Capstone Proposal
Wikimedia Foundation, Trust & Safety
Cyber Harassment Classification & Prediction of User Account Blocks

Students - Charu Rawat (cr4zy@virginia.edu)
Arnab Sarkar (as3uj@virginia.edu)
Sameer Singh (ss8gc@virginia.edu)

Adviser - Rafael Alvarado (rca2t@virginia.edu)

University of Virginia Master's in Data Science Program

Client and Sponsor - Wikimedia Foundation

Wednesday, September 26th, 2018

Summary

In today's day and age, online discussion has increasingly become an avenue for various kinds of abuse. This abuse inhibits the open exchange of ideas and resources amongst community members and impedes the growth of online platforms. As per the Wikimedia Community engagement insights it was observed that ~47% of the respondents who experienced harassment on their platforms reported a decrease in their contribution and engagement levels [1]. The Wikimedia foundation has no automated system of detecting this abuse online and identifying users who should be blocked from communities for their problematic behavior.

To this end, we aim to empower the Wikimedia foundation by enhancing their Trust and Safety processes using a data-driven approach. This will enable a safe and secure environment for their community users to engage in open exchange of ideas and resources. Our goal is to leverage the text data sourced from the Wikimedia platforms and analyze it to understand user behavior by leveraging machine learning algorithms. To that extent we will provide the Wikimedia foundation with a tool in the form of a model that:

- performs classification of the historical user account blocks on the basis of different types of user behavior and activity,
- predicts and flags users who are at risk of getting blocked in the future.

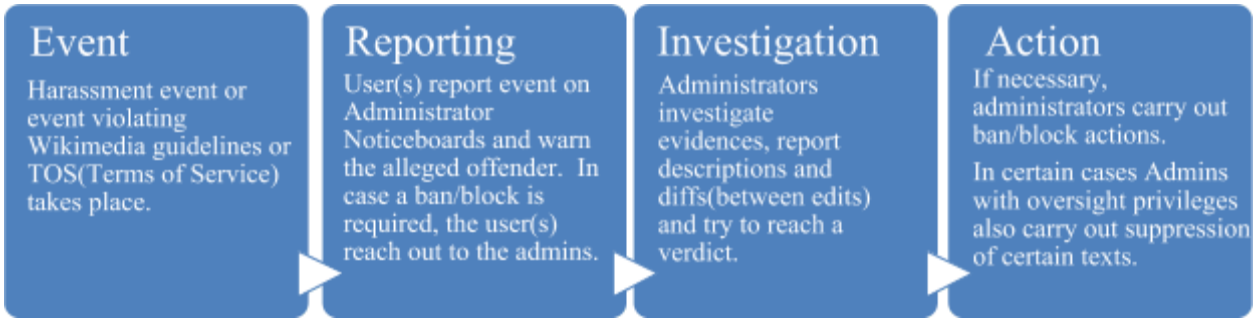
We expect that our tool will enhance Wikimedia Foundation's existing Trust & Safety processes and serve as a preemptive measure in combating harassment online. By providing a tool with increased accuracy in automatic harassment detection and the ability to flag problematic users in advance, the tool will help resolve conflicts in a shorter time to effectively protect the community users.

Background

Wikimedia strives to provide free education content to the world through its project and host websites such as Wikipedia, Wiktionary, Wikidata, and others [2]. Currently, Wikipedia is one of the most searched websites (Alexa rank-5), and is available in over 250 languages [3]. As of September 2018, the English Wikipedia, the largest language edition, has ~406 thousand editors on average per month and has had ~856 million edits to date. On a monthly basis, the community sees about ~150 thousand new users being registered every day [4].

