Systematically Understanding Reddit User Activity through Tensor Decomposition and LDA Topic Modeling Pipelines

Arzhang Valadkhani
UC Riverside
avala022@ucr.edu

Ben Treves
UC Riverside
btrev003@ucr.edu

Md Rayhanul Masud
UC Riverside
mmasu012@ucr.edu

Michalis Faloutsos
UC Riverside
michalis@cs.ucr.edu

1 ABSTRACT

Reddit is one of the fastest growing and most active social media platforms with high potential for groups cooperating in malicious behavior. The main question we want to determine is if there is a systematic way to compile data from Reddit and identify these malicious groups. We propose SubredditScanner a tool to help understand and find potential malicious activity on Reddit or just analyze overall user interactions and trends. Given a subreddit our tool produces metrics, visualizations, clusters of users and threads, as well as key topics discussed. Prior works aim to look more into the sentiment of user activity and few aim to look at summarizing what groups are discussing. We run SubredditScanner on a sample subreddit. (a) Acquire complete data sets from input subreddits, (b) find sharp drops in activity through statistical analysis and visualization (c) find potentially malicious clusters (d) use models to find key topics, and (e) narrow are results optimize our models to help further find more activity. Our work is a significant step into the ever-improving field of modeling group user behavior at scale.

2 INTRODUCTION

One of the most popular social media platforms today is Reddit. According to Reddit themselves they have 430 million monthly active users. Inside Reddit, users can conglomerate and join communities where in they can discuss specific subjects ranging from computer science, stocks, funny videos, etc. The interactions and activities of users is incredibly interesting and powerful. We've seen communities band together invest and pump stock prices up as much as 3000% adding more than 19 billion dollars to Gamestops market cap. This is just one of the many examples of the strength and influence within the herd mentality of these communities in Reddit.

The problem we tackle in this paper is identifying groups of users across subreddits. The input is a subreddit of choice, the output is user groups and the associated topics they discuss.

The communities of interest are those in the sphere of cybersecurity, hacking, and information security. These subreddits offer a platform for users to share information, ask for advice, and discuss emerging threats. However, moderation is done by users and bots which leads many of the forums to be heavily unregulated. Due to this nature forums are not immune to malicious users who may use them for nefarious purposes. A popular method for identifying sub groups in forums is clustering. The problem is finding the optimal number of clusters for each forum differs greatly. In this research we aim to offer methods for choosing the optimal number of clusters. Due to the dynamics of each subreddit, forum, and platform differing greatly, the number of important clusters will differ greatly as well. Therefore, it is incredibly important to test and compare the performance of different clusters sizes. Furthermore, the research also aims to find the key topics discussed in the clusters by utilizing Topic Modeling. By implementing NLP methods through topic modeling a higher level understanding of what users are discussing can be interpreted.

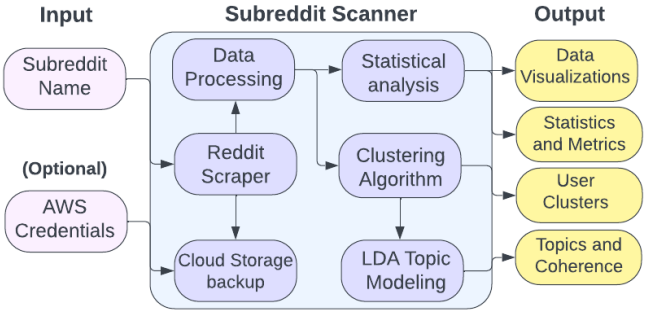

Figure 1: Visualization of the data pipelines and models in SubredditScanner

Identifying relevant groups of users is a non-trivial task, because of how dynamic and dimensional user activity is. Furthermore, clustering documents and text summarization through machine learning is a very nuanced task. Training and adjusting the parameters of an LDA model is an iterative process and run time for large sets is quite long. Most previous works on wide scale Reddit analysis don't have data dating back to the creation date of subreddits while ours does. Furthermore, most recent works utilize sentiment analysis as a means to classify posts [1]. On top of this even fewer works filter out to analyze by using clustering algorithms.

As our key contribution, we introduce SubredditScanner, a comprehensive approach for systematically understanding Reddit group user activity at scale. Specifically, our approach consists of: (a) Data scraping and processing (b) Statistical Analysis (c) Tensor clustering (d) LDA topic modeling (e) Extracting results.

A. Data scraping and processing. In this module we scrape subreddit specific data . Then we process, clean, and store it through various ways so that it can be pipelined for different models.

B. Statistical Analysis. This module produces various averages, CCDFs, and visualizations in order to make high level inferences on the given subreddit.

C. Tensor clustering. In order to identify groups of users we use a clustering tool Tenfor. [2] We use methods to find a correct rank for our model in the next step of our pipeline.

