cReddit: Misinformation Assessment Tool for Posts from Reddit

Anuraag Vankayala
BMW Research Silicon Valley,
2606 Bayshore Parkway,
Mountain View, CA, 94043
Anderson.vankayala@bmw.de

Abstract—
Fighting misinformation on social media is an uphill battle due to the rising scale of user participation and the polarizing modern political landscape. We introduce a content moderation algorithm that penalizes potential mis informative discourse before it is consumed by susceptible audiences. To do so we analyze the comments (n=88,954) garnered from 3 different posts pertaining to U.S political news discussion. We build a dataset containing 42 features of comment, network and user metadata. We fit and tune several Machine Learning models to perform a binary classification of the misleading nature of top-level comments. Furthermore, we reveal patterns in hyper-partisanship, profanity, references to external resources and their correlation to misinformation. We make all our code used to ingest data, extract relevant features, machine learning models, performance evaluation tools available on GitHub [9].

NOMENCLATURE

Table 1

Subreddit	A community for similarly themed "posts" Ex: r/news
Karma	The net score from the sum of crowd-sourced upvotes and downvotes.
Comment	A response to another comment or post (parent)
Thread	A list of comments that share the same parent
Gold	An award made to a comment or a post by another user
Top Level Comment	A comment that is directly in response to the post.

I. INTRODUCTION

Reddit has emerged as a strong social media platform in the United States by ranking as the 5th most popular for all internet traffic and engagement. It also touts a monthly userbase of 234 million unique users [1]. Majority of users use as a news source [2] and the effects of "fake-news" are therefore highly consequential.

The current methodology employed by reddit to suppress deliberately misleading content is that of moderators. Moderators are assigned to subreddits and exercise their privileges to hide and delete comments when they break the content policy [3] (commonly referred to as "*Rediquette*"). However, due to the large number of comments exchanged every second, real-time human moderation is not practical. Currently, fake news and misinformation need to be reported by the community and is much later redacted by the moderators. The sharing of fact-checking content lags misinformation typically by 10-20 hours [5].

It has been proven that community aggregated intelligence, even by laypersons can match professional fact-checkers [4]. We thus define the misleading characteristic by the large presence of net downvotes / karma and stark controversiality. We introduce a classification model for categorizing top level comments into 'misleading' and 'not misleading'.

II. BUILDING THE DATASET

We pick three mega-threads on r/politics, a community geared towards political discourse in the United States.

Table 2

Headline	**Comment Count**	**Unique authors**
Mueller files final report with Attorney General	32,066	15,503
Attorney General Releases Redacted Version of Special Counsel Report	28,163	11,568
Michael Cohen Testifies before House Oversight Committee	28,725	9,491
Total	88,954	32,394

The total user count is not a sum of the unique authors of each post since 4,168 authors have participated in the other post(s). We use PRAW [6] (The python Reddit API Wrapper) to collect historical metadata corresponding to the author of each comment and the most up-to-date score of the comment itself. It is a common pitfall during any post-hoc analysis to miss the deleted comments, as they are not retrievable by the official APIs'. Since PRAW [6] automatically hides several thousands of comments that were removed by moderators or deleted by the user himself/herself, we leverage PushShift.io [7] a Social Media Ingest service that is publicly available to query and retrieve these redacted comments..

The features used for the classification model stem from one of the four feature sets (See Table 3). It is important to note the User Ideology feature set analyzes the activity of the author on partisan subreddits. Our dataset showed interaction of authors across 48,121 unique subreddits. Since reddit has several partisan subreddits and there is no single

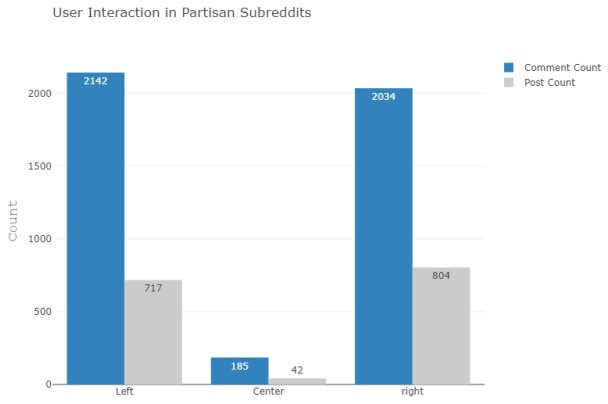
repository listing them, the determination of which subreddit is a partisan subreddit and a classification of 'center', 'left' and 'right' leaning was made with the help of Amazon Mechanical Turk. We create "assignments" ensuring that the classification task of partisanship goes to distinct workers who live in the United States. Each assignment contains 100 subreddits to classify. In each assignment, we provide a link to the subreddit and ask if it is a partisan subreddit or a community for generic discussion (by giving an example for both). If the worker responds by "This subreddit is a partisan subreddit". We then provide 3 options: "left leaning", "right-leaning" and "centrist". Furthermore, we poll the partisanship that the worker self identifies at the end. Using a majority likelihood criterion, we compile a list of partisan subreddits (n=104). The user interaction in partisan subreddits is shown in Figure 1. A list of this compilation is also available in the GitHub resource [9].

