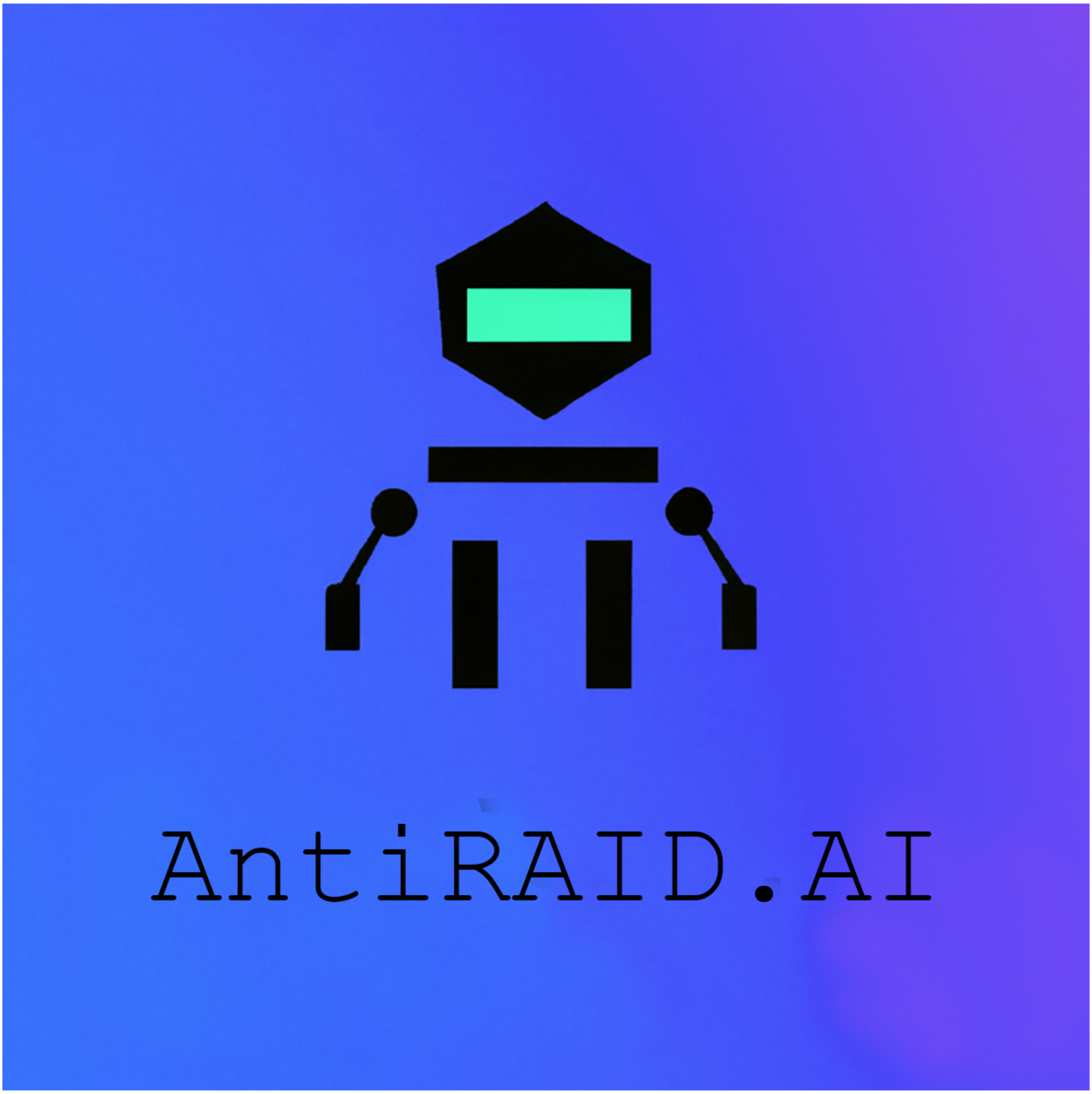
AntiRAID.AI White Paper

Introducing a new era of artificially intelligent community moderators.