In today's digital world, where information is readily available and everyone has the ability to easily connect with each other, there are immense possibilities for knowledge sharing, but at the same time there exists a dark side to it. With time there has been steady increase in online attacks to the extent that it can now be classified as a "feature" of online communities. According to the Pew research survey, ~41% of Americans have experienced personal attacks online and ~66% have observed attacks directed towards others [5]. Wikimedia has not remained untouched by that phenomenon. In the last few years, the community has seen a steady increase in both the number of incidents and variety of attacks

(Wikihounding, sock puppetry, user space harassment, posting personal information etc.) [6]. But, Wikimedia has realized that online harassment of its editors can have a detrimental effect on the growth of its platform, and so the community has taken proactive steps to create awareness around such issues and has put into place, a well-defined and structured "no personal attack" policy for all of its members [7]. When users misbehave there is a Wikimedia process by means of which they might be blocked, which means that their account loses the ability to post to Wikimedia projects. Prior to the block process they engage in some automatically recorded behavior in Wikimedia projects, such as a hostile interaction, spamming, or history of conflict. Human evaluation of that behavior confirms a blockable offense and enacts a block. Mentioned below, are the steps involved in detection and subsequent blocking of a user account within its current framework of combating harassment.



Wikipedians (or editors - the volunteers who write and edit Wikipedia articles) who experience harassment firsthand or witness harassment or violation of Wikipedia Terms of Service report such an event on the various Wikipedia Administrator Noticeboards [6]. They may discuss the event there with others and warn the offender in question. If they feel that a user needs to be banned or blocked, they reach out to the admins on the noticeboard. In certain cases, Wikipedians may also contact Admins, Arbitrators and Stewards directly through email if they don't want to report an event publicly (for example, in cases of "Doxxing"). Administrators then look through all the evidence posted by the reporter and try to arrive on a verdict. In certain scenarios they may need to escalate the case to users with higher permission privileges like Arbitrators and Wikipedia Stewards.

If admins find evidence of wrongdoing, depending on the severity of harassment action, they enact actions like blocks, range blocks ("blocks across multiple IPs" to prevent sock puppetry) [9]. Administrators may also carry out revision deletions which involve hiding content from all Wikimedia users without admin privileges [10]. Arbitrators and Stewards may carry out suppression of content (also called Oversight), which hides content even from Administrators, as well as suppression of blatant attack names classified as block-suppression (local level action) or lock-suppression (global level action) [10]. Stewards can also mitigate harassment across multiple Wikimedia platforms by using global lock on a registered account or global block of an IP [11]. In case of disputes, as a resolution, bans may also be awarded to certain users. This can be a site ban, page ban or an interaction ban. A ban is like a formal prohibition imposed by admins, an Arbitration committee or by community consensus and is usually enforced by issuing blocks. Users who have been site banned are no longer part of the Wikipedia Community and may not even access their Talk pages [12].

This process for reporting, investigation, and final action on events of harassment in Wikipedia is currently human driven and moves forward at every step through human deliberation. Relying on human evaluation works in some ways but is not a solution which scales with the growth of Wikimedia projects. There have been times when certain cases fall through the cracks of the process of human scrutiny due to lack of consistency in application of blocks, failure to identify bad behavior, bias to overlook such behavior and an inability to understand the context and scope of the problem pertaining to the blockable offense. The Google Ex Machina study in 2017 found that, on Wikipedia, blockable actions have been triggered only for one fifth of the personal attacks [13]. Overall, as a result of such behavior in the community, there has been a reported drop in the confidence and motivation levels amongst its members and in certain cases, this has even been the cause of Wikipedians leaving the platform completely. This is a major cause of concern for a growing global community like Wikipedia which sees an average of 4 million edits and 170 thousand new pages being created every month by its community members [4].

We believe that an automated harassment detection process that runs in parallel with the existing human driven approach will aid in combating harassment, aggressively and more effectively as the community continues to grow in its size and diversity.

