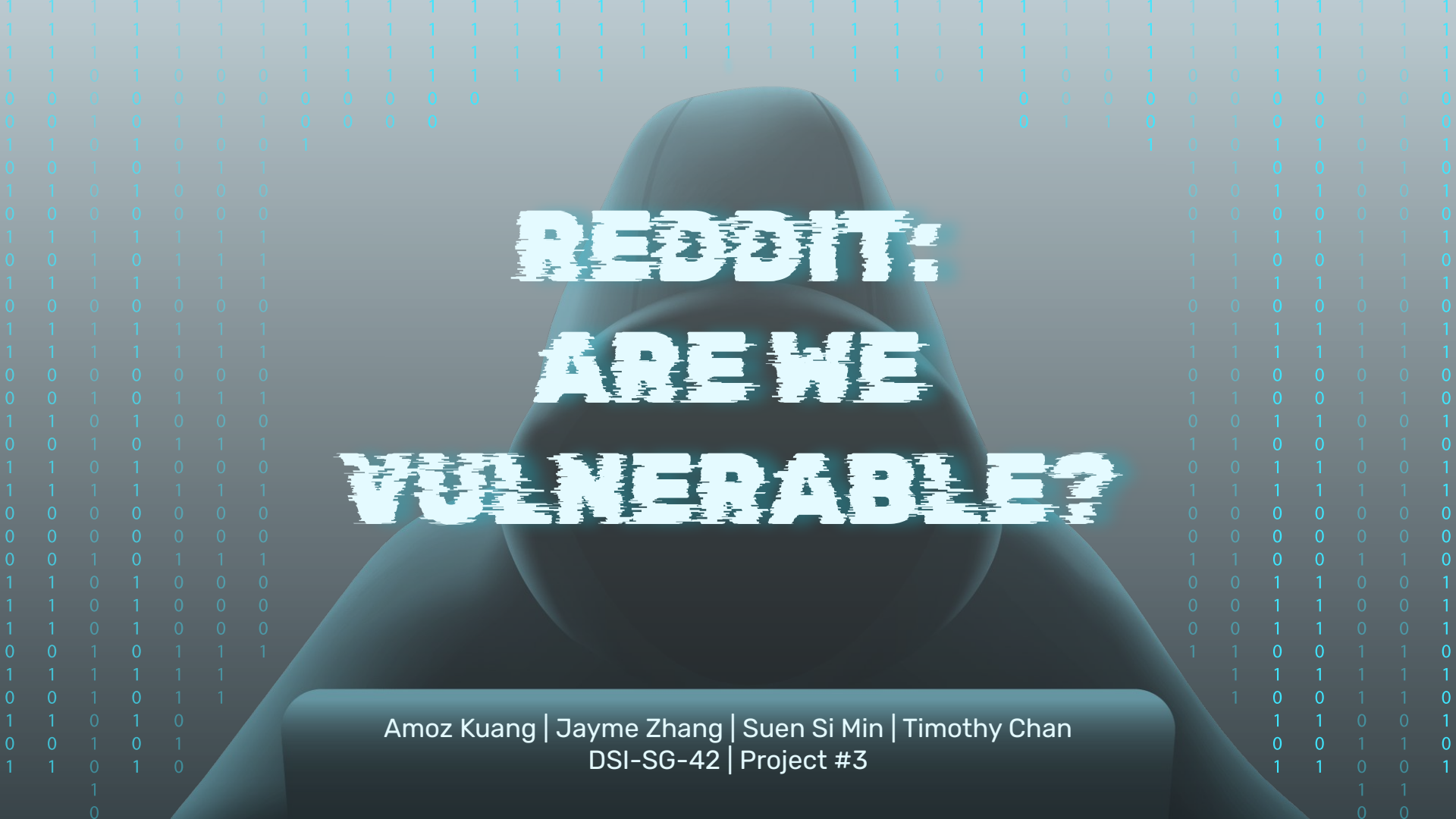


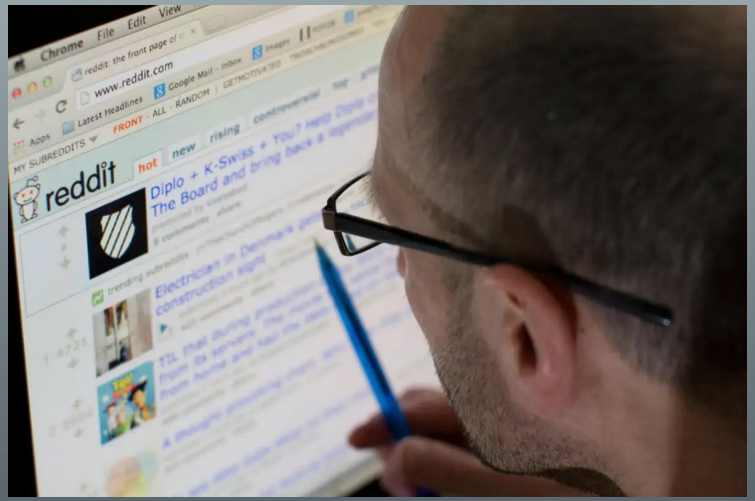
Amoz Kuang | Jayme Zhang | Suen Si Min | Timothy Chan
DSI-SG-42 | Project #3

WHO ARE WE?

NATIONAL CRIME PREVENTION COUNCIL

National Crime Prevention Council (Public Safety)

WHO ARE YOU?

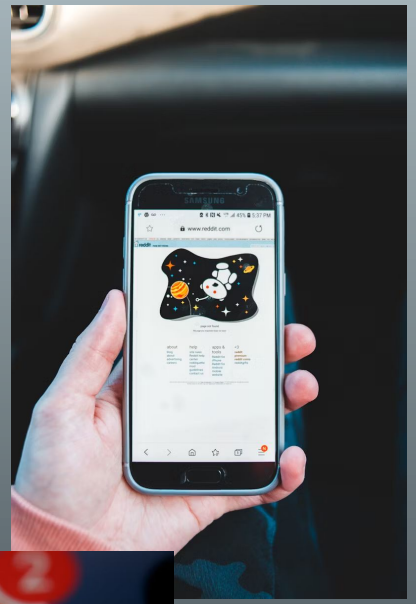


Reddit Moderators

Image Source: NCPC; New Scientist, 2022

ROLE OF A REDDIT MODERATOR

- Objective: Create a safe and relevant space for users
- Flag posts
- Offensive/Unsolicited/Unfavourable /Spam contributions
  - Warnings
  - Remove
  - Ban
- Ensure subreddit-relevant contents
- Constantly revise rules and regulations of subreddit



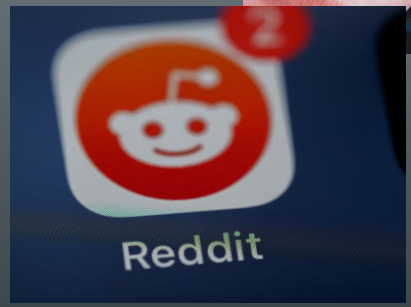


Image Source: Unsplashed

SAMPLE #1: SCAM OR RANDOM ACT OF KINDNESS?

u/KardGuru • Promoted

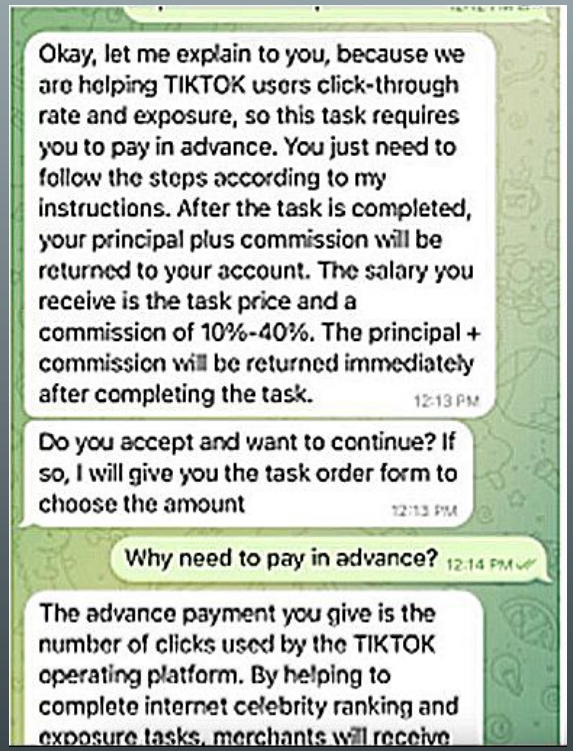
Experience the Fusion of Art and Robotics with 'Hong Kong Machines' Limited Edition Playing Card & Mystery Box - Support on Kickstarter Today!

Tiny Hong Kong Machines Playing Card
qpmarketnetwork.com

Learn More

Sort by: **Best**

SAMPLE #2: SCAM OR RANDOM ACT OF KINDNESS?



AGENDA

01 INTRODUCTION

02 PROBLEM STATEMENT

03 DATA FINDINGS

04 DATA MODELING

05 CONCLUSION

06 LIMITATIONS & RECOMMENDATIONS

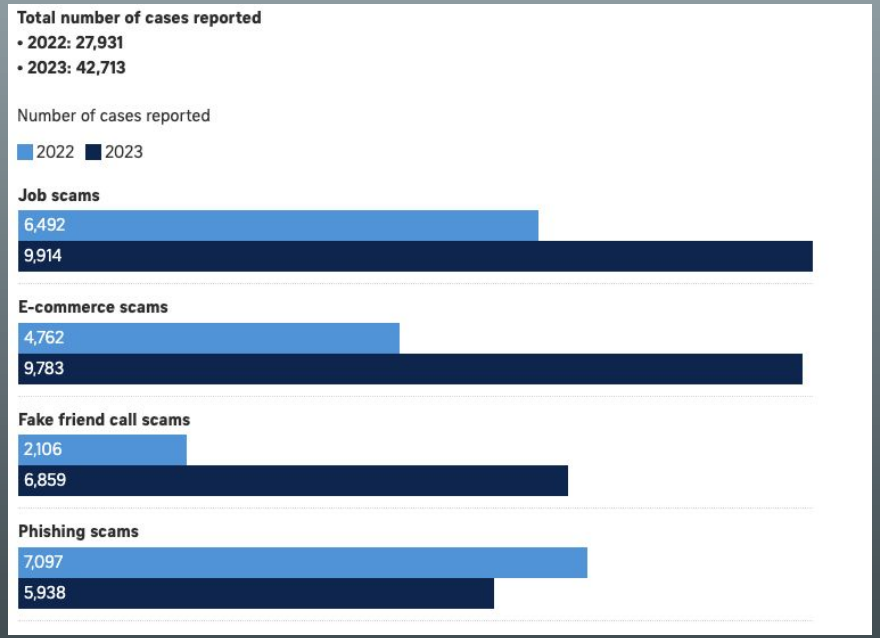
01

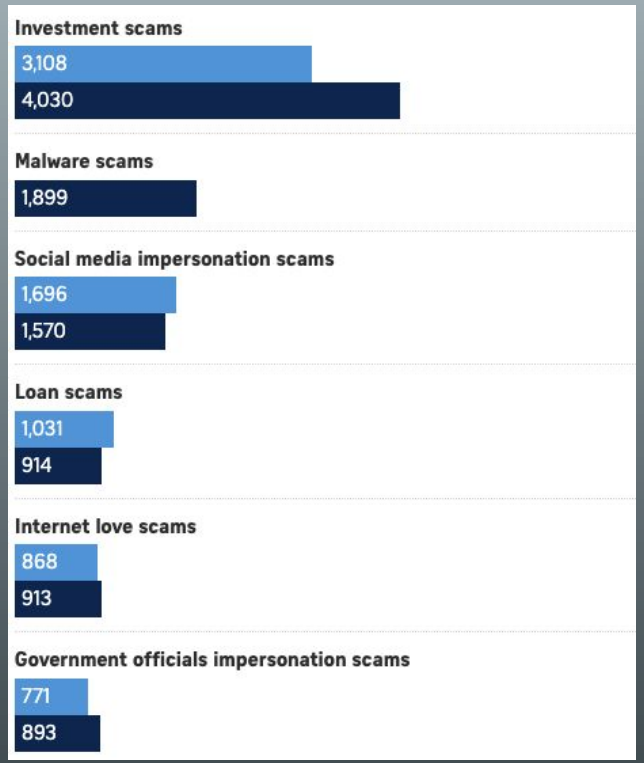
INTRODUCTION

- Defining the Context
- The Persona

CONTEXT

Scam victims in S'pore lost $651.8m in 2023, with record high of over 46,000 cases reported





Sources: The Straits Times, Feb 2024

CONTEXT

A new variant of e-commerce scams emerged in 2023 involving items such as durians and wagyu beef sold online at huge discounts. Victims would lose money when scammers asked for goodwill deposits, reservation fees or delivery fees.

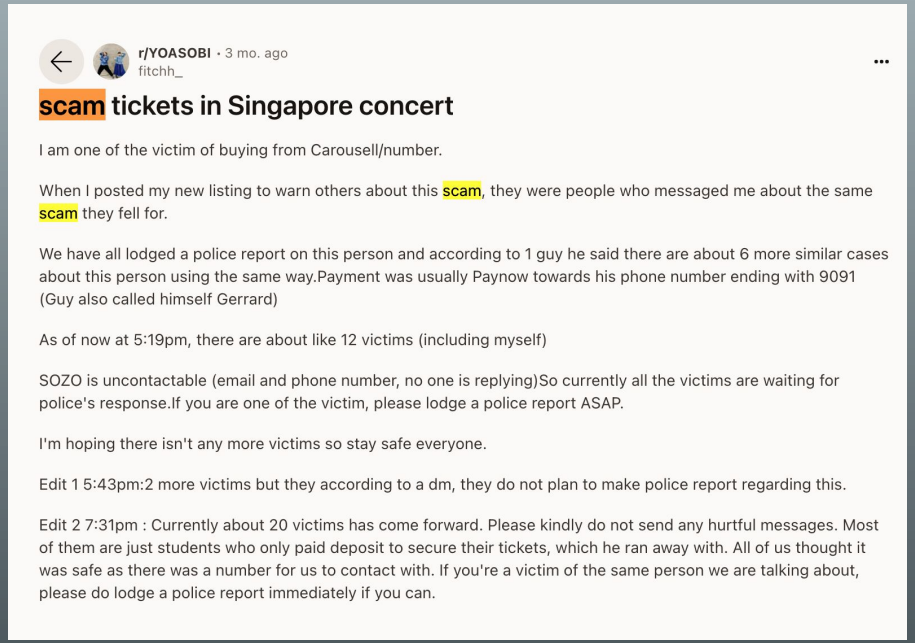
The police said most online scams are perpetrated by scammers based outside Singapore, and such cases are difficult to investigate and prosecute.