twitter.com/AntiRAIDAI | t.me/AntiRAIDAI

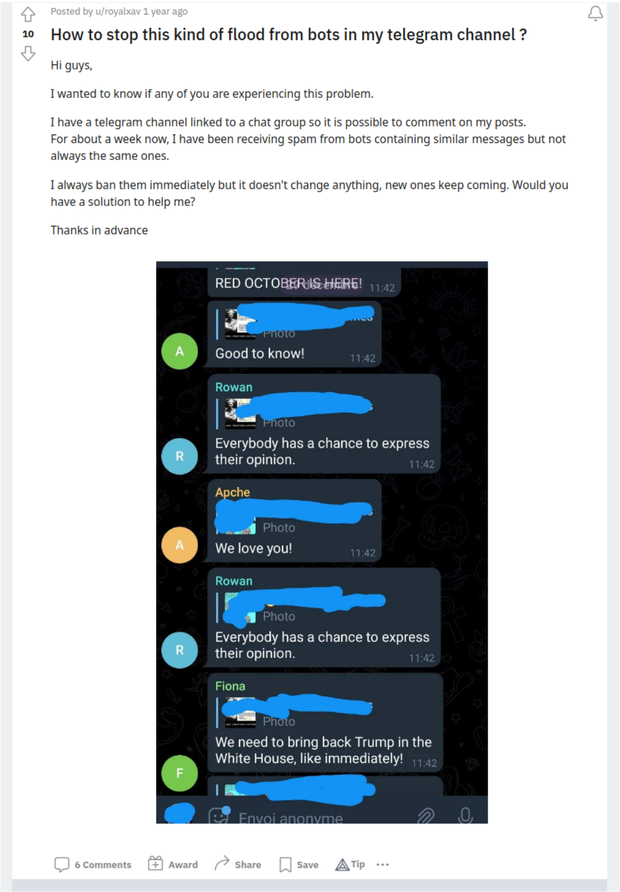
CAPTCHAS and security on Telegram community groups:

CAPTCHAS, the bane of existence for most people who spend a long time behind a computer. Most Telegram and Discord community groups nowadays rely on CAPTCHA puzzles to authenticate new users to ensure they aren’t spam bots looking to derail the chat.

In the crypto Telegram community, spam, raid and FUD bot attacks are particularly common, as bad actors look to blackmail devs by derailing their groups and demanding payment as a condition to stop. The formula is simple, deploy hundreds of bots to join your group and spam a message like “DM Dr JEET to stop the flood” and of course, when you message him, he sends you a BSC or ETH wallet to deposit funds into. Most crypto groups rely on ‘portals’ to protect themselves, channels which require the user to complete a short ‘CAPTCHA’ asking the user to correctly identify the contents of certain pictures before being allowed to join, to prove they are humans.

There has been much buzz in the news recently about artificial intelligence following OpenAI’s release of their public API allowing anybody to utilise the power of AI to solve any task the user can give it. This presents a serious problem for Telegram groups relying on CAPTCHAS to protect themselves from spam/raid attacks. The power of AI has become so great that asking it to identify which of a set of images contains a sidewalk or a bird in a cage is no longer a difficult task for a machine.

Therefore botters who have access to this technology could quite conceivably build a network of bots programmed to correctly solve CAPTCHA puzzles and invade your group regardless of any portals designed to protect it. Admins often use bots like Rose as a further means to protect the channel, by blacklisting certain words or phrases that the spammers or raiders may use. The issue with relying on this is that the spammer can regularly reword his sentences, or simply use synonyms of the banned keywords the bot might be looking for.



Posted by u/royalxav 1 year ago

10

How to stop this kind of flood from bots in my telegram channel ?

Hi guys,

I wanted to know if any of you are experiencing this problem.

I have a telegram channel linked to a chat group so it is possible to comment on my posts.
For about a week now, I have been receiving spam from bots containing similar messages but not always the same ones.

I always ban them immediately but it doesn't change anything, new ones keep coming. Would you have a solution to help me?