Technical Approach

Dataset Description
Our primary dataset consists of English Wikipedia talk namespaces and is a 15GB+ XML compressed file, available as part of the Wiki Dumps. The data is textual in nature, and contains records of all edits, comments, and blocks with associated timestamp from 2004 till date. Each log has multiple attributes associated with it. Some of the important attributes are –
Username (registered community user) / IP address (anonymous user) of the user who requested a block
Username (registered community user) / IP address (anonymous user) of the user who the block is against
Edit/comment added by the user
Timestamp for when the block was requested
Reason for the block being requested
Duration of the block

Below is the snippet of a log capturing user account block –

- 15:05, 23 September 2018 Bbb23 (talk | contribs) blocked XXXTENTOTINOSCIONESE (talk | contribs) with an expiration time of indefinite (account creation blocked, email disabled, cannot edit own talk page) *({{checkuserblock-account}}: Abusing multiple accounts: Please see: Wikipedia:Sockpuppet investigations/Arthur Fonzarelli MDCCXXXVIII)*
- 10:12, 23 September 2018 Alexf (talk | contribs) blocked Harrison, Benjamin USD (talk | contribs) with an expiration time of indefinite (account creation blocked) *(Suspected sock puppet of Franklin, Benjamin USD)*
- 10:10, 23 September 2018 Favonian (talk | contribs) blocked 36.71.233.241 (talk) with an expiration time of 31 hours (anon. only, account creation blocked) *(Vandalism)*

We would use the entire corpus data to create a user block data subset which will enable us to determine a) what accounts should be classified as a block b) what activity did the user engage in prior to the block.

General Methodology and Evaluation

We will leverage Python as the primary language for our analyses and model building process. Since the raw data is in XML format, our first step would be to use a data parser to extract contents from the files and store them in an appropriate format. The data will then undergo extensive pre-processing so as to improve its quality. In the pre-processing module, we will leverage the WordNet corpus and spell-correction algorithms to correct grammatical mistakes in raw sentences such as deleting repeated letters in words, deleting meaningless symbols amongst other things. In order to build a model that classifies user accounts as block or not, we would consider circumstances before the enacted block, including text of messages posted, a user's Wikimedia content edits such as posting spam links, interaction with other community users etc. To understand these dynamics of what constitutes an account block we will employ techniques of feature extraction onto every post associated with each account to measure different aspects of the user activity. This would include local feature extraction using text mining techniques such as TF-IDF to filter out unimportant words and reduce dimensionality of feature space as well as sentiment analysis techniques such as N-grams, Bag-of-Words, etc. to detect use of offensive words. Prior work in the field of online harassment detection has included the use of models built on logistic regression [14], neural networks (CNN, RNN) [15], SVM [16] etc. methods which we intend to explore and build upon. Our methodology in building a model to classify an account as a block or not would be to first train our model using the data corresponding to labelled account blocks, talk page comments, and their corresponding features using various modeling methods. We will then choose a model after testing each of them by performing cross-validation and comparing model metrics such as R-square, ROC, AUC, etc. The finalized model will be leveraged onto the test dataset. As the dataset is time series, we will predict the probability "at-risk" score for each user using the n-nearest user comments/edits. We will then choose two reasonable thresholds over this score which will determine the classification of a user account as a confirmed block, an "at risk" accounts or a normal account.

We will use the following metrics to broadly determine the thresholds and compare the different models:

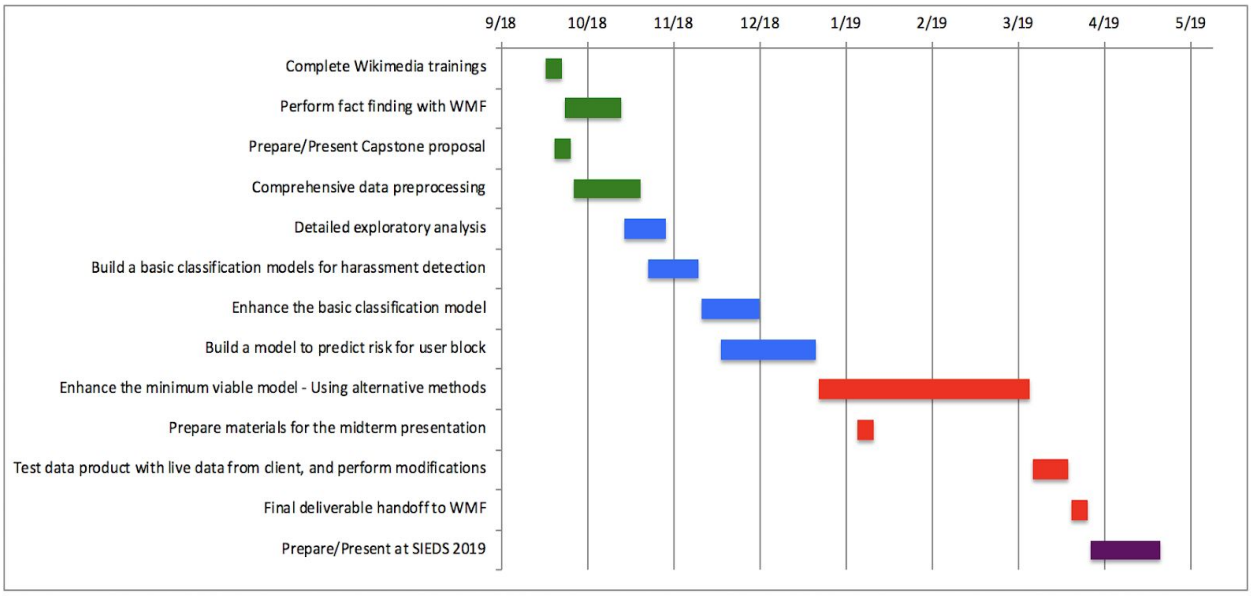
- ***Precision***: the percent of identified accounts that are truly blocked account
- ***Recall***: the percent of blocked accounts that are correctly identified
- ***F-measure***: the weighted harmonic mean of precision and recall

Project Deliverables

- A model that performs classification of Wikimedia user account blocks and predicts and flags users who are at risk of getting blocked in the future.
- A data product in the form of a list containing Wikimedia accounts which merit a block but currently do not have one enacted on them. This will be accompanied with evidence showcasing the basis on which the block is recommended.
- Proposal
- Data and Feature Engineering Presentation
- Technical Paper

- Presentation of research at the 2019 Systems and Information Engineering Design Symposium (SIEDS) Conference in Charlottesville, Virginia
- Research Poster
- Paper Detailing Ethical Implications of this project
- Data and Code Artefacts

Schedule



Budget

Item	Count	Cost per Item	Total
SIEDS 2019 Student Registration Fees	3	$ 75.00	$ 225.00
SIEDS 2019 Non-Student Registration Fees (Advisor & Client)	3	$ 150.00	$ 450.00
TOTAL			**$ 675.00**

Team Qualifications

The team is comprised of three students pursuing their master's degree in Data Science from Data Science Institute at University of Virginia. All the 3 students will collaborate and work together on developing this project. The students will work independently as well as a group on completing specific aspects of the project and in producing the project deliverables. The students in this team bring to the table, a diverse

range of skills and ideas, all of which will contribute towards achieving the goals outlined for this project. Below is some information on each student's background.

Charu Rawat obtained her Bachelor's degree in Mathematics from the University of Delhi, India. Post that, she worked in the financial services sector for 3 years as a data analyst. During that time, she was exposed to working with various big and complex alternative sources of datasets to analyze patterns in consumer behavior and leverage predictive analytics onto data to build models that could forecast company and industry wide trends. She is proficient in Python and SQL. She looks forward to contributing her analytical skills and ideas in developing the framework and methodology for this project and to gain a deeper understanding of analyzing user behavior through the medium of data.

Arnab Sarkar obtained his bachelor's degree in Electronics and Communications Engineering from SRM University, India. Following that he worked as a Systems Engineer with Tata Consultancy Services, having General Electric's Power and Water division as the primary client and working on their business solutions over a span of three projects and a period of four years. He has primarily worked on Relational Database Management systems such as Oracle databases and implemented business logic for Business Intelligence and ERP platforms like Oracle's E-Business Suite and Oracle Demantra using SQL and Oracle PL/SQL. Arnab brings his technical and consulting skills to the project and looks forward to learning more about how machine learning can be used to solve human centric problems in online communities.