Young adults aged 20 to 29 mostly fell prey to job scams, while those aged 30 to 49 mostly lost money to e-commerce scams.

The elderly, aged 65 and above, made up 7.1 per cent of scam victims. More than a third of them fell for fake friend call scams and over 13 per cent for investment scams.

Sources: The Straits Times, Feb 2024

CONTEXT (AS SEEN ON REDDIT)



r/YOASOBI · 3 mo. ago
fitchh_

scam tickets in Singapore concert

I am one of the victim of buying from Carousell/number.

When I posted my new listing to warn others about this scam, they were people who messaged me about the same scam they fell for.

We have all lodged a police report on this person and according to 1 guy he said there are about 6 more similar cases about this person using the same way.Payment was usually Paynow towards his phone number ending with 9091 (Guy also called himself Gerrard)

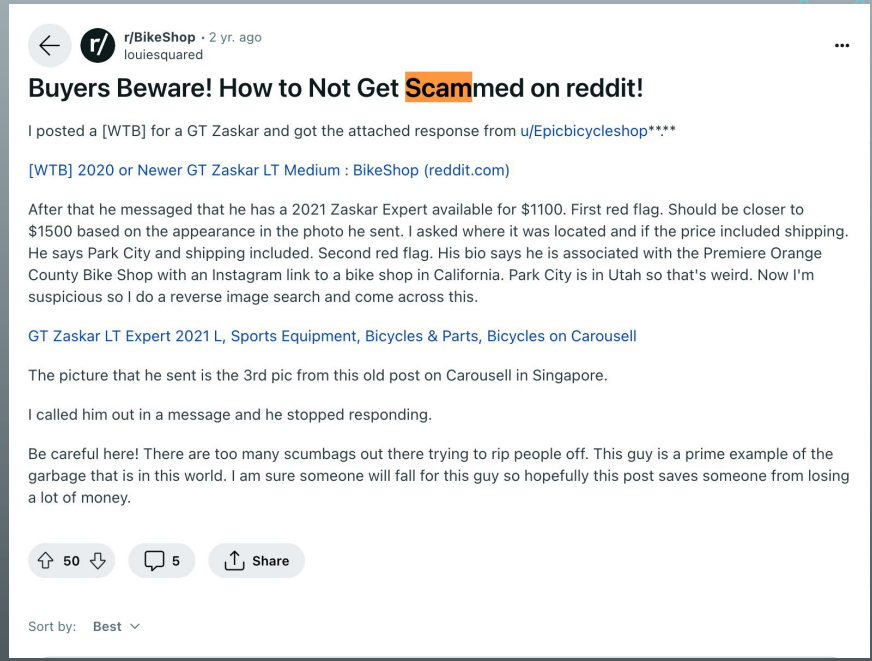
As of now at 5:19pm, there are about like 12 victims (including myself)

SOZO is uncontactable (email and phone number, no one is replying)So currently all the victims are waiting for police's response.If you are one of the victim, please lodge a police report ASAP.

I'm hoping there isn't any more victims so stay safe everyone.

Edit 1 5:43pm:2 more victims but they according to a dm, they do not plan to make police report regarding this.

Edit 2 7:31pm : Currently about 20 victims has come forward. Please kindly do not send any hurtful messages. Most of them are just students who only paid deposit to secure their tickets, which he ran away with. All of us thought it was safe as there was a number for us to contact with. If you're a victim of the same person we are talking about, please do lodge a police report immediately if you can.



r/BikeShop · 2 yr. ago
louiesquared

Buyers Beware! How to Not Get Scammed on reddit!

I posted a [WTB] for a GT Zaskar and got the attached response from u/Epicbicycleshop****

[WTB] 2020 or Newer GT Zaskar LT Medium : BikeShop (reddit.com)

After that he messaged that he has a 2021 Zaskar Expert available for $1100. First red flag. Should be closer to $1500 based on the appearance in the photo he sent. I asked where it was located and if the price included shipping. He says Park City and shipping included. Second red flag. His bio says he is associated with the Premiere Orange County Bike Shop with an Instagram link to a bike shop in California. Park City is in Utah so that's weird. Now I'm suspicious so I do a reverse image search and come across this.

GT Zaskar LT Expert 2021 L, Sports Equipment, Bicycles & Parts, Bicycles on Carousell

The picture that he sent is the 3rd pic from this old post on Carousell in Singapore.

I called him out in a message and he stopped responding.

Be careful here! There are too many scumbags out there trying to rip people off. This guy is a prime example of the garbage that is in this world. I am sure someone will fall for this guy so hopefully this post saves someone from losing a lot of money.

50 · 5 · Share

Sort by: Best

Sources: r/YOASOBI, 2024 ; r/BikeShop, 2022

NEW REDDIT MODERATOR

Bio

- Keith, 30 years old
- Day Job: Engineer
- Volunteer moderator for r/Ticketmaster ~1 year

Attitude & Behaviour

- Tech-savvy
- Scam-Adverse
- Advocate safe online space

Pain Points

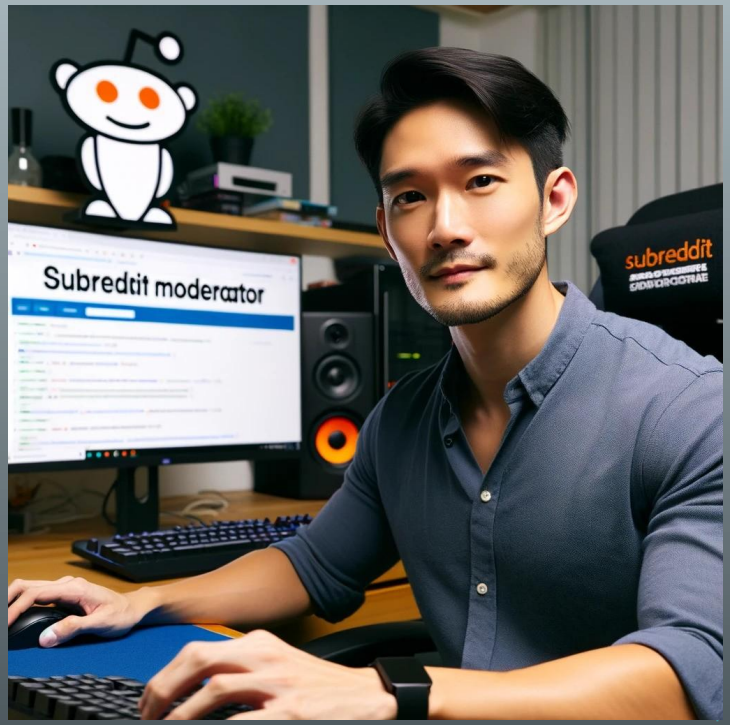
- Currently spends about 3 hours/week moderating

Scenario

He and his friend got scammed on C2C marketplace platform, while purchasing concert tickets

Motivation(s)

- Quick & Easy Moderating (reduce time spent by 50%)
- Less Manual Reviewing of Posts
- Create safe subreddit community



02 PROBLEM STATEMENT

"

How Might We Assist Reddit Moderators to Differentiate Online Scams from Legitimate Acts of Kindness within Reddit Posts using a Sensitive & Accurate Natural Language Processing Model

"

r/Scams

What?

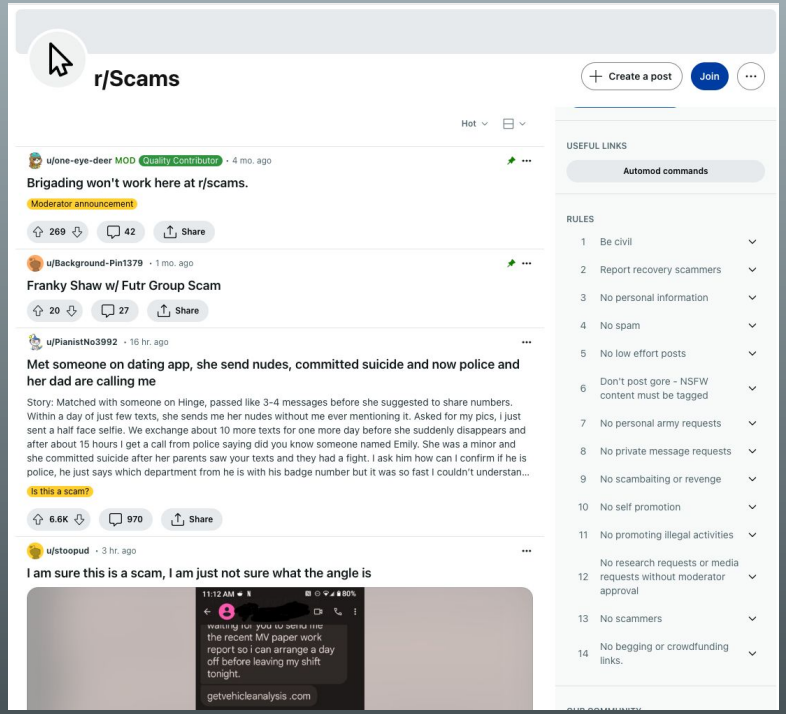
- **People educate themselves, and/or support space for scam victims**

Who?

- **Redditors who needs community confirmation or advice**

How?

- **Seeking advice for scam related content**
- **Sharing their scam experiences**



Sources: r/Scams

r/RandomKindness

What?

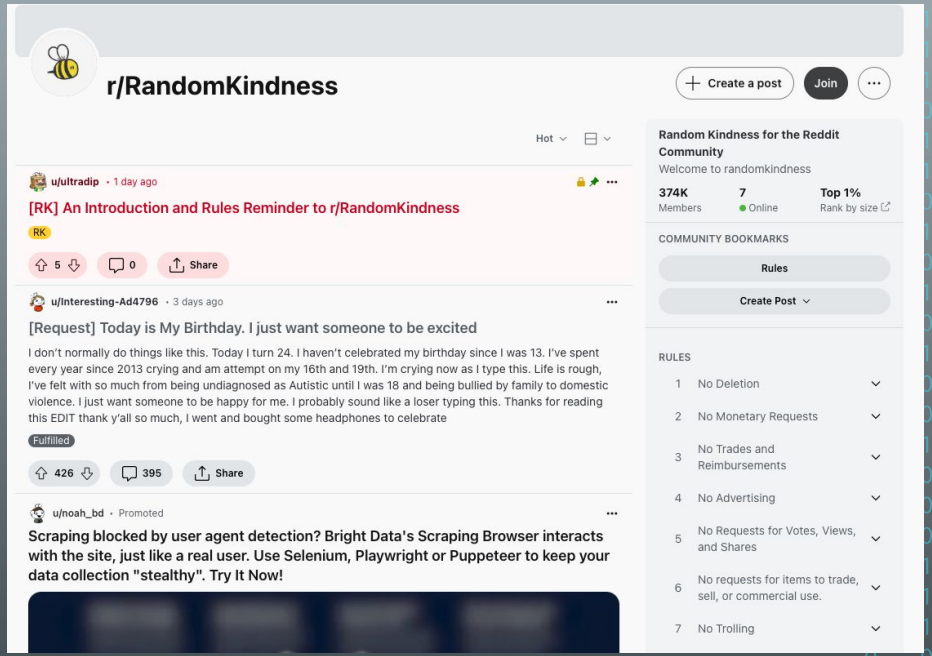
- **Spread the kindness and to ask for goodwill donations/help**

Who?

- **Redditor offering services**
- **Redditor who needs aid**
  - **Monetary**
  - **Grocery**
  - **Services**

How?

- **Create a post or reach out to OP post**



Sources: r/RandomKindness

Why?

r/Scams

- Spreading Awareness about new and old scamming techniques
- Rich in keywords commonly associated with fraudulent activities
- Community helps validate whether a situation is a scam

- Words used in r/RandomKindness may be present, allowing scammers to mimic and exploit good intentions

- An Active Community
  - r/Scams = 638K members
  - r/RandomKindness = 369K members

r/RandomKindness

- Showcases genuine acts of kindness and generosity

DATA PROCESS

STEP 1

COLLECTION

Scrapped 27k+ rows of data from r/RandomKindness, r/Scams

STEP 2

CLEAN / ANALYSIS

Analysis of distribution, and identifying of outliers

STEP 3

PRE-PROCESS

Using n-gram analysis to search for words

STEP 4

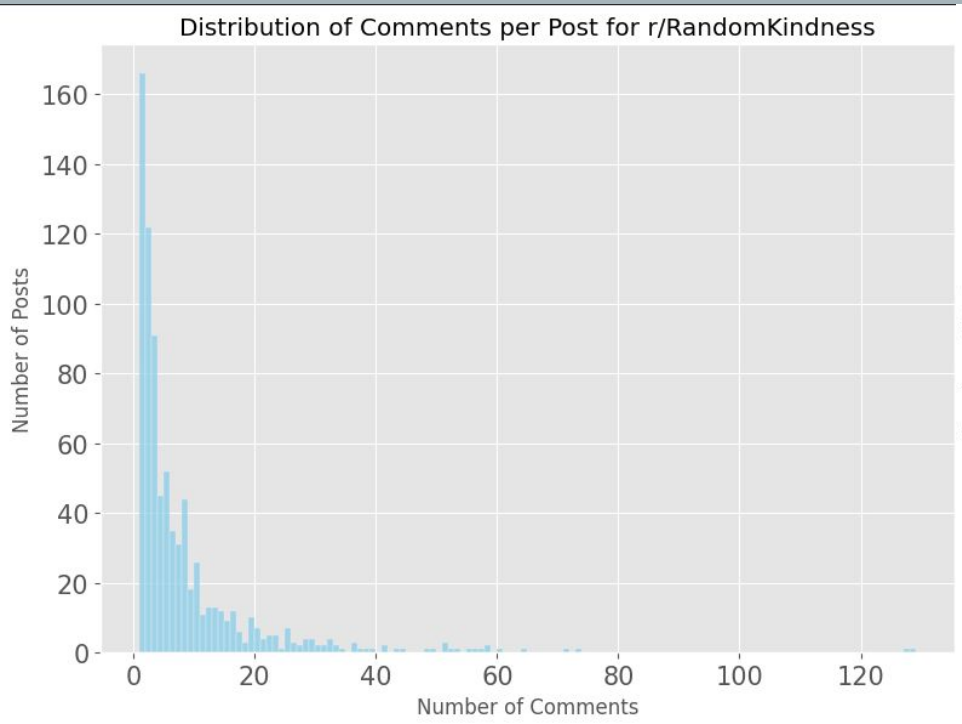
MODELING

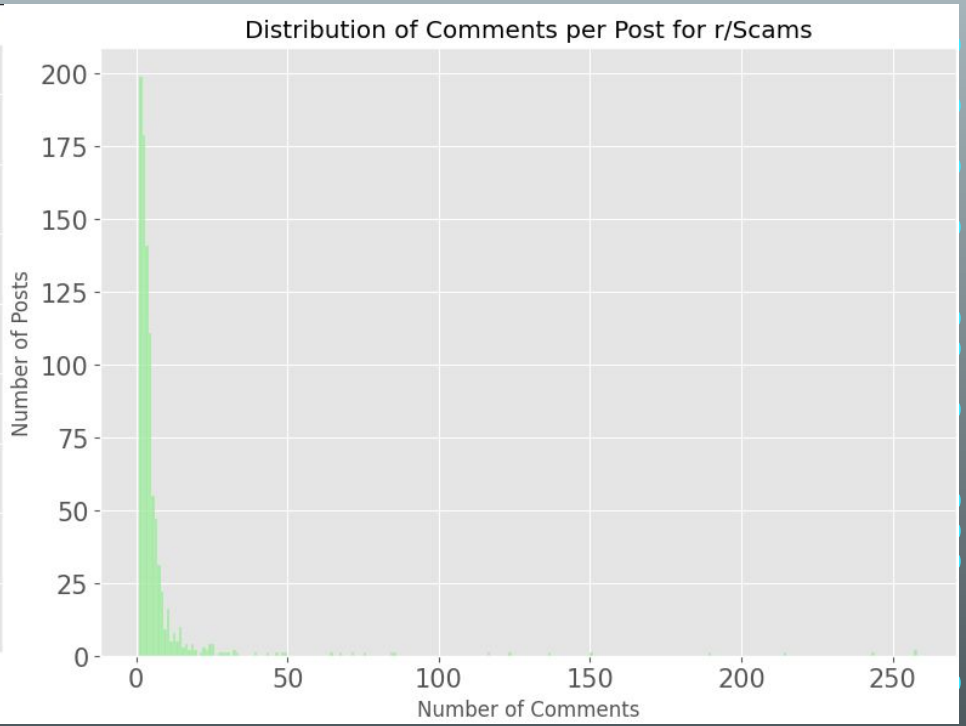
The Classifier-type Model churned out to give better prediction