D. LDA topic modeling. In this phase we evaluate the clusters extracted from Tenfor by applying them to our LDA model. A summarization and key topics are produced from the cluster and each topic is given a coherence score.

E. Extracting results. We use visualizations to compare and contrast the coherence score from our LDA model. This evaluation step helps identify rank sizes that produce quality clusters.

We demonstrate SubredditScanner's potential to help understand activity on subreddits.

Our key results are:

A. 2.9 million posts across 6 subreddits We process, clean, and store data from 6 subreddits ranging in size from high to lower in popularity.

B. Metrics and data visualizations. Averages, statistics, graphs pertaining to a given subreddit are presented to a notebook for easy access.

C. Clusters of Groups of Users and Posts. Lists with key groups containing thread ids, usernames, and associated dates are outputted. The lists are put into JSON files for scaliblity.

D. Topics and Coherence Score. A list of the relevant topics is outputted which contain the words that occur most often. Found groups of users talking about things like child porn amongst many other things.

E. Extracting results. Through comparisons of coherence scores and rank sizes we determine the ideal cluster sizes to examine. Any score above 0.60 is strong and offers opportunity to identify activity. We reached a peak score of .0725 in r/Hacking_Tutorials.

3 DATASET

In this section, we provide a brief description of the data set. We use compiled data from 6 total subreddits, basic statistics listed below in table 1. Subreddit data is classified into two different categories: submissions and comments. Submissions are posts made on a subreddit while comments are simply responses to submissions/posts.

Table 1: Dataset Overview

Subreddit	Active Users	Comments	Submissions
r/Hacking	171976	683306	158513
r/cybersecurity	96898	542103	102115
r/blackhat	10189	24114	8316
r/Hacking_Tutorials	30002	71725	21348
r/computerforensics	11625	62254	10760
r/privacy	143279	1048575	183393

4 METHODOLOGY

In this research we used a multitude of methods to systematically gain a higher level of understanding as to the activities going on in various subreddits. We broke down the various steps of the wide scale analysis into three key phases. In this research we used a multitude of methods to systematically gain a higher level of understanding as to the activities going on in various subreddits. We broke down the various steps of the wide scale analysis into four key phases. (a) Data Collection, (b) wide scale statistical analysis, (c) clustering using TenFor, (d) running Topic Modeling on given clusters. Through these four processes we have essentially built a framework to acquire understanding of user activity at a large scale.

A. Data Collection.

The first and most important phase in our research was data collection which came with a set of its own challenges. Scraping websites at scale can be a challenging and time consuming task. Many platforms and websites are quite against users being able to mine and scrape their data. Reddit's API is only made to scrape subreddit at very small scales. However PushshiftIO is an open source one maintained by a community of researchers. In this research a data scraping script was developed in order to create data sets with quality processed data. Furthermore, this script was developed to scale, and to extract data from as early as the inception of subreddit and as recent as the present day. Posts are differentiated into two categories, Submissions and Comments. Submissions are a essentially the parent thread while comments are any posts proceeding the parent thread. The script also merges and sorts by the Post ID. See figure below for communities of interest. We offer further scalabiltity of data by making it optional to pipeline storage to Amazon S3. This allows for users who want to make data accessible through the cloud possible.

Table 2: SubredditScanner's Scraping Capability

Subreddit Name	Number of Users	Creation Date
r/Hacking	2,609,255	4/26/2008
r/cybersecurity	457,508	5/22/2012
r/blackhat	80,184	2/12/2009
r/Hacking Tutorials	225,991	9/3/2012
r/computerforensics	56,429	6/14/2010
r/howtohack	387,414	3/5/2016
r/privacy	1,295,941	3/21/2008

B. Statistical Analysis.

In the second phase the main goal is to run analysis to gain a high level understanding of activity in each community. The basics statistics we gather include calculating and identifying the top active users, top active threads, the top active comments, and the number of inactive users. Manual inspection is done on each of the top fields in order to possibly identify foul play or malicious intent. The next set of analysis completed are several CCDFs paired with visualizations to gain further understanding of inter user activity as well as possible key events.

A visualization of the number of posts per month on every subreddit helps us identify peaks in user activity. The CCDFs used include the time over the percentage of total posts made from the start of subreddit to its most current time. Next a CCDFs of user posting frequency, and one for the number of comments on post along the percentage of posts with that number of comments. The analysis is streamlined very efficiently from the data collection stage allowing strong reusability. Visualizations and calculations were all done on Jupiter Notebook using Python.