Thanks in advance

RED OCTOBER IS HERE! 11:42

[redacted] Photo
Good to know! 11:42

Rowan
[redacted] Photo
Everybody has a chance to express their opinion. 11:42

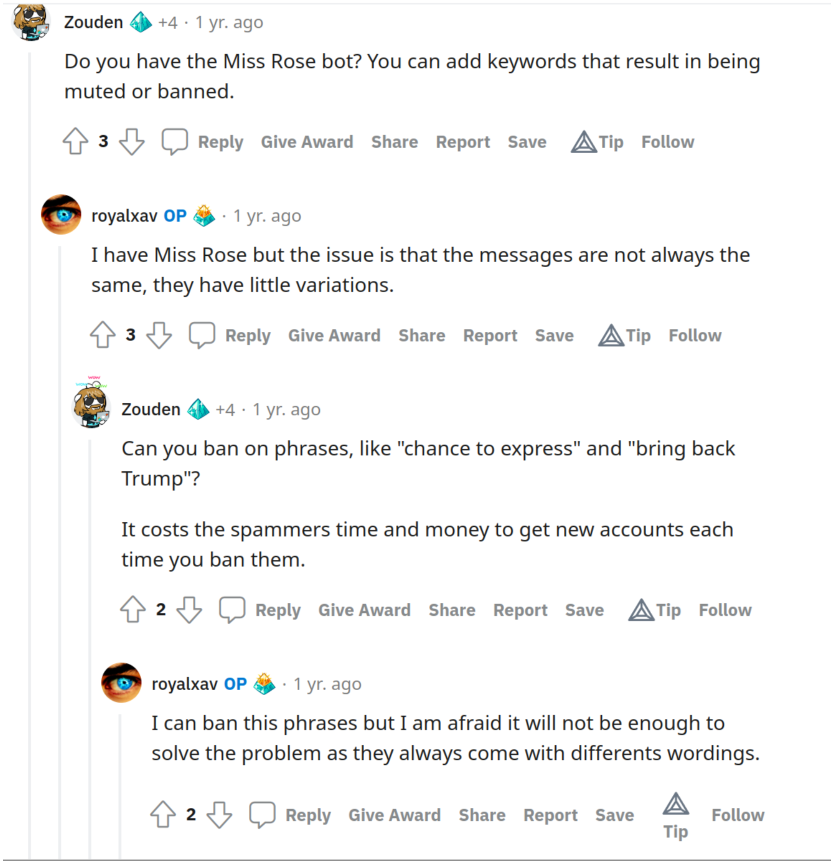
Apche
[redacted] Photo
We love you! 11:42

Rowan
[redacted] Photo
Everybody has a chance to express their opinion. 11:42

Fiona
[redacted] Photo
We need to bring back Trump in the White House, like immediately! 11:42

[redacted]

[redacted] Envoi anonyme



Zouden +4 · 1 yr. ago

Do you have the Miss Rose bot? You can add keywords that result in being muted or banned.

3 Reply Give Award Share Report Save Tip Follow

royalxav OP · 1 yr. ago

I have Miss Rose but the issue is that the messages are not always the same, they have little variations.

3 Reply Give Award Share Report Save Tip Follow

Zouden +4 · 1 yr. ago

Can you ban on phrases, like "chance to express" and "bring back Trump"?

It costs the spammers time and money to get new accounts each time you ban them.

2 Reply Give Award Share Report Save Tip Follow

royalxav OP · 1 yr. ago

I can ban this phrases but I am afraid it will not be enough to solve the problem as they always come with differents wordings.

2 Reply Give Award Share Report Save Tip Follow

(A Reddit user – venting his frustrations at spam bots raiding his channel, and complaining about the ineffectiveness of using basic word filters such as Miss Rose to stop them)

AntiRAID.AI stepping in with a solution:

AntiRAID.AI aims to solve this problem by ‘fighting fire with fire’ to some extent and using the power of AI to fight back against bad actors entering your group who may be using AI to bypass captcha requirements and join with bots en-masse.

AntiRAID.AI is a bot on Telegram which can be added to any community group or channel and used to protect it from raid or spam attacks by any bad actors, from rival projects looking to FUD (crypto slang for baseless fear mongering) or spread rumours, to blackmailers looking to spam the group with annoying messages demanding the admin message them to stop the raid. Blackmailers may also attempt to use FUD or negative sentiment as a means of scaring the dev into paying up, as Telegram groups are a particular asset to cryptocurrency projects launched on decentralized exchanges, as this is where most of their exposure happens.