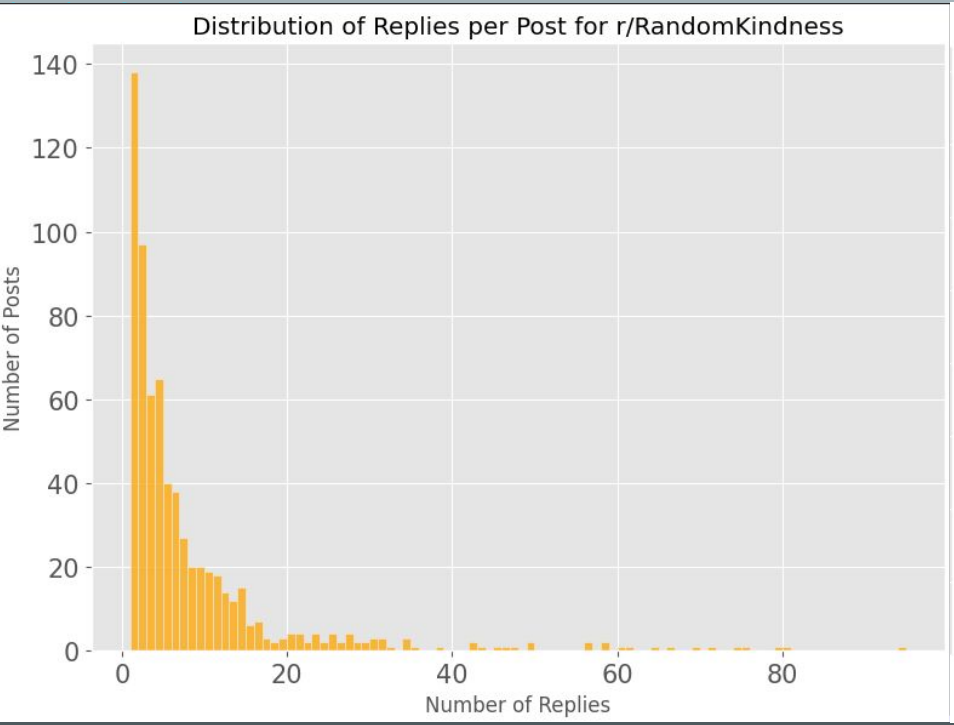
03

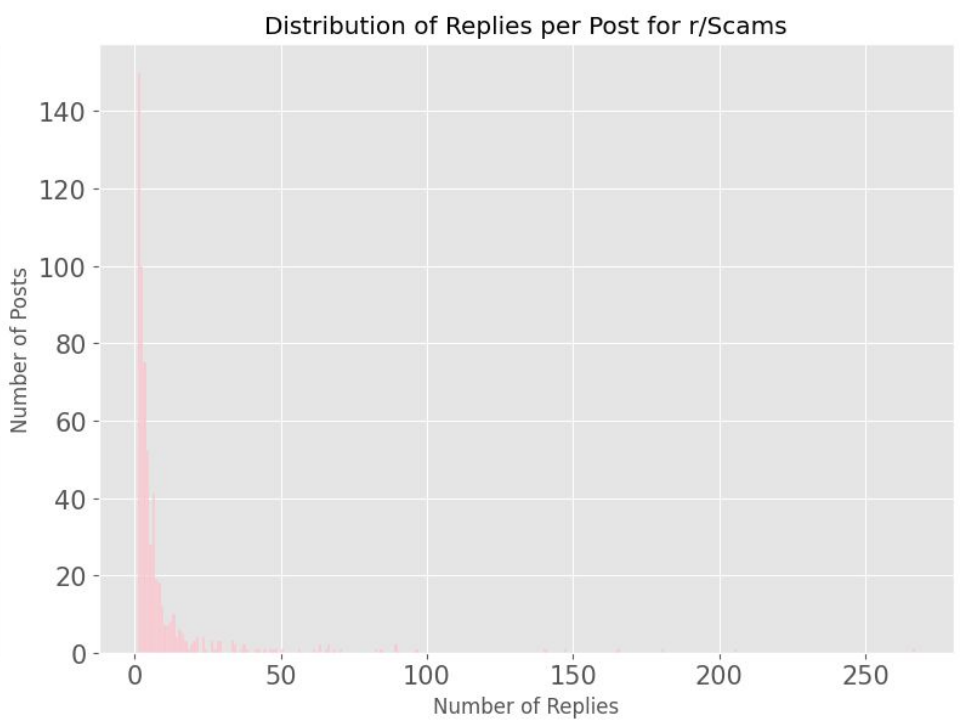
DATA FINDINGS

- EDA
  - Word vs. Character Count
  - Outliers
- Data Pre-processing


Distribution of Comments per Post for r/RandomKindness
Number of Posts
Number of Comments
0
20
40
60
80
100
120
140
160

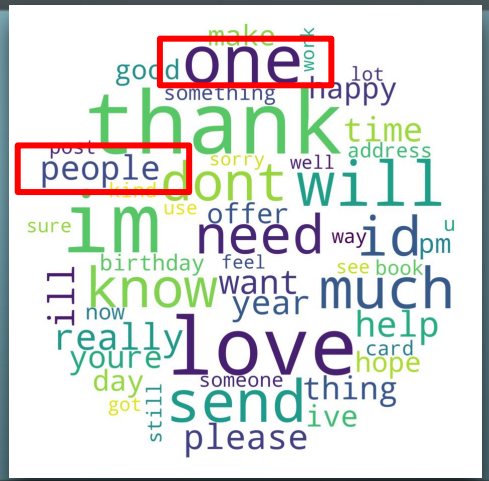

Distribution of Comments per Post for r/Scams
Number of Posts
Number of Comments
0
25
50
75
100
125
150
175
200
250


Distribution of Replies per Post for r/RandomKindness
Number of Posts
Number of Replies
0
20
40
60
80
100
120
140

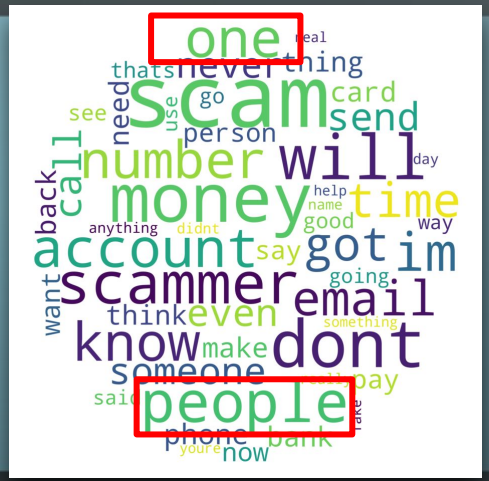

Distribution of Replies per Post for r/Scams
Number of Posts
Number of Replies
0
50
100
150
200
250
20
40
60
80
120
140