C. Clustering using TenFor.
In the third phase we utilize the a machine learning clustering tool that utilizes tensor decomposition written by Risul Islam in the paper TenFor [2]. TenFor takes into account three fields of interest those being Date, Username, and Thread ID. While TenFor has been shown to work well and identify meaningful clusters its use comes with some challenges as well. Changes to the code must be made in order for it to fit a given set of data. Furthermore, an implementation of sparse matrix was introduced in order for the script to work on bigger data sets. Despite this further optimizations still need to be made as TenFor is a bottleneck in the pipeline of our analysis. Output of the clusters have been translated and organized appropriately into JSON format for more use in the future. This allows for TenFor's data to be more scalable and data to be usable for other uses. In our case the data is pipe lined for use for our next module.

D. Running Topic Modeling on output clusters.
In this stage of the pipeline we analyze the clusters that TenFor has produced. In order to summarize what users are specifically talking about we take an NLP approach and introduce Latent Dirichlet Allocation Topic Modeling algorithm to our resulting clusters. We compile the post content from every single cluster into its own document. Every post has counts as one document in our model. Standard NLP text processing is then applied to our list of documents. Every single word is lower cased, and the all the words in the document are then tokenized. Numbers are removed but words containing numbers aren't as they might contain important information. Furthermore, we filter out documents with words that occur in more than 50 percent of them, and words that occur in less than 20 documents. We then apply lemmetazation to the remaining documents, and the from there we apply it to our LDA model implemented through the Gensim library.

With text processing completed we iterate, and use different clusters as input in order to check the coherence and accuracy of our model. By increasing the number of potential topics we test and look at potentially how many different subjects users are talking about in our cluster. Through manual investigation, we conclude that a coherence score of above 0.6 is required for generating meaningful clusters.

E. Extracting Results
In the last module of our pipeline we compare and contrast different ranks and cluster sizes. We make visualizations comparing the coherence scores on various topics across multiple ranks. This helps us identify if a cluster is meaningful or not. Furthermore we analyze the average number of users, threads, and dates across in clusters across all ranks completed. Through this we can gain insight into what an optimal group of users, threads, or dates look like.

5 RESULTS

We completed our methodology from beginning to finish on one subreddit. However we completed section A. and B. respectively on all subreddits. We demonstrate SubredditScanners abilities completely on r/Hacking_Tutorials below are the key results we found.

A. Data collection and processing. We discuss the results of our Reddit data scraper along with our processing.

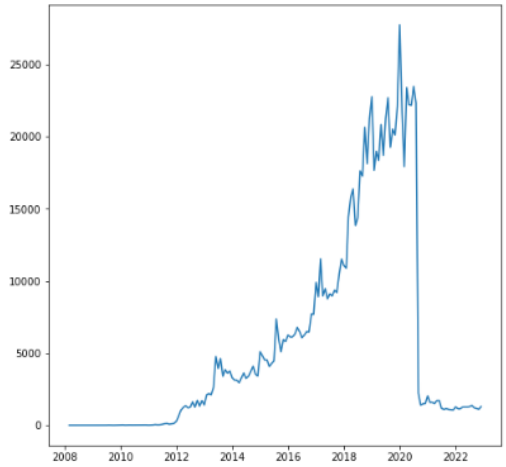

Figure 2: The number of posts over of time. Note the sharp decline in early 2020.

In the data scarping portion of our research we amassed a total of over 2.9million posts total. We had compiled a total of 2,432,077 comments and 484,445 submissions respectively. The total number of users compiled totals at 463,969 active users. Refer to 2 for an in depth breakdown for each subreddit. Important to note is our data is extensive and complete from the very inception of the subreddits themselves. As shown in Table 1 etc.

We processed the large swaths of data, and removed deleted users from our data set. We narrowed down our initial raw JSON data with 57 dimensions to 9 in our data model.

Table 3: Data Model

Post ID	Comment ID	Username	Body
Subreddit	Date	Score	URL

B. Statistical Analysis. In our statistical analysis module we find various interesting results. Across all subreddits we find the top most active users. In most cases the most active users were also moderators. Although we focus on running our pipeline completely on r/Hacking_Tutorials we found interesting results in r/privacy subreddit. Pictured in figure 2. In March 2020 we can see a drastic drop in posts and user activity. We hypothesize that stricter rules on posting, and greater increase in moderation led to the sharp decline. In comparison to Hacking_Tutorials, pictured in figure **??**.

C. Tensor clustering. After running TenFor on our subreddit as output we get meaningful clusters based on the rank. We found in the m below that meaningful clusters tend to have higher coherence. We examined the average size of clusters across the three fields: users, threads, and dates. Interestingly, we found no trends in the size of clusters as rank increases.