.

Table 3

Feature Set	**Features**	**Num Features**	**Explanation**
User	Email Verified, User Age, Total Post Karma, Total Post Count, Total Comment Karma, Total Comment Count	6	Whether user has verified his email The age of the user's account The total post and comment karma received by the user The total post and comment count made by the user
User Post Network	Top Comment Count, Thread Comment Count, Total Comment Count	3	Network features for the author within the post The number of comments made within the thread the total number of comments made in the post the total number of top-level comments made by the author
User Ideology	Left partisan subreddit interaction, Right Partisan Subreddit Interaction, Centre subreddit interaction	12	Historical interaction of the user aggregated in left partisan subreddits, right partisan subreddits and centrist themed subreddits. Each partisan subreddit aggregation has 4 features (post count, post karma, comment count, comment karma)
Subreddit Specific	r/news and r/politics subreddit interaction	8	Historical interaction of the user only in r/news and r/politics subreddits. News and politics each have 4 features (post count, post karma, comment count, comment karma)
Comment Network	Max Depth, Network Size, Network First Children	3	Since comments can be interpreted as trees, max depth is the height of the tallest child comment. Network size is the total number of comments (and sub children, etc.) in the entire tree Network First Children are the number of comments whose parent is the comment being analyzed.
Comment	Num Spelling Errors, Char Count, Profanity, User referenced, Markdown Citation of text, Links Referenced, Sentiment Polarity, Objectivity, Golds, comment Time Delta	10	Tracks the number of spelling errors for the text Presence of profanity, user mentions and URLs If the comment has received "Golds"/" Awards" The time delay between when the comment and post creation The length of the comment The sentiment expressed by the comment
	Total	42	

Figure 1



A.

We build the dataset containing the features listed in Table 3 for all the top-level-comments (n=27,430). We categorize the top-level comments that have been removed by moderators as "mis informative".

III. Modelling

We train our models on SVM with multiple kernel choices, random forest, and with boosted trees: ***XGBoost***. We also evaluate dimensionality reduction tools to breakdown our very large feature set into smaller principal components. Since the spread of our data is non-linear, we use kernel principal component analysis (***kPCA)***, an unsupervised algorithm for dimensionality reduction. We notice that by setting the desired number of components to 2 for higher interpretability, the new features do not capture the dataset accurately. We see a significant loss in accuracy.

By sticking to the original dataset and feature scaling we train our models and show our findings in Figures 3 and 4. To evaluate the performance of our models we capture the cumulative accuracy profile of the base model and compute the Confusion matrix. We also run k-fold cross validation to identify the variance profile of the classifier.

SVM appears to be a very poor model for this classification problem. However, decision trees show promising results as seen in the CAP Curve (Cumulative Accuracy Profile) in Figure (3). We use ***GridSearch*** a popular Model Selection Tool for hyper-parameter tuning. The models referred to in Figure (3) are the tuned models. The best model that was computed was ***XGBoost*** with a learning rate of 0.21 and max-depth 4.

Since most of our dataset does not contain misinformation, a very high accuracy rate tends to have little meaning. We show our type I and type II errors from the tuned and un-tuned models in Figure 4. We notice that model selection gave us 50% lower false positives and negatives with ***XGBoost*** while keeping the maximum height of the tree same.

Figure 3.