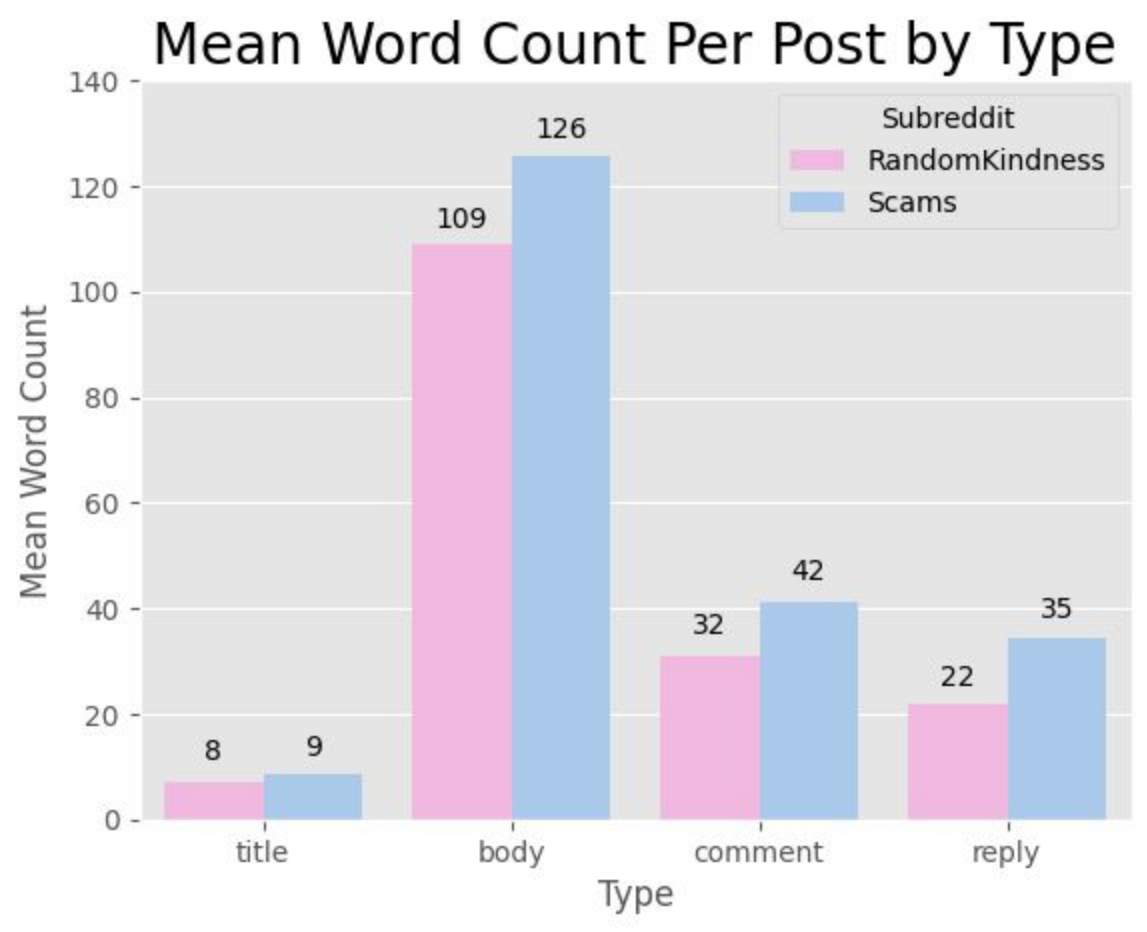
WORD COUNT
vs.
CHARACTER COUNT

r/RandomKindness

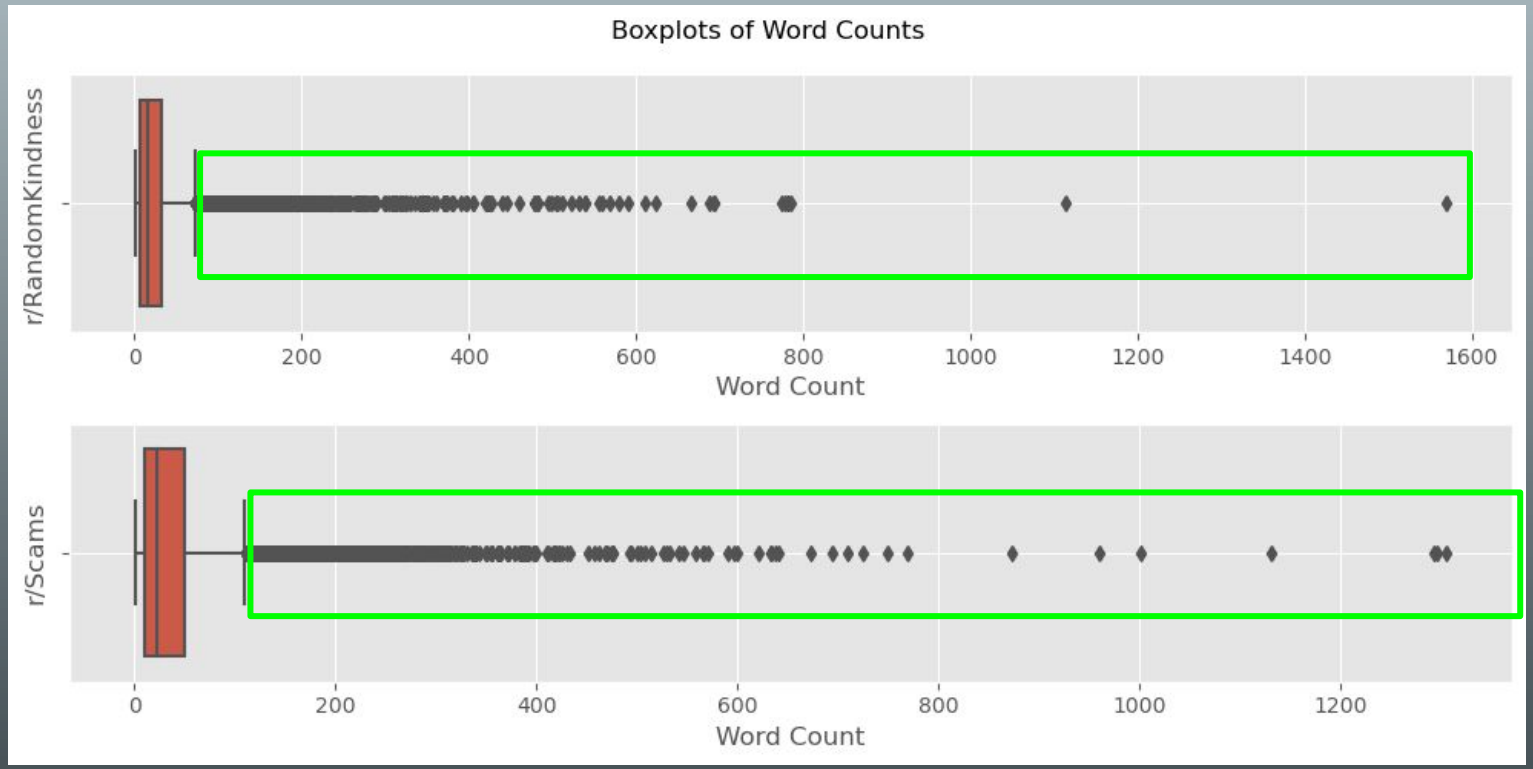


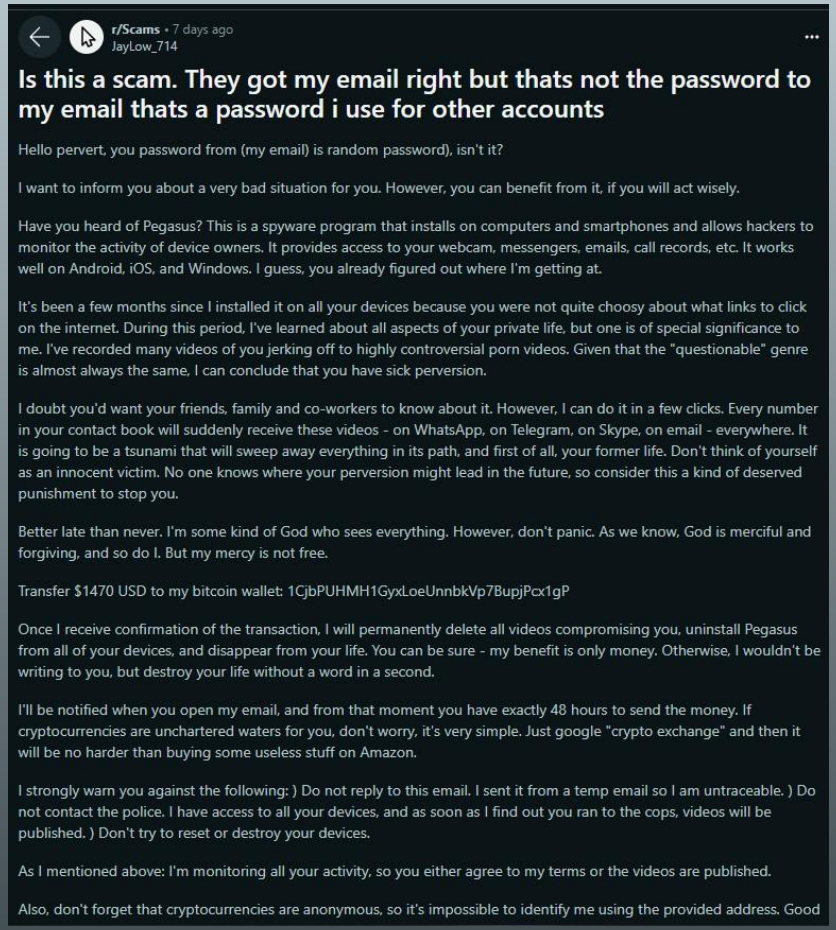
r/Scams




Mean Word Count Per Post by Type
Subreddit
RandomKindness
Scams
Mean Word Count
Type
title
body
comment
reply
8
9
109
126
32
42
22
35
0
20
40
60
80
100
120
140

OUTLIERS


Boxplots of Word Counts
r/RandomKindness
0
200
400
600
800
1000
1200
1400
1600
Word Count
r/Scams
0
200
400
600
800
1000
1200
Word Count



r/Scams • 7 days ago
JayLow_714

Is this a scam. They got my email right but thats not the password to my email thats a password i use for other accounts

Hello pervert, you password from (my email) is random password), isn't it?

I want to inform you about a very bad situation for you. However, you can benefit from it, if you will act wisely.

Have you heard of Pegasus? This is a spyware program that installs on computers and smartphones and allows hackers to monitor the activity of device owners. It provides access to your webcam, messengers, emails, call records, etc. It works well on Android, iOS, and Windows. I guess, you already figured out where I'm getting at.

It's been a few months since I installed it on all your devices because you were not quite choosy about what links to click on the internet. During this period, I've learned about all aspects of your private life, but one is of special significance to me. I've recorded many videos of you jerking off to highly controversial porn videos. Given that the "questionable" genre is almost always the same, I can conclude that you have sick perversion.

I doubt you'd want your friends, family and co-workers to know about it. However, I can do it in a few clicks. Every number in your contact book will suddenly receive these videos - on WhatsApp, on Telegram, on Skype, on email - everywhere. It is going to be a tsunami that will sweep away everything in its path, and first of all, your former life. Don't think of yourself as an innocent victim. No one knows where your perversion might lead in the future, so consider this a kind of deserved punishment to stop you.

Better late than never. I'm some kind of God who sees everything. However, don't panic. As we know, God is merciful and forgiving, and so do I. But my mercy is not free.

Transfer $1470 USD to my bitcoin wallet: 1CjbPUHMH1GyxLoeUnnbkVp7BupjPcx1gP

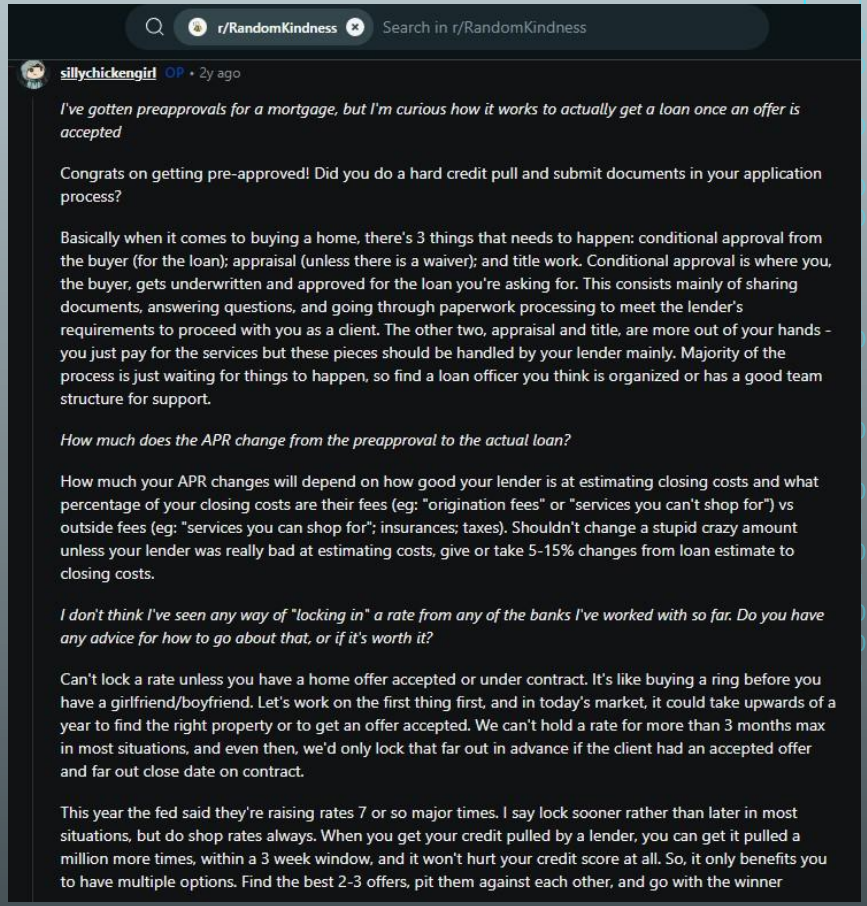
Once I receive confirmation of the transaction, I will permanently delete all videos compromising you, uninstall Pegasus from all of your devices, and disappear from your life. You can be sure - my benefit is only money. Otherwise, I wouldn't be writing to you, but destroy your life without a word in a second.

I'll be notified when you open my email, and from that moment you have exactly 48 hours to send the money. If cryptocurrencies are unchartered waters for you, don't worry, it's very simple. Just google "crypto exchange" and then it will be no harder than buying some useless stuff on Amazon.

I strongly warn you against the following:) Do not reply to this email. I sent it from a temp email so I am untraceable.) Do not contact the police. I have access to all your devices, and as soon as I find out you ran to the cops, videos will be published.) Don't try to reset or destroy your devices.

As I mentioned above: I'm monitoring all your activity, so you either agree to my terms or the videos are published.

Also, don't forget that cryptocurrencies are anonymous, so it's impossible to identify me using the provided address. Good



r/RandomKindness Search in r/RandomKindness

sillychickengirl OP • 2y ago

I've gotten preapprovals for a mortgage, but I'm curious how it works to actually get a loan once an offer is accepted

Congrats on getting pre-approved! Did you do a hard credit pull and submit documents in your application process?

Basically when it comes to buying a home, there's 3 things that needs to happen: conditional approval from the buyer (for the loan); appraisal (unless there is a waiver); and title work. Conditional approval is where you, the buyer, gets underwritten and approved for the loan you're asking for. This consists mainly of sharing documents, answering questions, and going through paperwork processing to meet the lender's requirements to proceed with you as a client. The other two, appraisal and title, are more out of your hands - you just pay for the services but these pieces should be handled by your lender mainly. Majority of the process is just waiting for things to happen, so find a loan officer you think is organized or has a good team structure for support.

How much does the APR change from the preapproval to the actual loan?

How much your APR changes will depend on how good your lender is at estimating closing costs and what percentage of your closing costs are their fees (eg: "origination fees" or "services you can't shop for") vs outside fees (eg: "services you can shop for"; insurances; taxes). Shouldn't change a stupid crazy amount unless your lender was really bad at estimating costs, give or take 5-15% changes from loan estimate to closing costs.

I don't think I've seen any way of "locking in" a rate from any of the banks I've worked with so far. Do you have any advice for how to go about that, or if it's worth it?

Can't lock a rate unless you have a home offer accepted or under contract. It's like buying a ring before you have a girlfriend/boyfriend. Let's work on the first thing first, and in today's market, it could take upwards of a year to find the right property or to get an offer accepted. We can't hold a rate for more than 3 months max in most situations, and even then, we'd only lock that far out in advance if the client had an accepted offer and far out close date on contract.

This year the fed said they're raising rates 7 or so major times. I say lock sooner rather than later in most situations, but do shop rates always. When you get your credit pulled by a lender, you can get it pulled a million more times, within a 3 week window, and it won't hurt your credit score at all. So, it only benefits you to have multiple options. Find the best 2-3 offers, pit them against each other, and go with the winner

PRE-PROCESSING

Top 10 N-gram r/RandomKindness

1-gram	2-gram	3-gram
to	thank you	thank you for
you	if you	thank you so
and	in the	you so much
the	love to	would love to
for	would love	id love to
of	so much	me your address
my	this is	let me know
it	for the	love to send
in	to be	so much for
is	you for	be able to
so	you are	if you have

1-gram	2-gram	3-gram
love	happy birthday	amazon wish list
thank	sorry loss	happy birthday hope
ill	feel free	thank kind offer
please	last year	message moderator mail
happy	year old	please feel free
thanks	years ago	feel free pm
hope	thank kind	please read rules
offer	please pm	happy birthday spiderthom
birthday	please let	thank thank thank
pm	wish list	please message moderator
kind	please dm	feel free message

1-gram	2-gram	3-gram
love	happy birthday	amazon wish list
thank	sorry loss	happy birthday hope
ill	feel free	thank kind offer
please	last year	message moderator mail
happy	year old	please feel free
thanks	years ago	feel free pm
hope	thank kind	please read rules
offer	please pm	happy birthday spiderthom
birthday	please let	thank thank thank
pm	wish list	please message moderator
kind	please dm	feel free message

Top 10 N-gram r/Scams

1-gram	2-gram	3-gram
the	in the	there is no
to	of the	you need to
and	if you	be able to
you	to be	this is the
it	this is	block and ignore
is	on the	this is scam
of	to the	is this scam
that	it was	you have to
they	to get	thank you for
for	they are	and move on
in	you can	good to be

1-gram	2-gram	3-gram
scam	phone number	pig butchering scam
money	gift cards	watch recovery scammers
account	block ignore	buy gift cards
number	social media	name phone number
email	gon na	file police report
scammers	bank account	reverse image search
phone	red flag	scam block ignore
call	years ago	social security number
pay	task scam	pig butchering scams
bank	friends family	long story short
scammer	red flags	stop sending nudes

1-gram	2-gram	3-gram
scam	phone number	pig butchering scam
money	gift cards	watch recovery scammers
account	block ignore	buy gift cards
number	social media	name phone number
email	gon na	file police report
scammers	bank account	reverse image search
phone	red flag	scam block ignore
call	years ago	social security number
pay	task scam	pig butchering scams
bank	friends family	long story short
scammer	red flags	stop sending nudes

1-gram	2-gram	3-gram
scam	phone number	pig butchering scam
money	gift cards	watch recovery scammers
account	block ignore	buy gift cards
number	social media	name phone number
email	gon na	file police report
scammers	bank account	reverse image search
phone	red flag	scam block ignore
call	years ago	social security number
pay	task scam	pig butchering scams
bank	friends family	long story short
scammer	red flags	stop sending nudes

- Scam-related actions: *pay, block ignore*
- Payment method: *money, gift cards, bank account*
- Contact information: *email, phone number, social media*

04

DATA MODELING

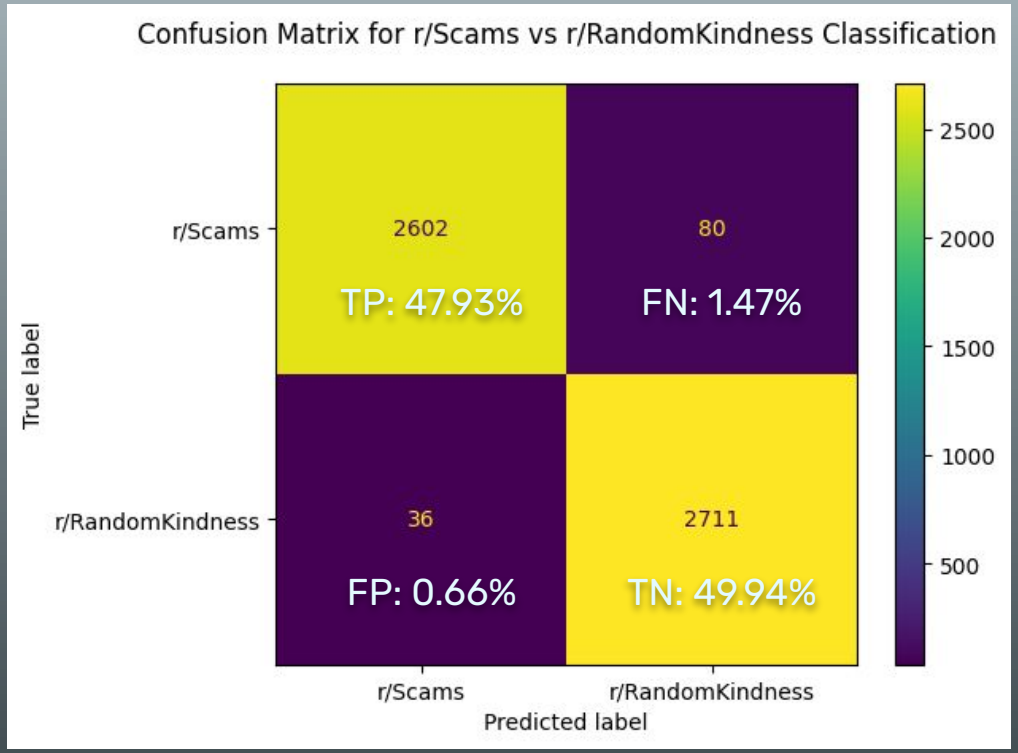
MODEL COMPARISON

Rank	Model	Vectorizer	CV Score
1	Multinomial Naive Bayes	TF-IDF	0.972553
2	Multinomial Naive Bayes	CountVectorizer	0.972037
3	Logistic Regression	TF-IDF	0.956158
4	Logistic Regression	CountVectorizer	0.958369

CONFUSION MATRIX
for Naive Bayes with TF-IDF Vectorizer

Accuracy Score = 0.975

Our model correctly classified 97.5% of posts into its respective subreddit.

