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

Charu Rawat, Arnab Sarkar, Sameer Singh, Rafael Alvarado, and Lane Rasberry

UNIVERSITY OF VIRGINIA
DATA SCIENCE INSTITUTE

Introduction

In this paper, we propose a framework to understand and detect abuse in the English Wikipedia community. We analyze multiple publicly available data sources provided by Wikipedia. We propose a web scraping methodology to extract user-level data and perform extensive exploratory data analysis to understand the characteristics of users who have been blocked for abusive behavior in the past.

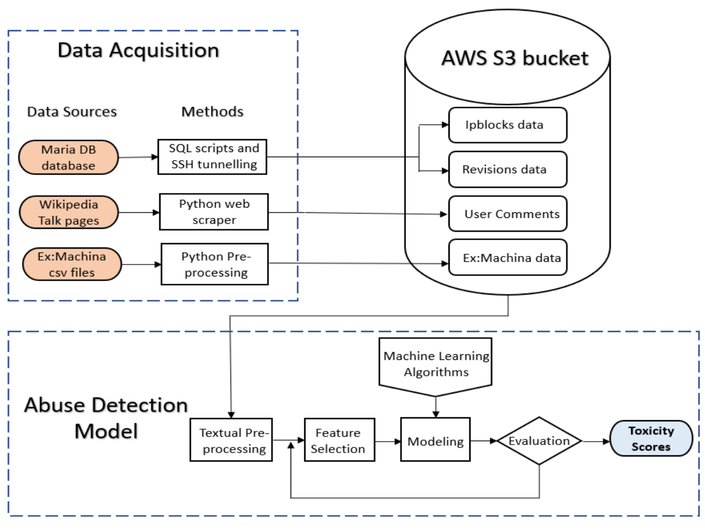
We further build upon these insights to develop an abuse detection model that leverages Natural Language Processing techniques, such as character and word n-grams, sentiment analysis, and topic modeling, to generate features that are used as inputs in a model based on machine learning algorithms to predict abusive behavior.

Data

A major challenge in our research was the lack of a unified dataset for all our analyses. We identified four sources of data, both structured and unstructured, that could be leveraged for our research. We then built a pipeline to extract, process and store all the data. Below is a summary of the data we used:

- Block Data – Record of users blocked between Feb 2004 - Nov 2018 in the English Wikipedia, ~1M unique records.
- Revision Data - Metadata for user revision edits for the period Jan 2018 – Aug 2018, 95.7M rows for 6.1M users.
- User Comments – This textual data can be accessed through the public XML dumps or by scraping the website. We developed approaches to acquire data through both. We discovered that the XML approach was computationally heavier, hence we developed a Web Scraper to scrape the data. The scrapped corpus consists of 503.4K comments for 50.6K users.
- Google Ex:Machina annotated comments – Annotated dataset of 197.6K user comments.

Fig I - Data Pipeline



Data analysis

Block Data

We analysed block data to study block trends over the years. Mentioned below are the key findings:

- We observed that out of the total blocked users 91.5% are registered users whereas 8.5% are anonymous users.
- We discovered a previously unknown sharp increase in the number of users getting blocked in 2018. Our analysis suggested that this spike was due to “proxy” blocks, and majority of these blocks were enacted by one admin.
- We found that vandalism, spam, sock puppetry are dominant reason of blocks for registered users whereas for anonymous users it is proxy, webhost, and school blocks.

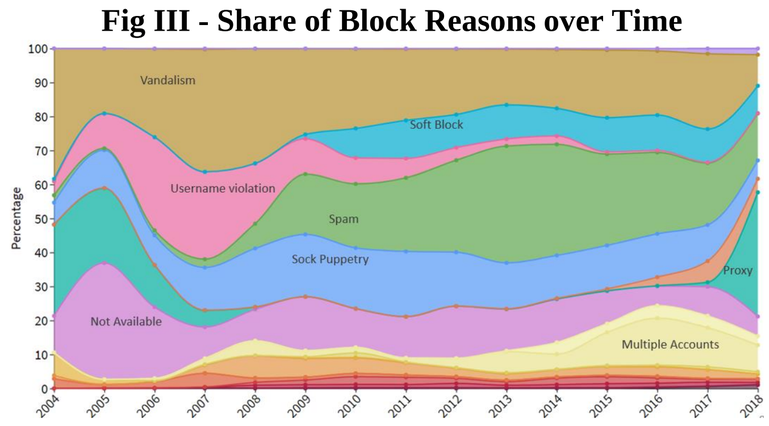
Fig II - User Block Trends

Revision Activity Data

We analysed the revision activity data to gain insights into how users behaved in the weeks leading up to them getting blocked. Mentioned below are the key findings:

- Users blocked for reasons such as vandalism, proxy, spam tend to make anywhere between 5-15 revisions on average every week within their recent 8-week window.
- Non-blocked users generally make a higher number of major and minor edits on a daily basis compared to blocked users.
- Blocked Users have a higher proportion of deleted edits.
- The exploratory data analysis on the revision activity data highlighted the difference in editing patterns of users who get blocked vs users who don’t.

Fig III - Share of Block Reasons over Time



Modeling

Methodology

- Corpus Aggregation
- Data Pre-processing
- Corpus Annotation
- Feature Extraction
  - ➢ Orthographical Features
  - ➢ Char/Word n-gram
  - ➢ NLTK Sentiment analyser
  - ➢ Topic Modelling
  - ➢ Username Based Features
- Implementing Machine Learning Algorithms
  - ➢ 75% - 25% split between train and test data
  - ➢ k-fold cross-validation
- Model Comparison with Google Ex: Machina data
- Model Threshold Selection
  - ➢ XGBoost – AUC of 84%

Table 1 – Unique User Counts

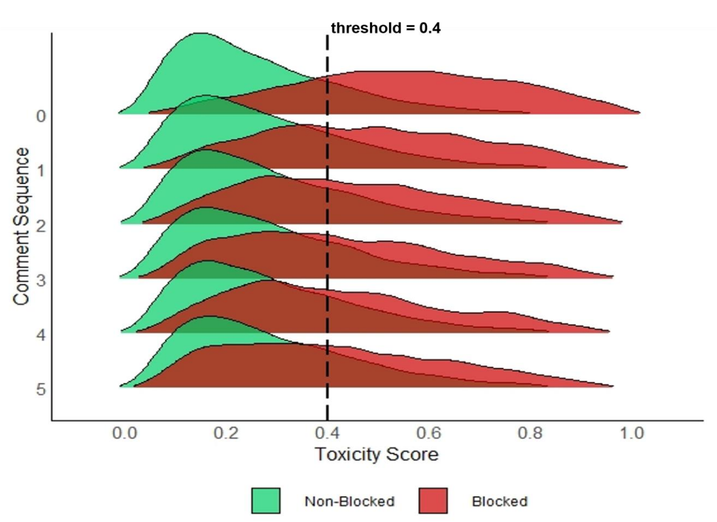
Data	Blocked Users	Non- Blocked Users
Initial Corpus	19,580	31,066
After pre-processing	12,186	25,748

Model Output

A toxicity score, ranging from 0.0 to 1.0 (least to most toxic), was generated for every user comment.

Figure IV below depicts the evolution of the toxicity of user comments with time. We can can observe that the toxicity score for blocked users skews towards 1.0 as the comments get more closer to them being blocked, whereas for non-blocked users, the toxicity score remains skewed towards left.

Fig IV - Toxicity score evolution



Conclusion

In this paper, we developed a framework for a data-driven approach to detect abuse in the English Wikipedia community.

- We established two methods to acquire the user comment data through different sources.
- We uncovered previously unknown interesting dynamics within the blocking ecosystem of the English Wikipedia.
- We drew insights that highlighted the difference in the editing behavior of blocked versus non-blocked users
- We built an abuse detection model trained on the text corpus generated from the user comments. And we found that XGBoost Classifier gave the best performance with an AUC of 84%.

Table 2 - AUC Scores for different algorithms

Model	AUC
Logistic Regression	65.36 %
Linear SVM	66.03 %
Random Forest Classifier	63.54 %
Gradient Boosting Classifier	66.74 %
XGBoost Classifier	68.26 %
XGBoost Classifier (with username features)	83.10 %
XGBoost Classifier (with threshold of 0.4)	84.00 %

Acknowledgment

The authors would like to thank Patrick Earley and Claudia Lo from the Trust & Safety team at the Wikimedia Foundation for their encouragement and helpful insights during the course of this project.