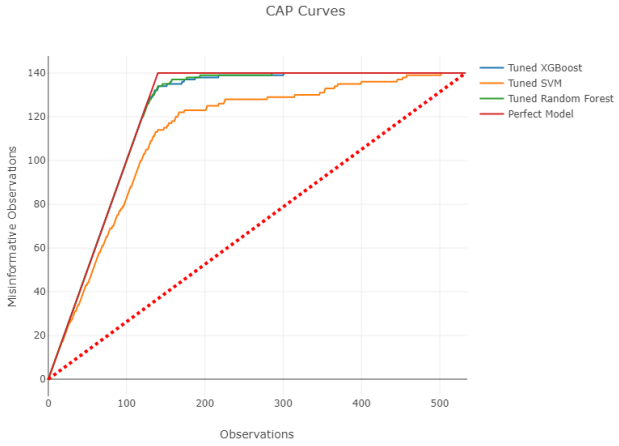
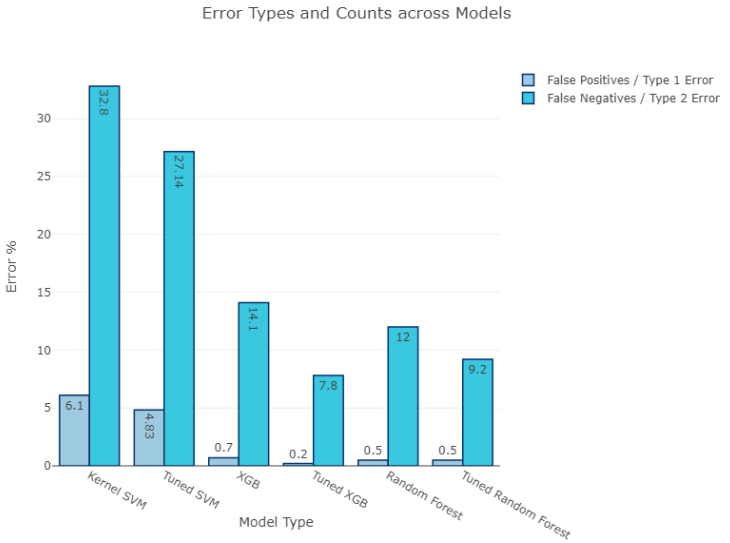


Figure 4



IV. Inference

The data in Table 4 shows the top 5 features sorted by their importance (descending) from our classifier. There are a few nuances that explain the very low importance of some features. One example is the "Gold" feature, which takes a non-zero value only in 32 comments of our entire dataset. Since "gilding" (act of awarding gold) costs real money, this variance is expected. We also notice that the significant contributing features are historical metadata collected from the user. Furthermore, participation in partisan subreddits is a significant factor towards the propensity for dissemination of misinformation. This finding aligns with previous research [8] on political discourse on reddit which discusses the civility or the lack thereof in partisan subreddits during the 2016 Presidential Elections.

Table 4

Feature	**Explanation**
Politics subreddit comment karma	Karma accumulated by commenting in r/politics subreddit (political news discussion)
Right Subreddit Comment Karma	Karma accumulated by commenting on right wing subreddits
Left Subreddit Post Karma	Karma accumulated by posting content on left wing subreddits
Right Subreddit Post Karma	Karma accumulated by posting content on right wing subreddits
Network Top Level Comment Count	The number of top-level comments made under the same post.

V. CONCLUSION & FUTURE WORK

Through our findings, it is apparent that misinformation is typically spread by authors who have historically posted controversial content via comments or posts pretending to be news (table 4, feature 0). We believe that cReddit can help classify misinformation with a good accuracy. This is especially useful during the onset of a crisis, where fact checking is slow [5] and content moderation is essential. cReddit can be used as a plugin within Reddit or by the end-user to automatically hide potential misinformation. Since cReddit has a very low false positive rate (0.5%), cReddit will not come in the way of posting credible content.

Since cReddit requires significant user metadata that is historical, it will vastly benefit when paired with an in-memory data-store such as Memcached. cReddit can be made more accurate by tracking the URLs referenced in the comment and it's crowd sourced reputation as tracked by Web of Trust [9]. Since users' loyalties and political ideologies shift over time, cReddit can be trained to learn patterns showing change in trends and alliances. One way to achieve this is by weighing content temporally closer to the comment more than the ones further in the past. cReddit restricts its' scope to top-level comments since they are the most visible. This restriction can be eliminated by re-building and normalizing certain network related features in the dataset.