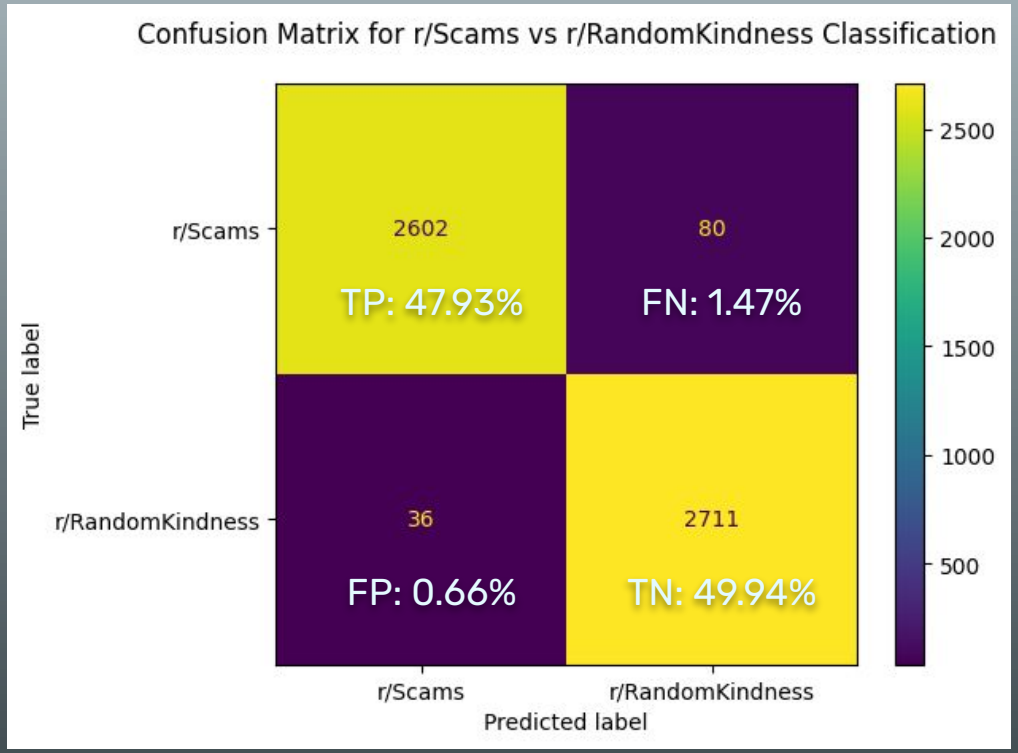


TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative

CONFUSION MATRIX
for Naive Bayes with TF-IDF Vectorizer

Sensitivity Score = 0.964

Among all the scam posts, our model correctly classified 96.4% of them as scams.



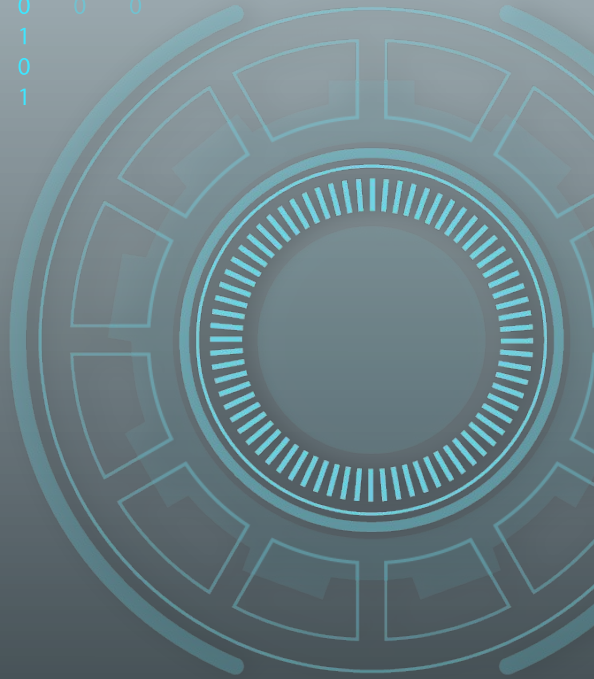
TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative

IMPACTS

- Wrongfully flagging a scam post as a random act of kindness can **reduce credibility of the subreddit** (i.e. misinformation → mislead people)
- Hence, accurate detection of a scam post, can…
  - **Reduce users' exposure** (and occurrence) to scams
  - Create **safer and more positive online community** by identifying scams while promoting acts of kindness
  - Boost **moderator's effectiveness in gatekeeping** the community from occurrence of potential scams

05

CONCLUSION



CONCLUSION

Metric(s)	Score (%)	Description
Sensitivity	96.4	Detects scam posts 96.4 out of 100 text-based entries as Scams
Accuracy	97.5	Classifies 97.5 out of 100 entries as Scams **AND** Random acts of Kindness

06
LIMITATIONS & RECOMMENDATIONS

LIMITATIONS

- Model to be further trained on other types of text-based data
  - Other forums (e.g. HardwareZone, askSingapore)
  - Social Media
  - Instant Messaging
- Vernacular terms in Singapore to be considered

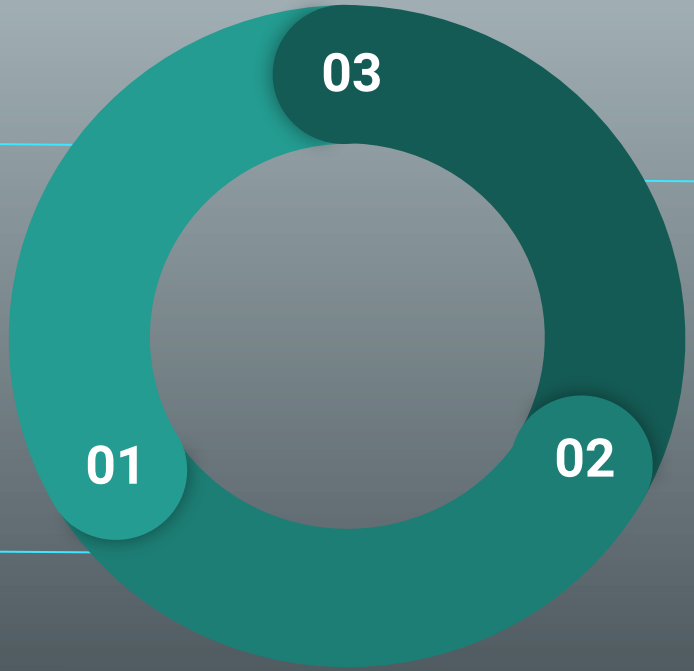
RECOMMENDATIONS

Text-based detection tool

- Note: App still in beta v1.1 stage and is complementary
- Cuts down the manual reviewing

Transparency Notice

- Pre-empt subreddit users about the use of a NLP model in subreddit



Feedback Loop

- Focus Groups with users + moderators
- Constant monthly review to improve modeling

‘SCAM OR NOT’ DETECTION TEST

SAMPLE #1

u/KardGuru · Promoted

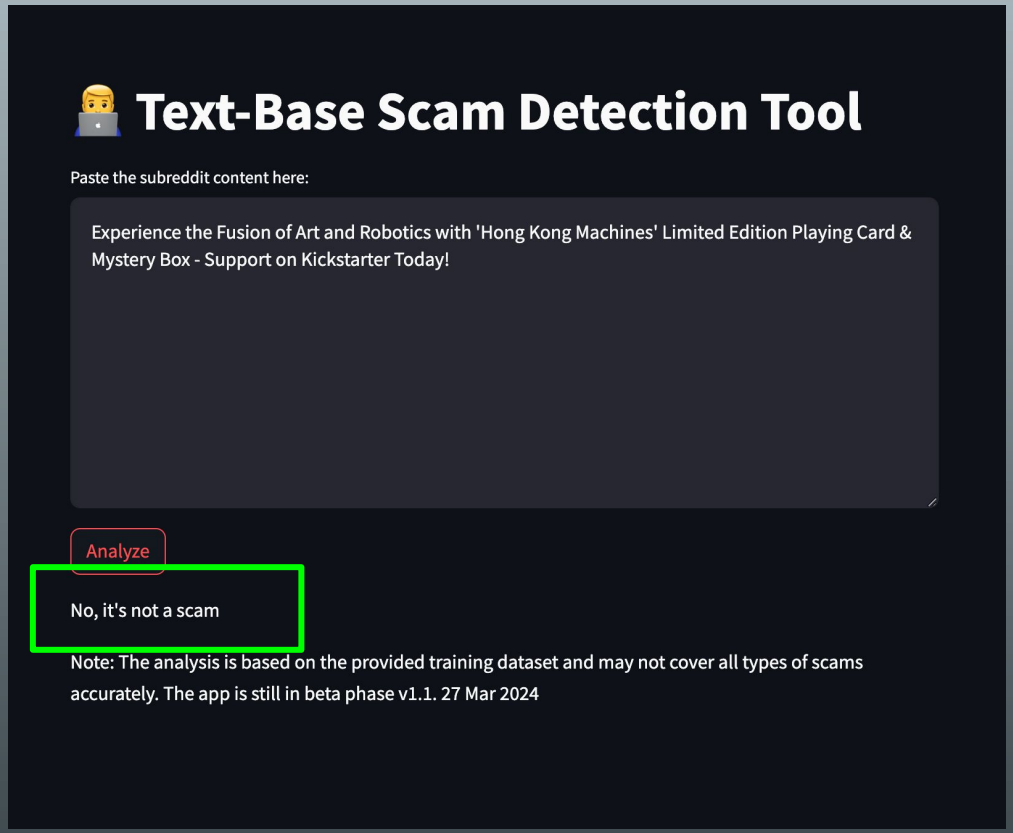
Experience the Fusion of Art and Robotics with 'Hong Kong Machines' Limited Edition Playing Card & Mystery Box - Support on Kickstarter Today!