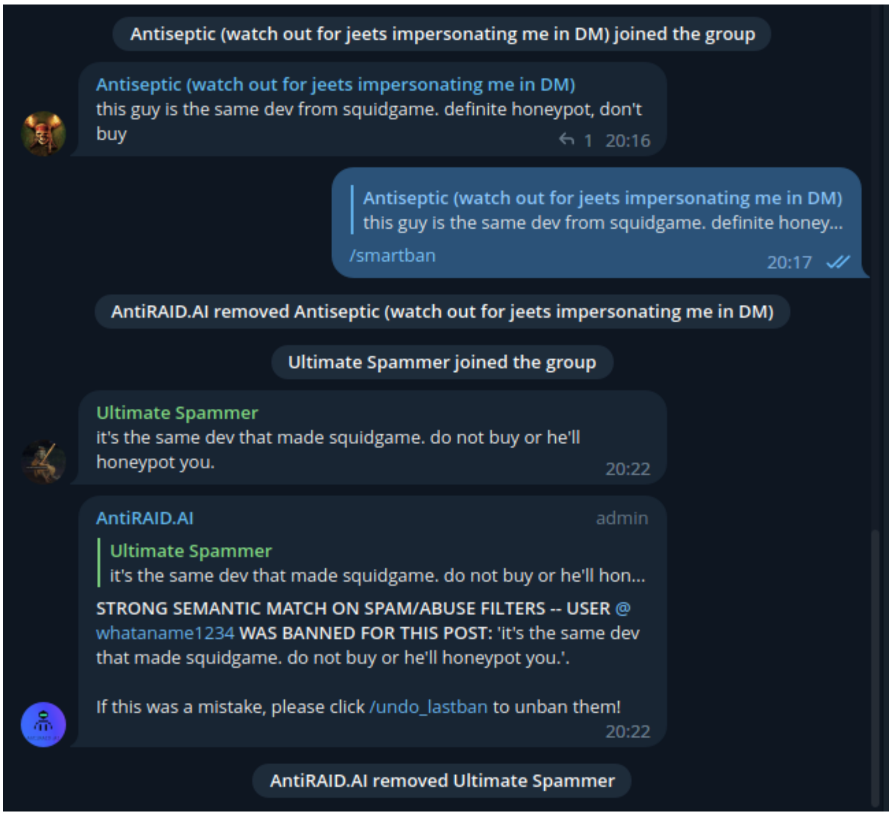
AntiRAID.AI bot uses the OpenAI API to access GPT-3 natural language models. These are OpenAI’s most advanced engines to date and have the ability to generate text on a particular topic or answer questions in completely fluent English and in great detail, such to the extent that one wonders whether a computer or a thinking, conscious being wrote it.

OpenAI also has the ability to analyze the semantic content of text, i.e. understand and interpret its given meaning. And this is where AntiRAID.AI comes in. As I mentioned earlier, admins of groups will sometimes use bots like Rose to blacklist certain words the spam bots use, so the spammers simply alter their sentences slightly. AntiRAID.AI takes this style of auto-moderation to the next level: when an admin sees a spammer trying to derail the chat, he can report him to AntiRAID.AI, who will automatically ban him, and add his messages to a blocklist of filtered phrases. Once a given message has been added to the blocklist, its content is sent to the OpenAI API and we request a vector value representing its given meaning. This returns us a vector made up of thousands of decimal numbers representing that word or phrase’s actual meaning.

When anybody in the chat posts a message, their message is also then sent to the OpenAI API and it too will return back a vector quantity consisting of hundreds of decimal numbers. Using something called cosine similarity, we can compare these two vectors and return a value between 0 and 1 that represents how strongly they are correlated. In the case of these vectors representing the semantic meaning of a piece of text, this value then represents how related the new message posted in the chat is compared to all of the words or phrases on the blocklist.

