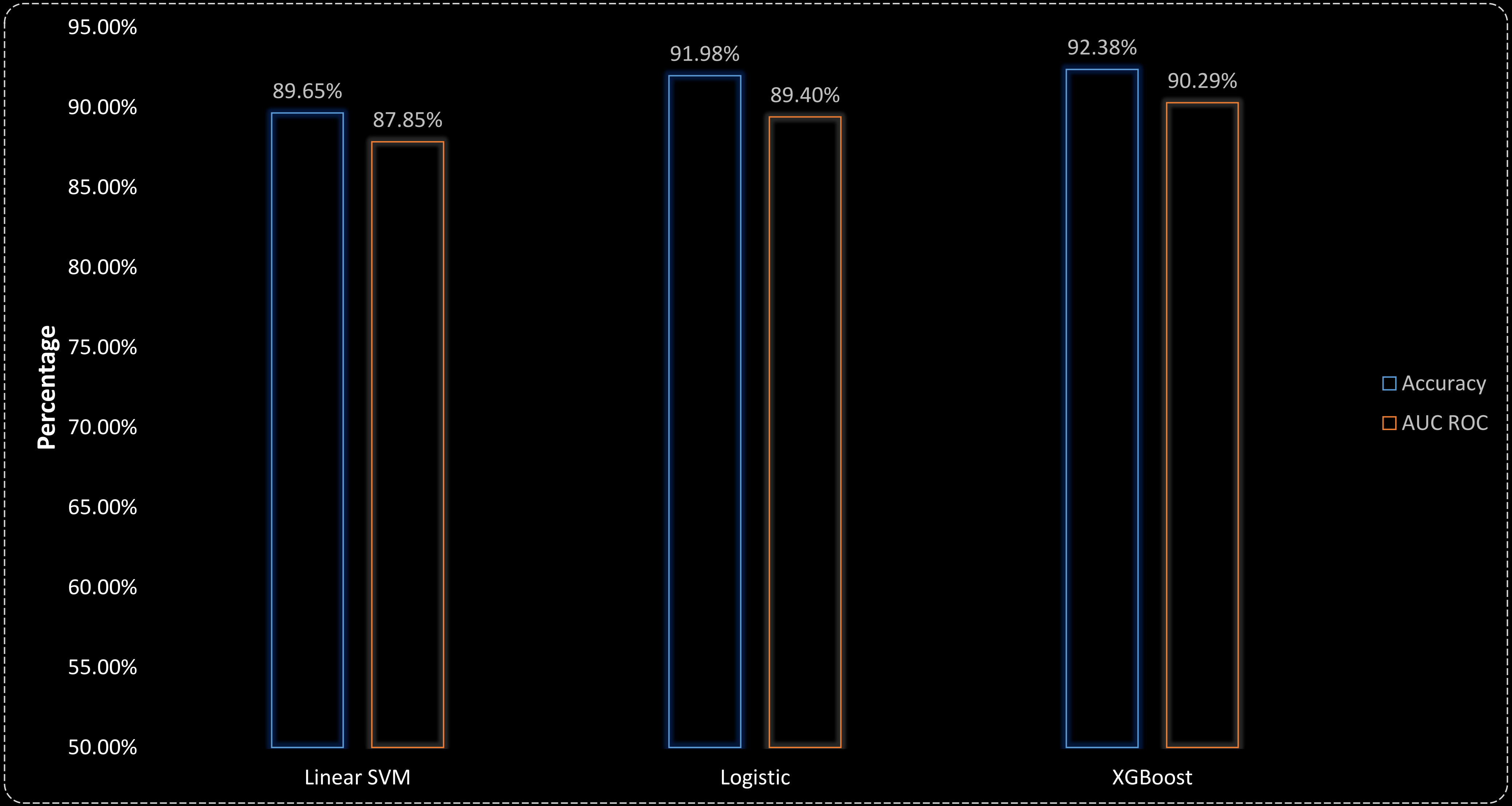
- **Inputs**
  - ➢ User Blocks, User Comment Text, User Activity, Toxicity Scores, and ORES Scores

- **Methodology**
  - ➢ Take into account n-1 recent scores and activity features of each user to predict the propensity for their n^{th} comment to be abusive
  - ➢ Leveraging Naïve Bayes approach with different Machine Learning Algorithms