Sameer Singh pursued a Bachelor of Technology degree in Electronics and Communication Engineering at Vellore Institute of Technology, India. He has prior work experience at OlaCabs, ZS Associates, and IQVIA (spanning four years). At ZS Associates, he performed sales and marketing analytics for pharmaceutical firms leveraging advanced analytics techniques such as multivariate linear regression, logistic regression, decision trees, clustering, and machine learning among others. At OlaCabs, he was part of team creating strategy for Ola's venture into Electric vehicles, and he analyzed Ola datasets across different product categories such as Auto, Car (Rental, City taxi), for top-8 cities in India. He has extensive experience in Python, R, SAS, and SQL. Sameer will be contributing his analytical and technical skills to this project, and would look forward to increasing his knowledge about cyber harassment, and his technical expertise in NLP, Text Mining among others.

References

[1] “Wikimedia Community Engagement Insights, 2018 report.” Available: https://meta.wikimedia.org/wiki/Community_Engagement_Insights/2018_Report/Support_%26_Safety [Accessed: 09/24/2018]

[2] “Wikipedia Official Website”. Available: https://www.wikimedia.org/. [Accessed: 09/22/2018]

[3] “Wikipedia. Wikimedia: Wikimedia Foundation”. Available: https://en.wikipedia.org/wiki/Wikimedia_Foundation. [Accessed: 09/24/2018]

[4] “Wikipedia. Wikipedia Statistics”. Available: https://en.wikipedia.org/wiki/Special:Statistics. [Accessed: 09/23/2018]

[5] M. Duggan. “Online harassment. Pew Research Center, 2017”. Available: http://www.pewinternet.org/2017/07/11/online-harassment-2017/. [Accessed: 09/22/2018]

[6] “Wikipedia. Wikimedia: Wikimedia Harassment”. Available: https://en.wikipedia.org/wiki/Wikipedia:Harassment. [Accessed: 09/23/2018]

[7] “Wikipedia. Wikipedia: No personal attacks”. Available: https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks [Accessed: 09/24/2018]

[8] “Wikipedia. Wikipedia: Requests for administrator attention”. Available: https://en.wikipedia.org/wiki/Wikipedia:Requests_for_administrator_attention [Accessed: 09/22/2018]

[9] “Wikimedia Training. Dealing with online harassment: Fundamentals-Blocking Users”. Available: https://outreachdashboard.wmflabs.org/training/support-and-safety/dealing-with-online-harassment-fundamentals/blocking-users. [Accessed: 09/22/2018]

[10] “Wikimedia Training. Dealing with online harassment: Fundamentals-Revision deletion or suppression”. Available: https://outreachdashboard.wmflabs.org/training/support-and-safety/dealing-with-online-harassment-fundamentals/revision-deletion-or-suppression . [Accessed: 09/24/2018]

[11] “Wikimedia Training. Dealing with online harassment: Fundamentals-Cross-Wiki blocking and tracking”. Available: https://outreachdashboard.wmflabs.org/training/support-and-safety/dealing-with-online-harassment-fundamentals/cross-wiki-blocking-and-tracking. [Accessed: 09/23/2018]
[12] “Wikipedia. Wikimedia: Banning Policy”. Available: https://en.wikipedia.org/wiki/Wikipedia:Banning_policy. [Accessed: 09/22/2018]
[13] Elery Wulczyn, Nithum Thain, Lucas Dixon. “Ex Machina: Personal Attacks Seen at Scale”. Available: https://arxiv.org/pdf/1610.08914.pdf. [Accessed: 09/24/2018]
[14] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. “Antisocial behavior in online discussion communities. In ICWSM, 2015”. Available: https://cs.stanford.edu/people/jure/pubs/trolls-icwsm15.pdf. [Accessed: 09/22/2018]
[15] Ying Chen & Yilu Zhou. “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”. Available: https://faculty.ist.psu.edu/xu/papers/Chen_etal_SocialCom_2012.pdf. [Accessed: 09/24/2018]
[16] Theodora Chu, Kylie Jue & Max Wang. “Comment Abuse Classification with Deep Learning”. Available: https://web.stanford.edu/class/cs224n/reports/2762092.pdf. [Accessed: 09/23/2018]