When any new message returns a very high cosine similarity with any message added to the blocklist, the user is automatically kicked and banned from the chat. This means that if a spammer is trying to blackmail you by posting lots of messages in your chat claiming you are a scammer, you can add his messages to a blocklist, and if he rejoins on any new accounts, he cannot simply reword his previous comments slightly to evade the filter. Any message he sends which strongly resembles a message on the blacklist will be detected and he will be automatically banned, making life much more difficult for any would-be raiders. Similarly, someone may conduct a raid by posting messages dozens/hundreds of times reading “DM @DRJEET TO STOP THE FLOOD”, or something similar. Now, rewording the message slightly will not evade detection with this bot. Whenever the spammer comes up with a message that differs enough to evade the message on the blocklist, you can blocklist any further comments they make until they have nothing left to say.

You can also add custom phrases to your blocklist to ensure that anybody who says anything similar in the future is automatically purged by the bot. To do this simply go to your group in which AntiRAID.AI is an admin and type the command “/smartblock [your text]” where [your text] is any word or phrase you want to be automatically filtered in future.



Antiseptic (watch out for jeets impersonating me in DM) joined the group

Antiseptic (watch out for jeets impersonating me in DM)
this guy is the same dev from squidgame. definite honeypot, don't buy
1 20:16

Antiseptic (watch out for jeets impersonating me in DM)
this guy is the same dev from squidgame. definite honey...
/smartban
20:17

AntiRAID.AI removed Antiseptic (watch out for jeets impersonating me in DM)

Ultimate Spammer joined the group

Ultimate Spammer
it's the same dev that made squidgame. do not buy or he'll honeypot you.
20:22

AntiRAID.AI admin
Ultimate Spammer
it's the same dev that made squidgame. do not buy or he'll hon...
STRONG SEMANTIC MATCH ON SPAM/ABUSE FILTERS -- USER @whataname1234 WAS BANNED FOR THIS POST: 'it's the same dev that made squidgame. do not buy or he'll honeypot you.'.

If this was a mistake, please click /undo_lastban to unban them!
20:22

AntiRAID.AI removed Ultimate Spammer

Example: a user intent on causing trouble and fudding in our group is banned by the admin using the /smartban command. When he rejoins on an alt-account and tries to post the same FUD with a re-worded message, AntiRAID.AI detects the similarity and automatically bans his alt account. ***NOTE: in the first release the offending messages are automatically deleted, but kept here only for demonstration***

‘ChatGPT’ analysis:

While embeddings may be extremely useful for text comparison purposes, in order to stop spammers posting small variations of the same message, they may not be so effective at pre-empting spam attacks before they happen. As of course, it’s difficult to know the content of a spam message before it is posted.

This is why, after the release of Stage 2 of our roadmap below, we will be implementing ‘ChatGPT’ analysis to work in conjunction with text embeddings. ‘ChatGPT’ analysis refers to the infamous chatbot, ChatGPT by OpenAI, which can be queried on almost any subject to generate text and give accurate answers that would be indistinguishable from a human expert. In this context, ‘ChatGPT’ analysis refers to the use of the OpenAI text completion endpoint. In much the same way you would ask ChatGPT a question about an essay or some code, we use the OpenAI text completion endpoint and send a request with a) the message posted by the user in your group b) some context regarding the name of your group and what it is about and c) a series of questions about whether the post contains spam, FUD, solicitation or DM requests.

We then request that the API returns an answer as a piece of code which our program can understand, detailing whether each of these elements are detected, and giving a reason for its judgement.