User Risk Model - Early Detection of Problematic Users



User Risk Model - ML Algorithms

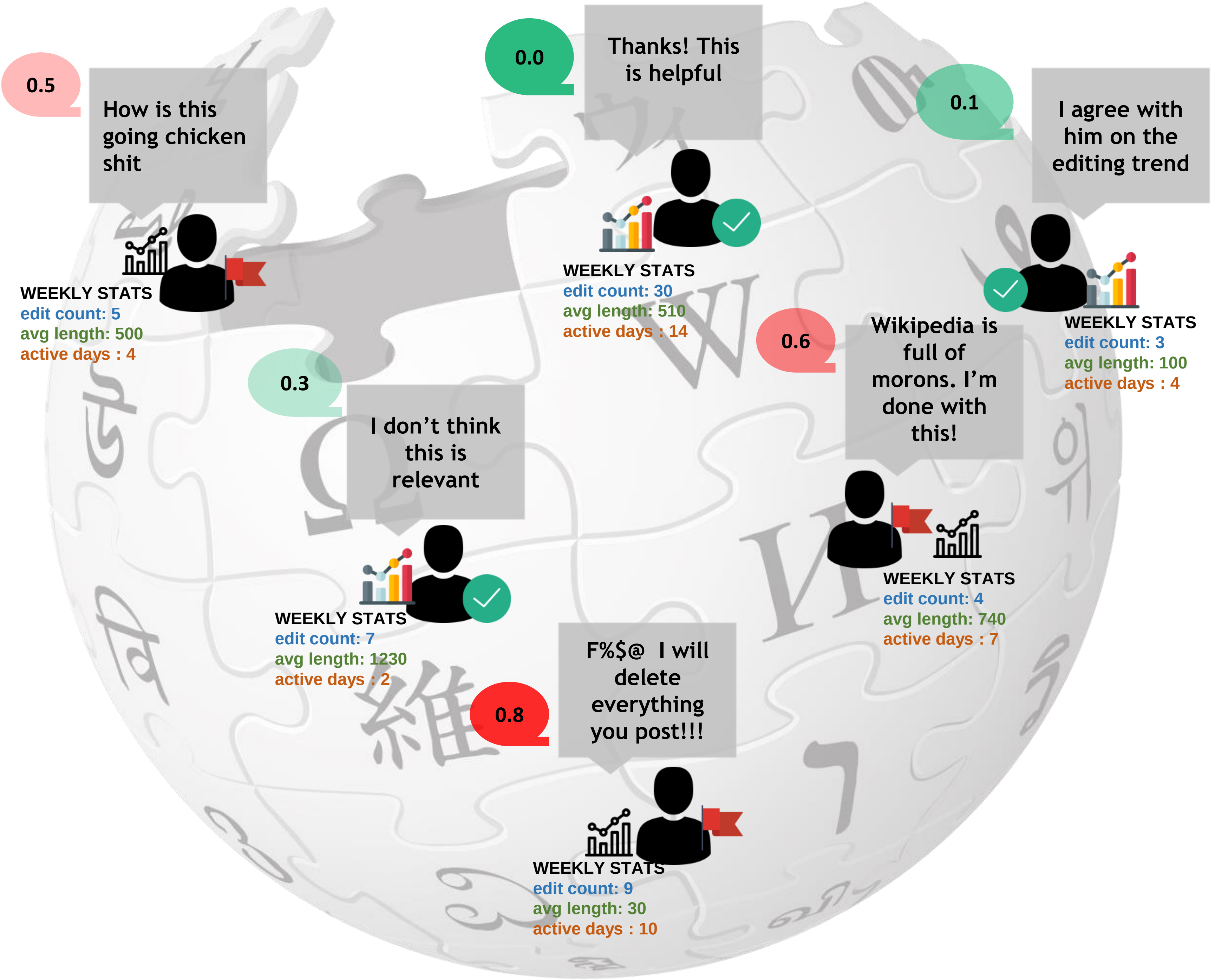


UNIVERSITY OF VIRGINIA
DATA SCIENCE
INSTITUTE

Performance of XGBoost model at different thresholds

- Optimizing for higher Recall values instead of Precision

Threshold	Accuracy	Precision	Recall	F1	AUC
0.1	91.17%	0.81	0.94	0.87	91.97%
0.2	92.29%	0.86	0.91	0.88	91.97%
0.3	92.60%	0.88	0.89	0.89	91.60%
0.4	92.71%	0.90	0.87	0.89	91.17%
0.5 *(default)*	92.38%	0.91	0.84	0.88	90.29%
0.6	91.94%	0.93	0.82	0.87	89.26%
0.7	91.59%	0.94	0.79	0.86	88.44%
0.8	91.10%	0.95	0.77	0.85	87.43%
0.9	89.94%	0.96	0.72	0.82	85.30%



UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE

Thank you!

UNIVERSITY OF VIRGINIA
DATA SCIENCE
INSTITUTE