Tiny Hong Kong Machines Playing Card
qpmarketnetwork.com

Learn More

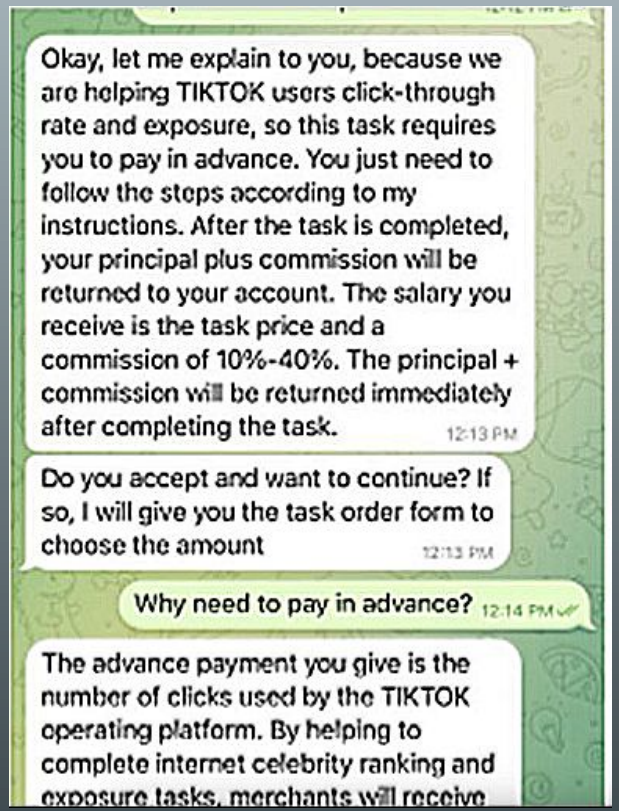
Sort by: Best

SAMPLE #1

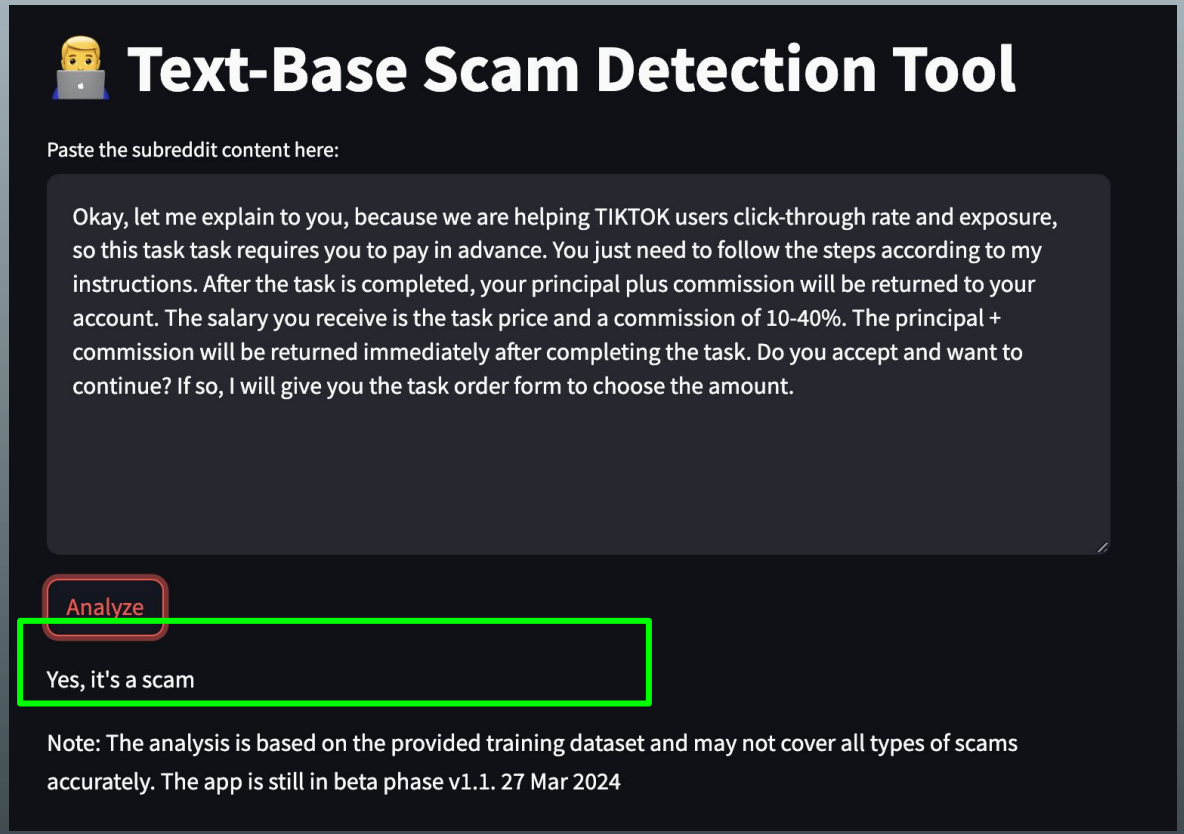


Sources: r/KardGuru

SAMPLE #2

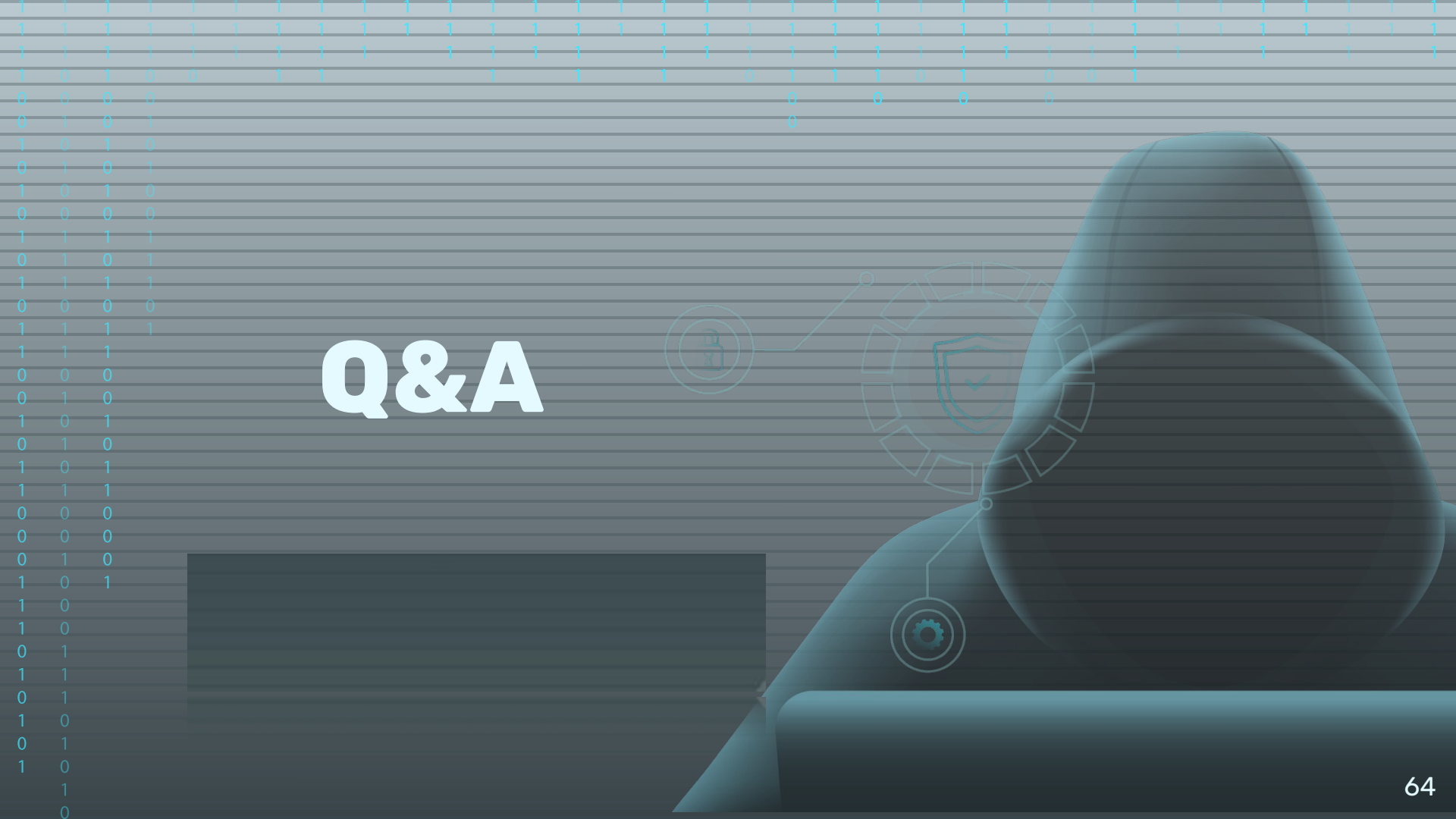

Okay, let me explain to you, because we are helping TIKTOK users click-through rate and exposure, so this task requires you to pay in advance. You just need to follow the steps according to my instructions. After the task is completed, your principal plus commission will be returned to your account. The salary you receive is the task price and a commission of 10%-40%. The principal + commission will be returned immediately after completing the task. 12:13 PM
Do you accept and want to continue? If so, I will give you the task order form to choose the amount 12:13 PM
Why need to pay in advance? 12:14 PM
The advance payment you give is the number of clicks used by the TIKTOK operating platform. By helping to complete internet celebrity ranking and exposure tasks, merchants will receive