D. LDA topic modeling. In the LDA topic modeling module we found various interesting events and measurements. A key topic discovered in a rank 7 cluster was users discussing about child pornography. The key words identified were 'CP' an abbreviation for child pornography and 'porn'. Upon further inspection the

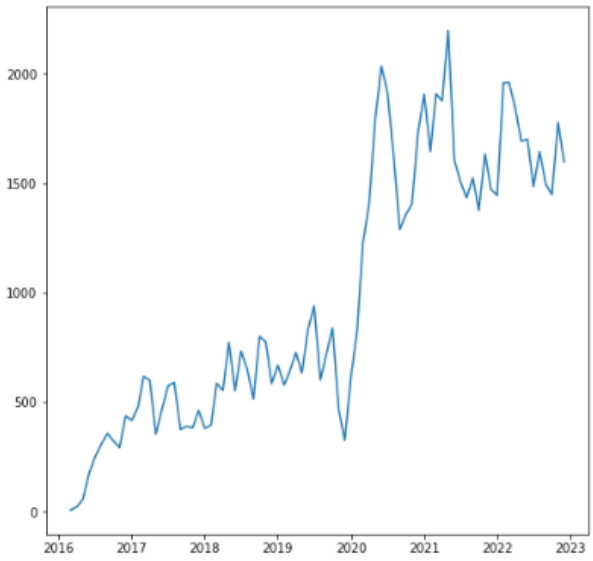

Figure 3: The number of posts over of time for Hacking Tutorials generally a natural progression of growth in activity.

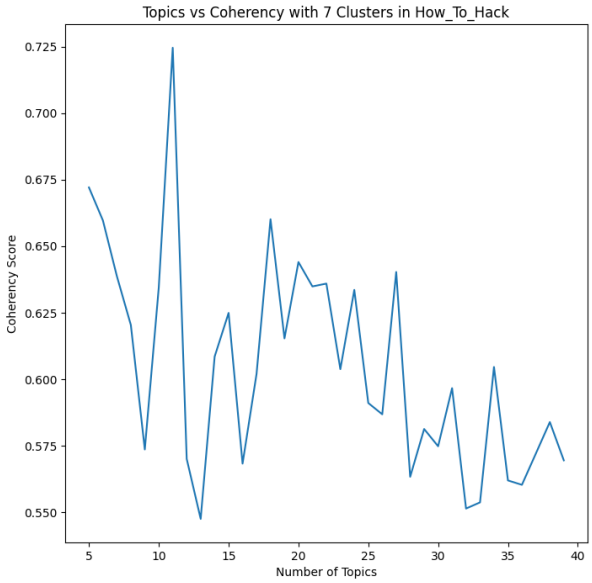

Figure 4: At 10 topics we have a score of .725 indicating a very coherent model and cluster.

conversation was benign, however it demonstrates that our model is capable of identifying cooperative malicious behavior. In the future we plan to implement a module that automatically finds malicious key words in clusters.

E. Extracting results.

In this module we go onto examine some of the data visualizations that we produce after running LDA on our clusters. The coherence score spikes to 0.725 at 10 topics for a cluster of rank 7 as seen in 4. This is the same cluster we used, and were able to find the potential malicious activity as stated in the results above. Furthermore, we cross examine the coherence score across all the various rank, and output there visualizations as well.We also found rank 9 to have strong coherence scores with the highest 0.68 at 16 topics. We can conclude that in the case of Hacking_Tutorials that ideal ranks are 7 and 9 respectively. Therefore we can conclude that LDA Topic modeling is good at measuring how meaningful clusters can be.

6 DISCUSSION

Through developing SubredditScanner a lot of thought and development was put into making open to possible changes. A large community of researchers, data scientists, and hobbyists work on ways to systematically understand user activity on forums. We want left the door open to scale SubredditScanner further and open source the project. Currently our work is available on GitHub for use.

Another means to improve our NLP approach with LDA topic modeling is introducing further scoring measures. UMass coherence scoring is another popular metric to measure the coherence of models. We want to implement UMass coherency at some point in the future to further measure the accuracy of our model and identify more ideal number of topics and clusters.

7 CONCLUSION

In conclusion SubredditScanner a tool we've built has proven successful at identify trends and user activity on subreddits. SubredditScanner successfully scrapes data at scale and processes the data and stores backups in Amazon S3 if so desired. Furthermore by combining the clustering abilities of Tenfor with the LDA Topic Modeling helps gain deeper knowledge and quantify user activity. The key results demonstrate the success of our methodology by finding users discussing something potentially malicious. In future work we would like to scale our pipelines, and more thoroughly investigate algorithms that could help find further malicious activity occuring.

REFERENCES

[1] Archit Aggarwal, Bhavya Gola, and Tushar Sankla. 2021. Data mining and analysis of reddit user data. In *Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA 2020*. Springer, 211–219.

[2] Risul Islam, Md Omar Faruk Rokon, Evangelos E Papalexakis, and Michalis Faloutsos. 2020. Tenfor: a tensor-based tool to extract interesting events from security forums. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 515–522.