VI. REFERENCES

[1] A. Mitchell, (2016, Feb.) Seven-in-Ten Reddit Users Get News on the Site, Pew Research Center, Journalism & Media, Available at https://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

[2] Alexa Rank, (2019, May) Reddit.com Traffic, Demographics and Competitors Available at https://www.alexa.com/siteinfo/reddit.com

[3] Reddit Content Policy, Available at https://www.redditinc.com/policies/content-policy

[4] P. Gordon, D.G Rand (2018), Fighting Misinformation on social media using crowdsourced judgements of news source quality

[5] G.C Luca, F. Alessandro, et. al (2016), Hoaxy: A Platform for Tracking Online Misinformation

[6] The Python Reddit API Wrapper, Available at https://praw.readthedocs.io/en/latest/

[7] PushShift.io https://pushshift.io/api-parameters/

[8] S. Brian, P. Gill, N. Rishab Online Political Discourse in the Trump Era Available at https://arxiv.org/pdf/1711.05303.pdf

[9] cReddit Repository https://github.com/andersonpaac/Creddit

[10] Web of Trust https://www.mywot.com/en/aboutus

Figure 1

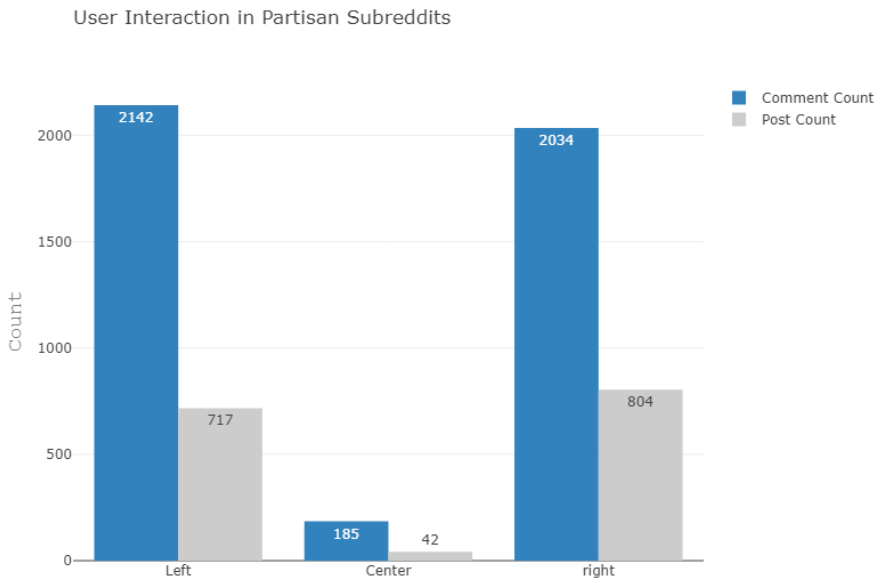


Figure 2

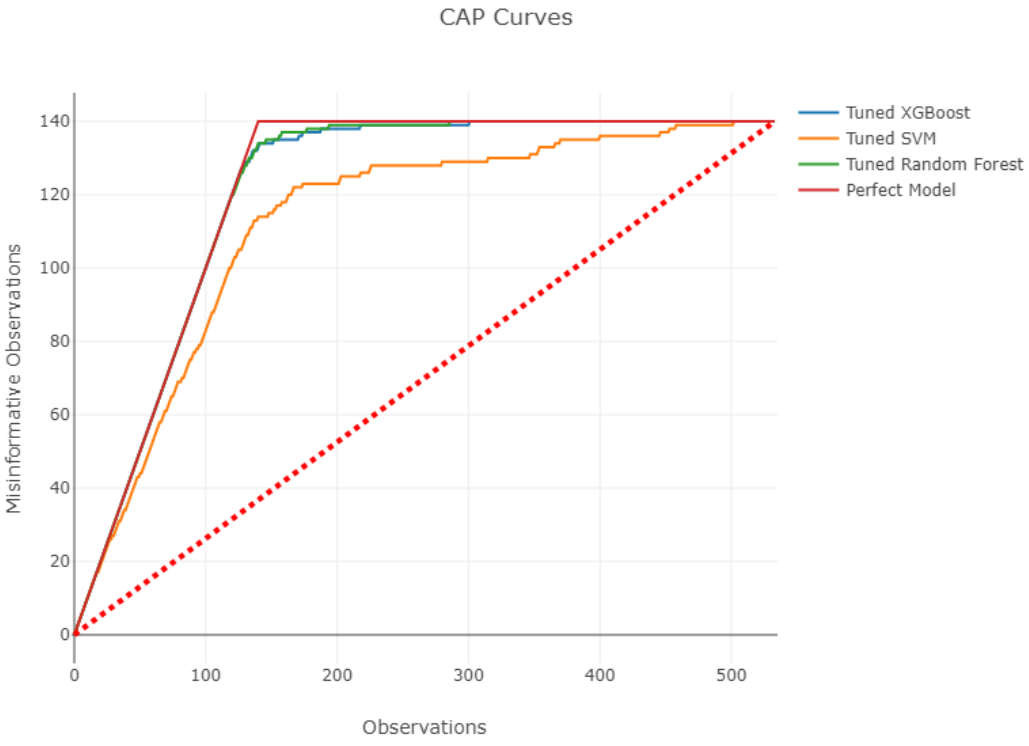
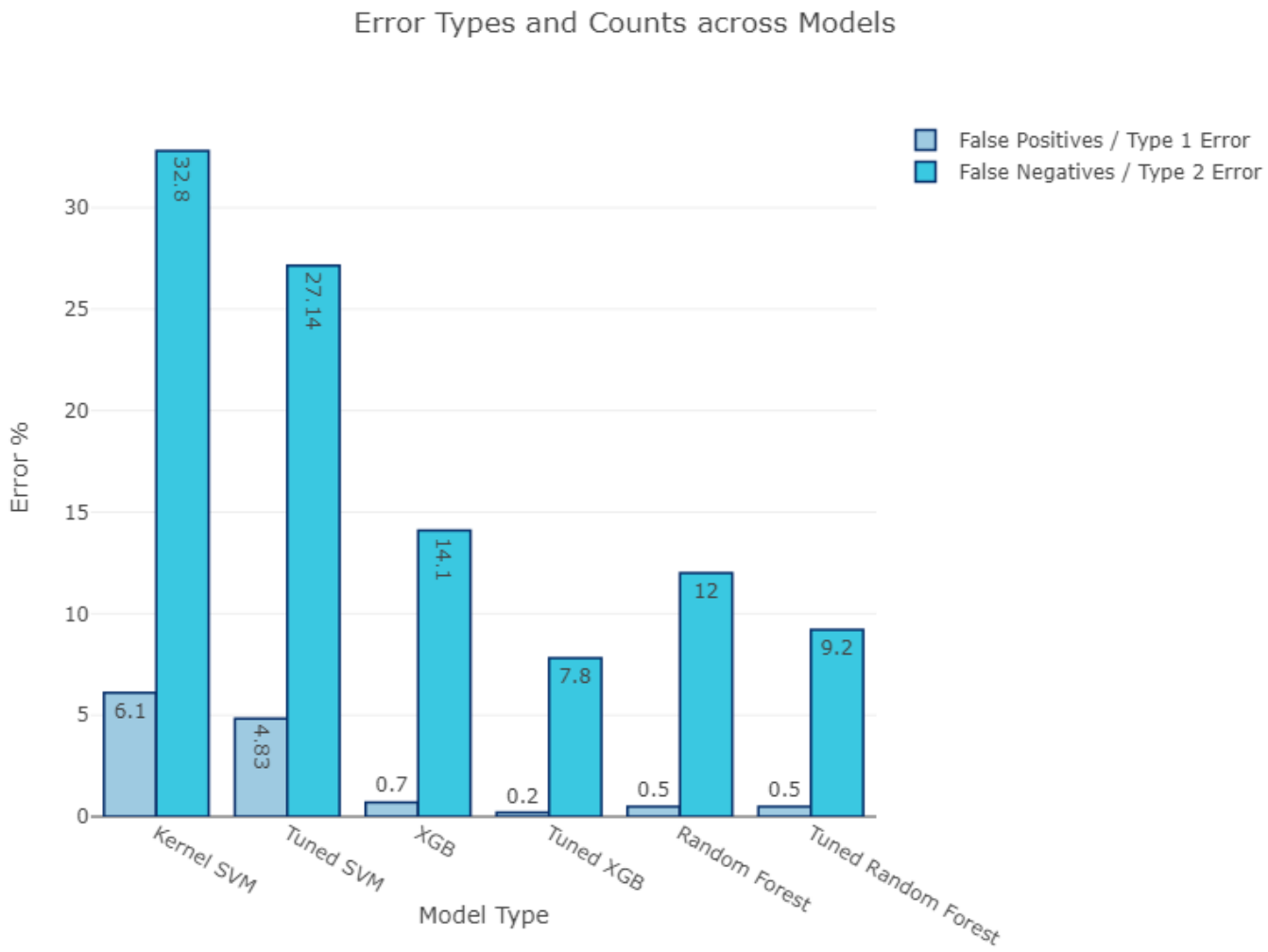


Figure 3


Error Types and Counts across Models
False Positives / Type 1 Error
False Negatives / Type 2 Error
Error %
30
25
20
15
10
5
0
32.8
27.14
14.1
12
9.2
7.8
6.1
4.83
0.7
0.2
0.5
0.5
Kernel SVM
Tuned SVM
XGB
Tuned XGB
Random Forest
Tuned Random Forest
Model Type