SAMPLE #2



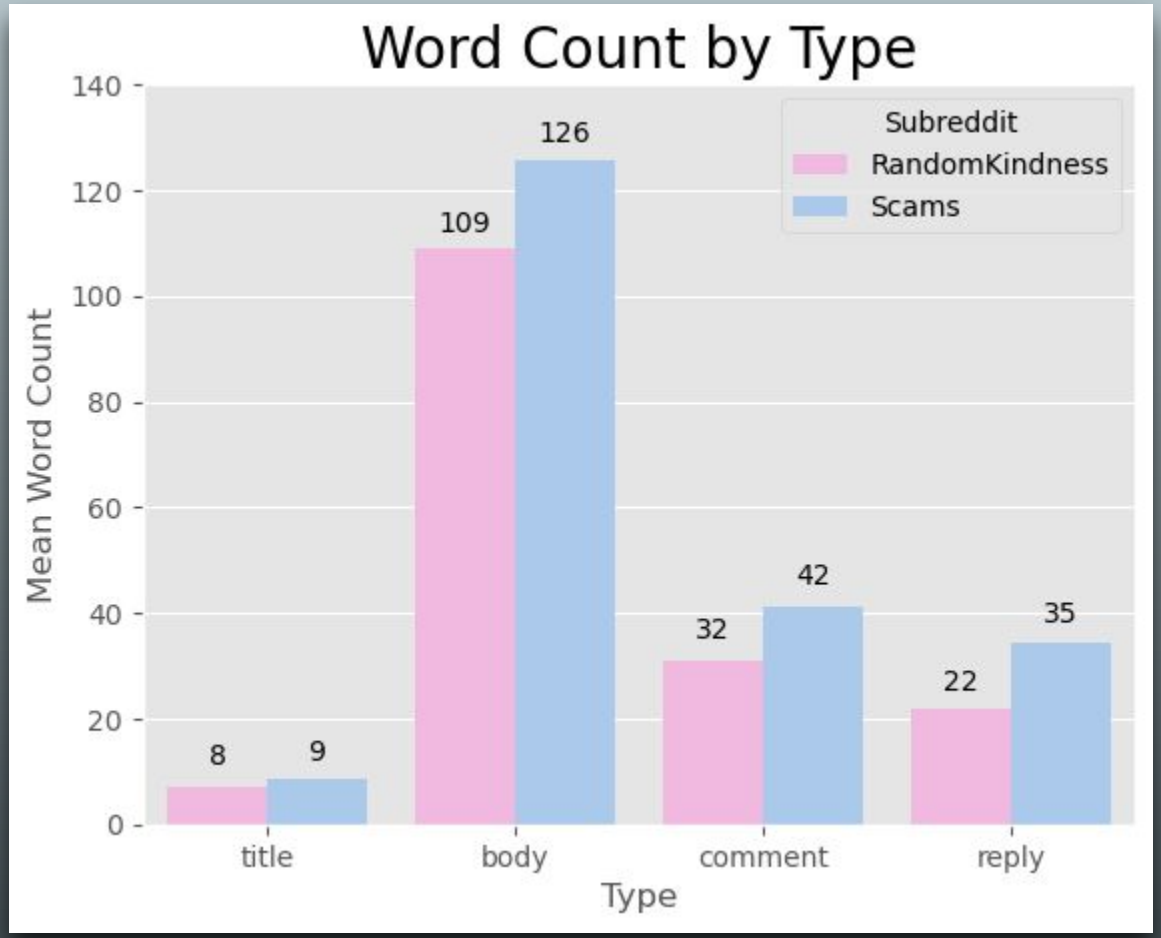
Sources: Singapore Police Force, 2024

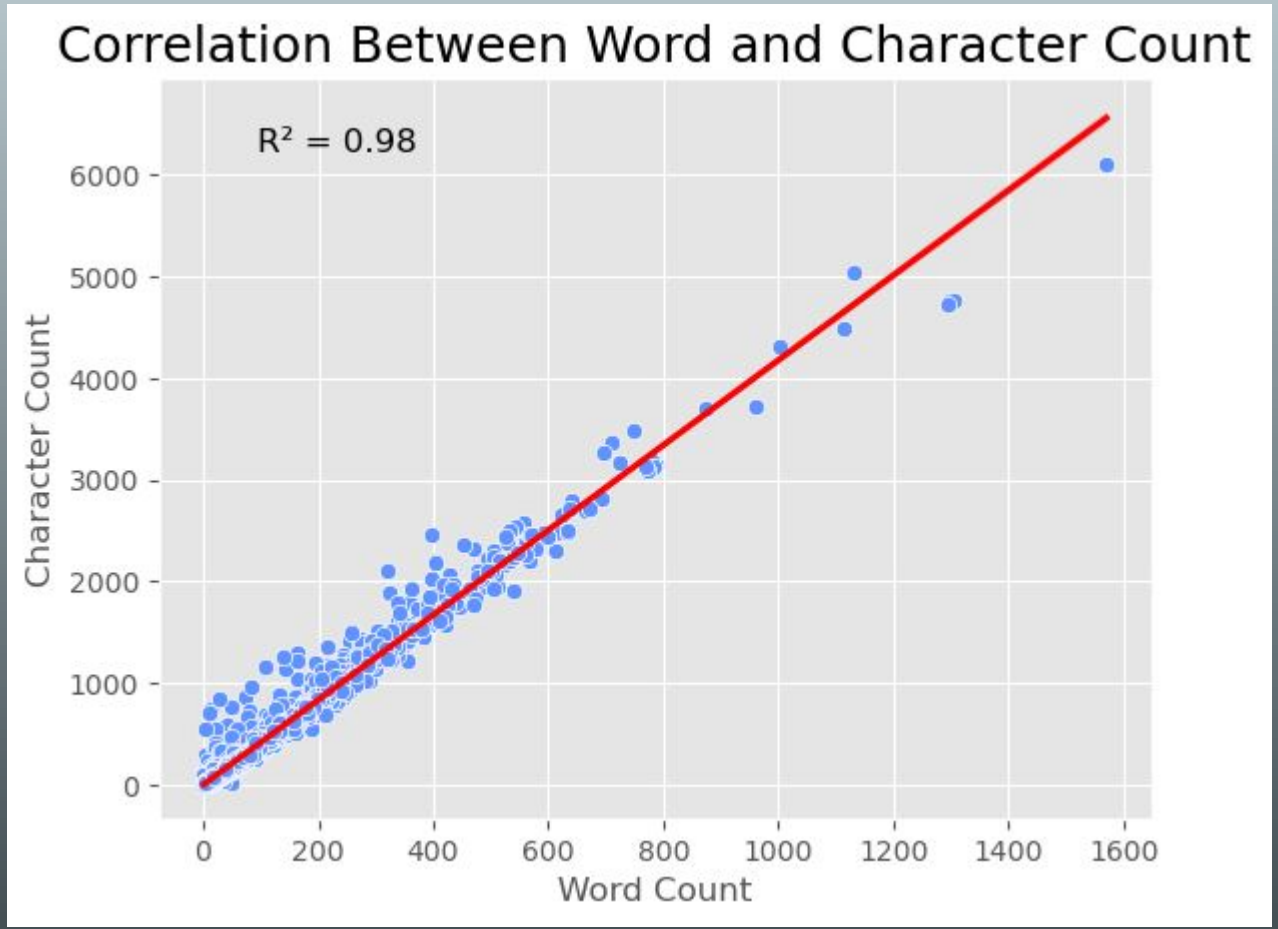
Q&A

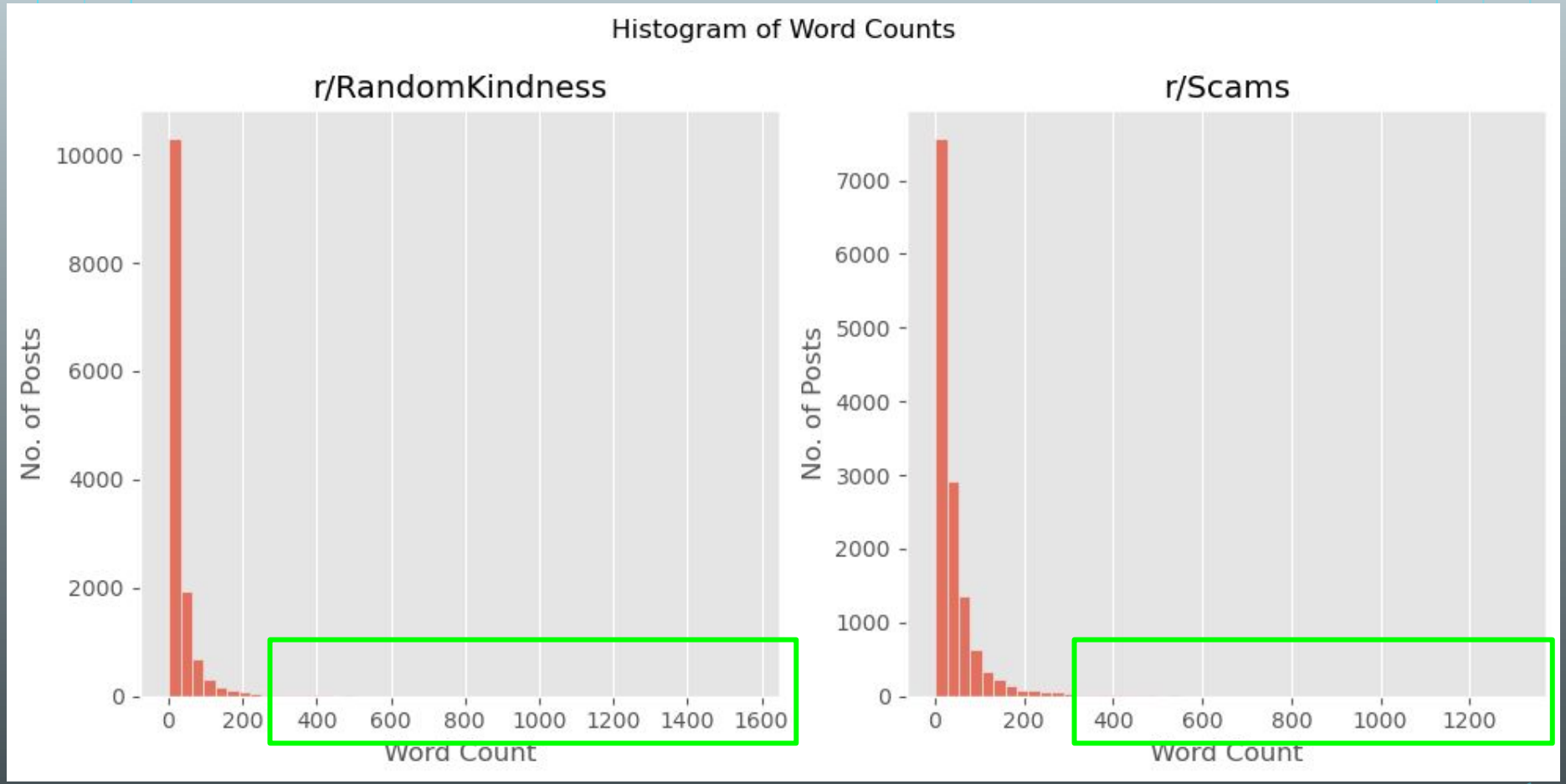


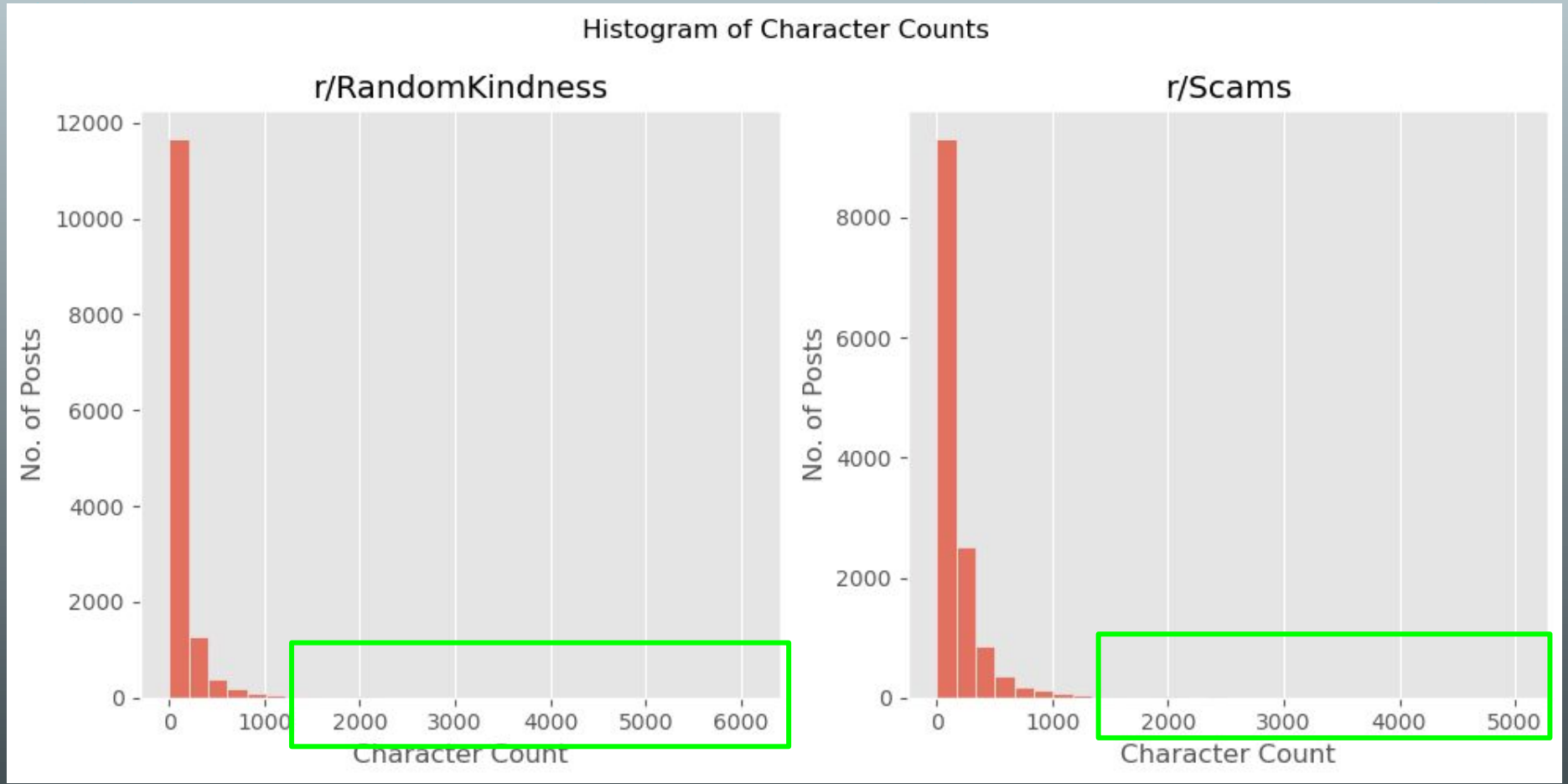
07

APPENDIX


Word Count by Type
Subreddit
RandomKindness
Scams
Mean Word Count
140
120
100
80
60
40
20
0
8
9
109
126
32
42
22
35
title
body
comment
reply
Type

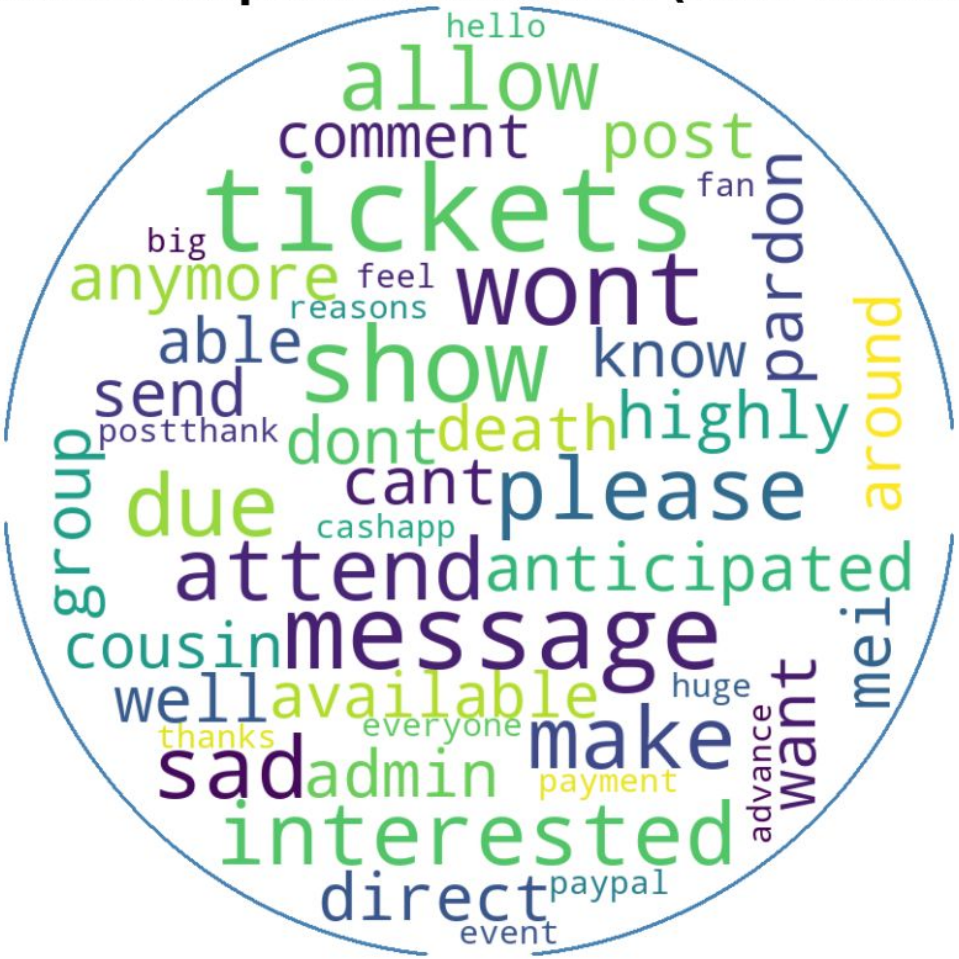

Correlation Between Word and Character Count
$R^2 = 0.98$
Character Count
Word Count
0
1000
2000
3000
4000
5000
6000
0
200
400
600
800
1000
1200
1400
1600


Histogram of Word Counts
r/RandomKindness
No. of Posts
10000
8000
6000
4000
2000
0
0
200
400
600
800
1000
1200
1400
1600
Word Count
r/Scams
No. of Posts
7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
1000
1200
Word Count

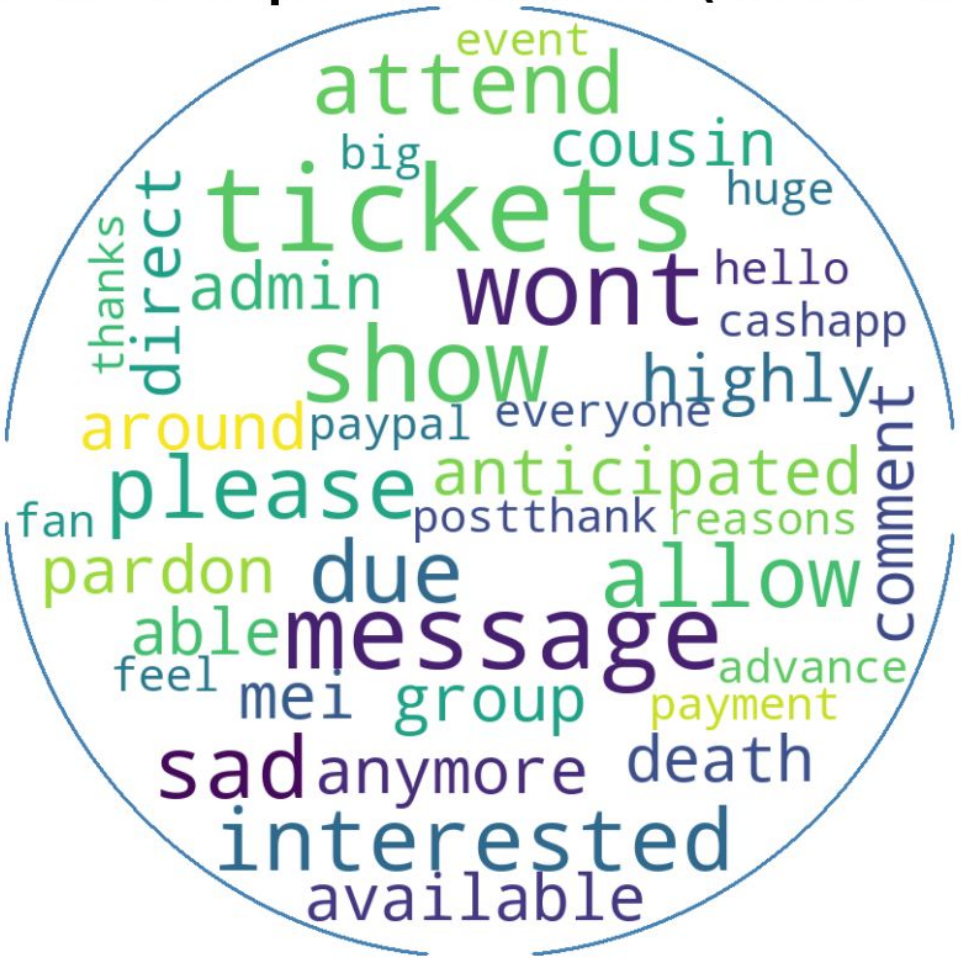

Histogram of Character Counts
r/RandomKindness
No. of Posts
Character Count
0
2000
4000
6000
8000
10000
12000
0
1000
2000
3000
4000
5000
6000
r/Scams
No. of Posts
Character Count
0
2000
4000
6000
8000
0
1000
2000
3000
4000
5000

Common words r/Scams & r/RandomKindness - Round 3 N-Gram

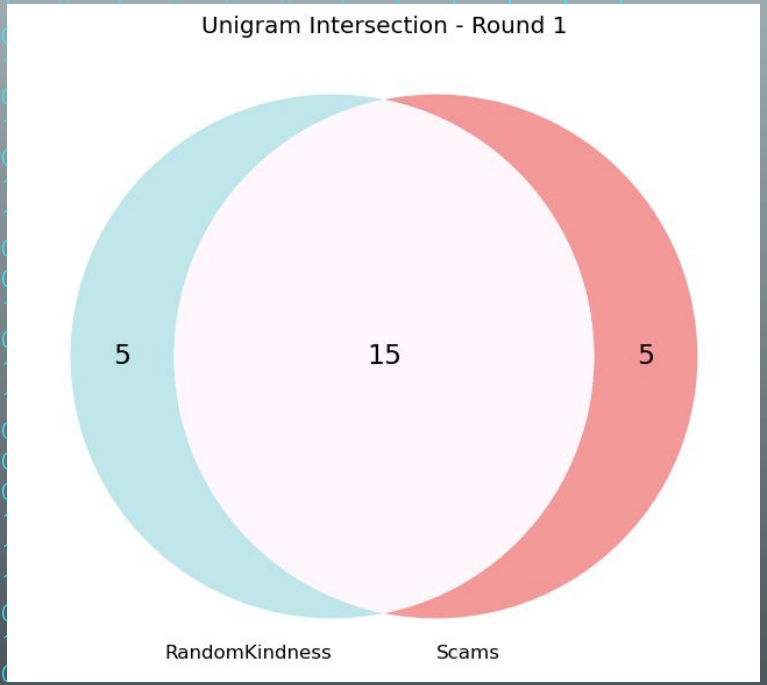
Standard Stopwords Removed (after Round 1)

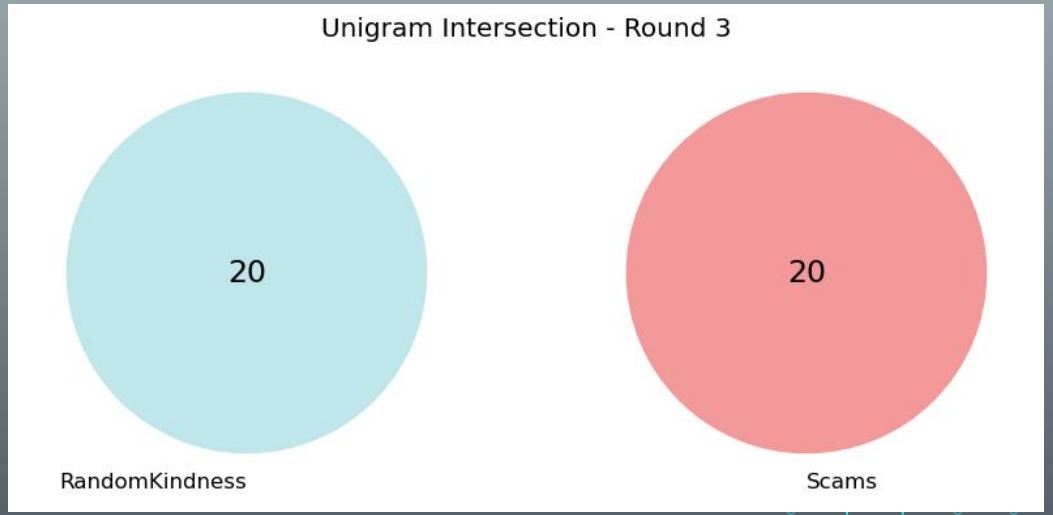


Customised Stopwords Removed (after Round 2)

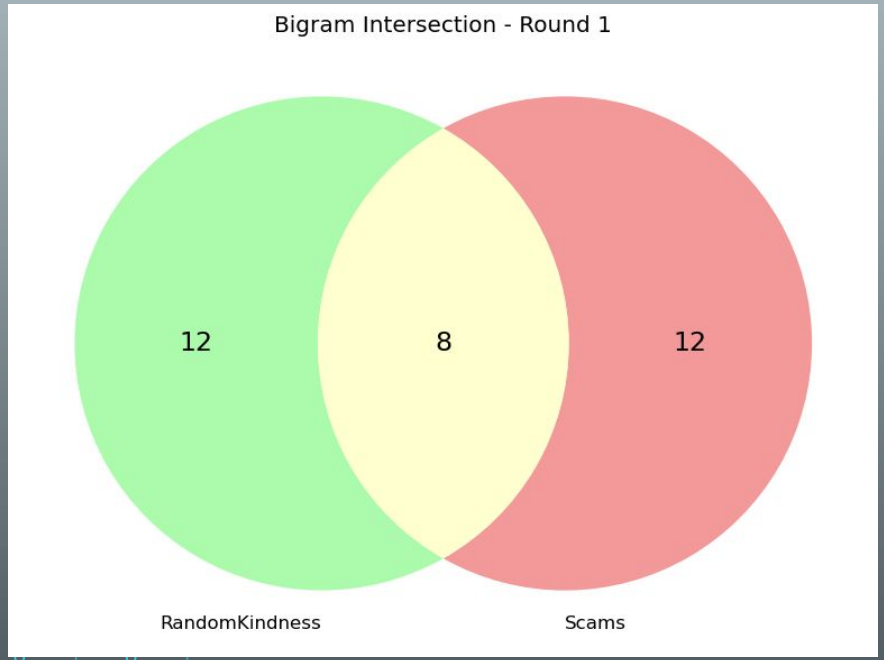


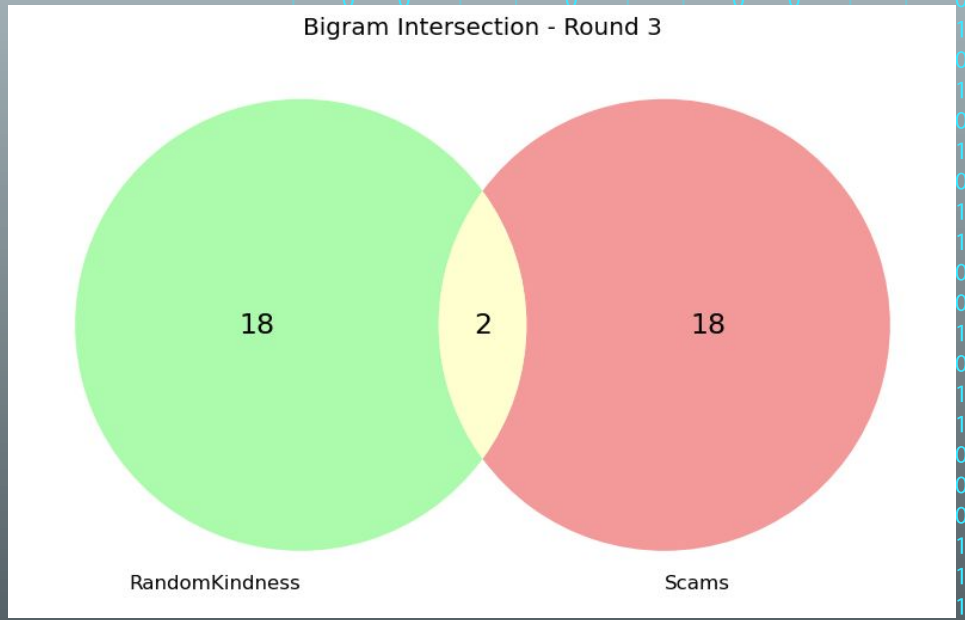
1-GRAM ANALYSIS



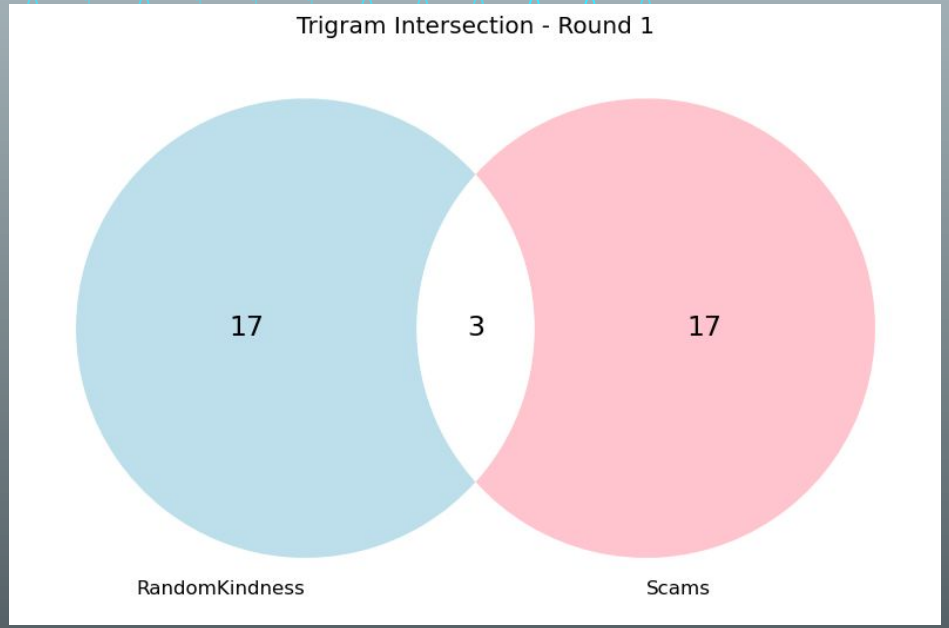


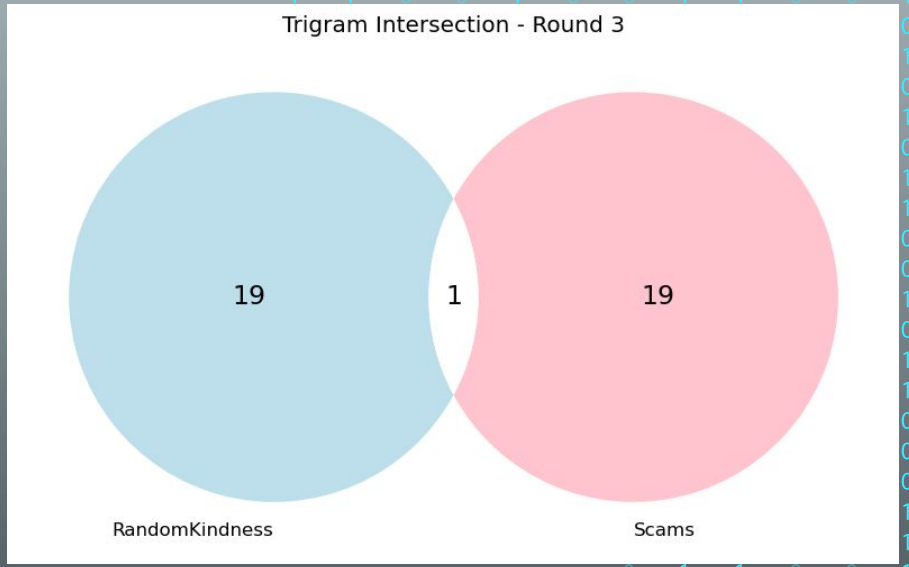
2-GRAM ANALYSIS



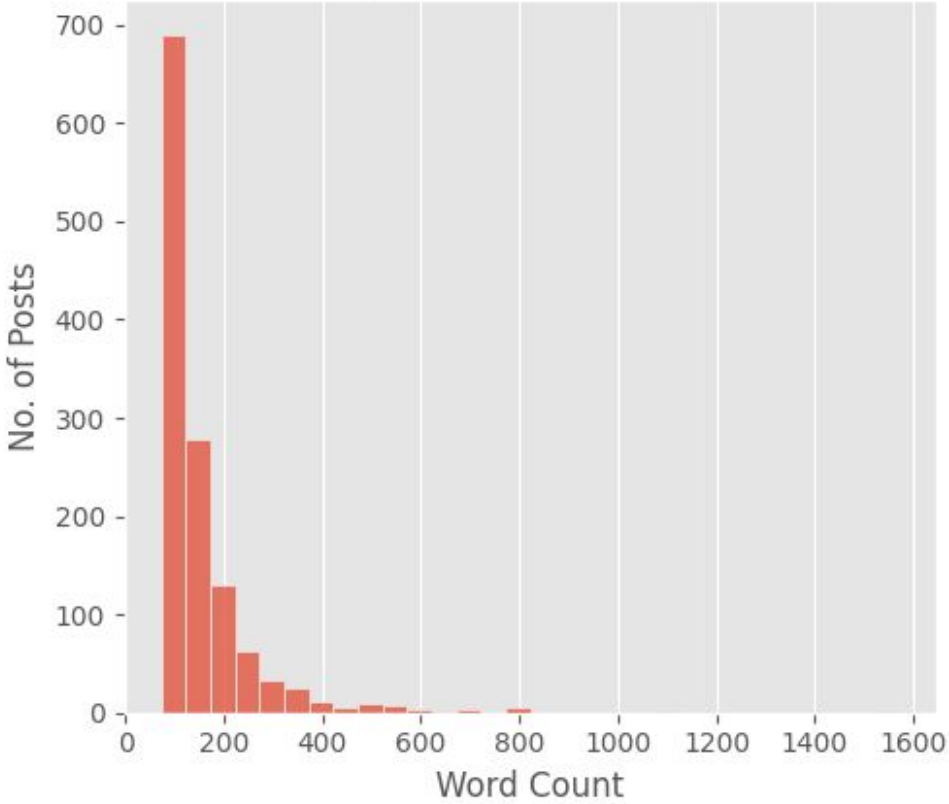


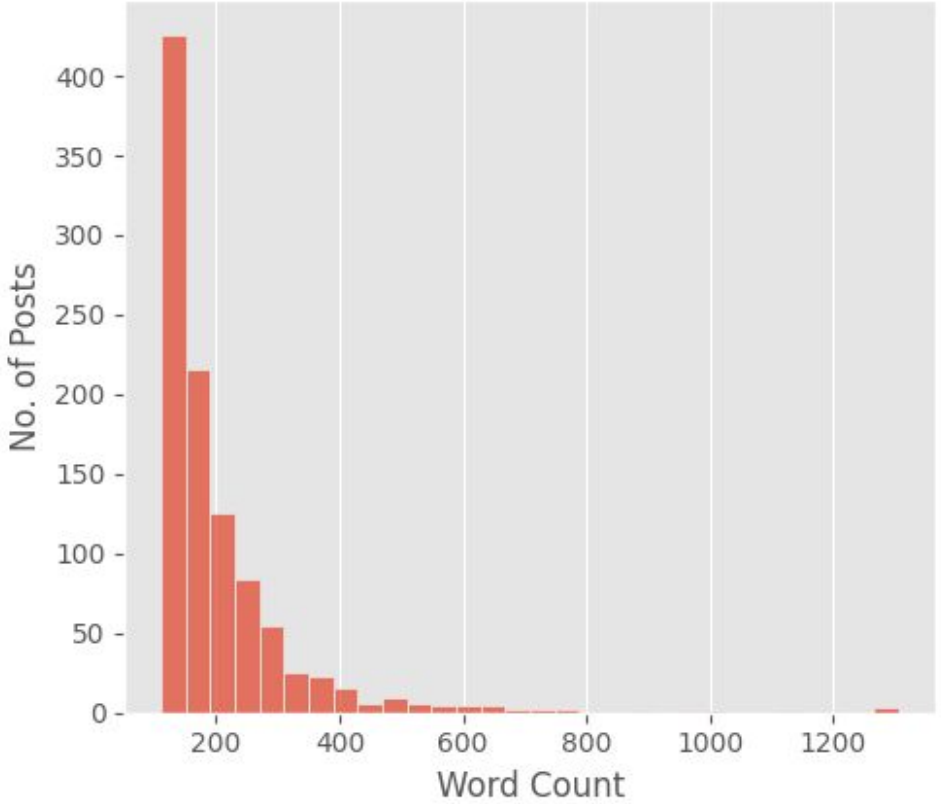
3-GRAM ANALYSIS





Histogram of Outlier Word Counts





Histogram of Outlier Word Counts