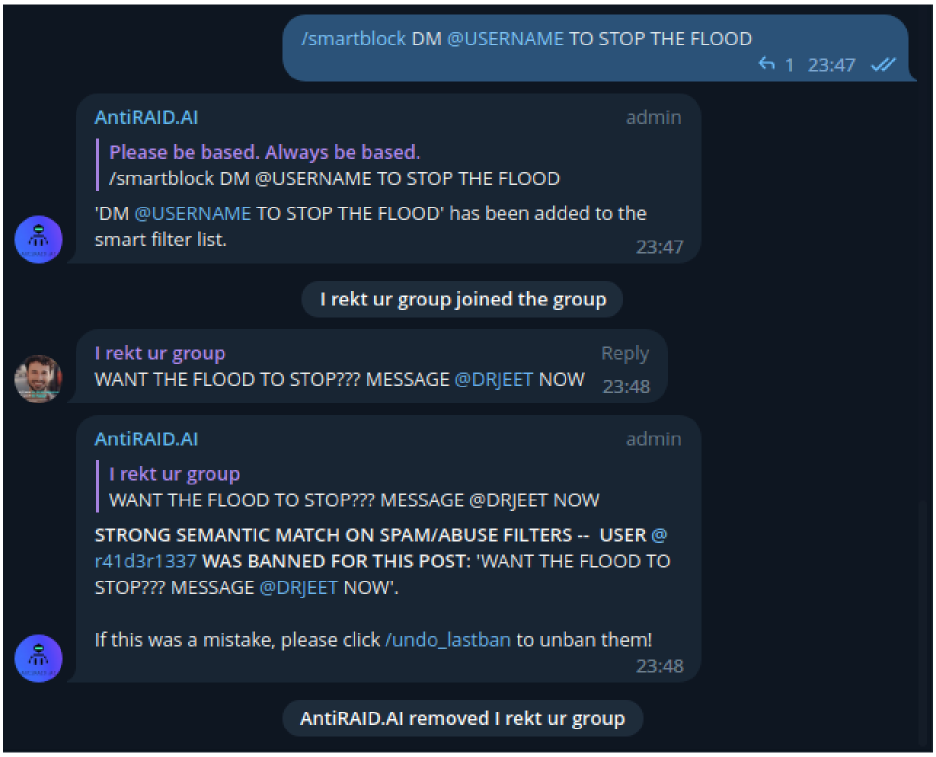
This means that, without knowing any context in advance, if a marketer or a spammer joins your group to promote their product or service in your group, their message will be queried against the OpenAI API and it will be specifically asked to determine whether the message contains, FUD, solicitation, spam, or DM requests.

If any of those elements are detected by OpenAI, the user’s message will be deleted and the user will either be warned, muted or banned according to user settings.

One particularly neat feature about this is that AntiRAID.AI will post a message in the group stating that the user has been muted/warned/banned, containing a preview of their original post with any links removed. It will then state whether the user was muted/warned/banned for FUD, solicitation, spam or DM requests, and give an AI generated reason as to WHY it believes the user’s post constitutes one of these.

This allows AntiRAID.AI to become a tool that can increasingly automate the moderation process and replace human labour, potentially saving you money. By using AI to analyse text for a broad range of undesirable content, and generating a reason for its decisions, AntiRAID.AI can begin to display human-level intelligence and reasoning for its decisions. Allowing you to sit back and relax, as you watch spammers and scumbags get automatically removed from your group.

ROADMAP AND DEVELOPMENT:



/smartblock DM @USERNAME TO STOP THE FLOOD

1 23:47

AntiRAID.AI admin

Please be based. Always be based.

/smartblock DM @USERNAME TO STOP THE FLOOD

'DM @USERNAME TO STOP THE FLOOD' has been added to the smart filter list.

23:47

I rekt ur group joined the group

I rekt ur group Reply

WANT THE FLOOD TO STOP??? MESSAGE @DRJEET NOW

23:48

AntiRAID.AI admin

I rekt ur group

WANT THE FLOOD TO STOP??? MESSAGE @DRJEET NOW

STRONG SEMANTIC MATCH ON SPAM/ABUSE FILTERS -- USER @r41d3r1337 WAS BANNED FOR THIS POST: 'WANT THE FLOOD TO STOP??? MESSAGE @DRJEET NOW'.

If this was a mistake, please click /undo_lastban to unban them!

23:48

AntiRAID.AI removed I rekt ur group

Example: an admin anticipates a flood attack demanding they DM a user to stop it. He uses the /smartblock command to add a custom phrase and when a user joins trying to post something similar he is automatically banned.

At the time of release, AntiRAID.AI is currently in proof-of-concept phase to demonstrate the efficacy of using AI semantic comparison to filter messages that are *similar* to banned text. We have a vision for this bot to be something even greater than what it's starting out as, and we have a five-stage roadmap outlining how we hope the bot will develop as it grows. In short, we hope to develop AntiRAID.AI into the Swiss-army knife of Telegram bots, fulfilling the functions of community moderator bots like Rose or SHIELDY with smart features utilising AI to make it much easier for devs to manage and monitor their groups.

Stage 1 – Launch and Proof-of-Concept release

On the 18th of December, we will conduct a stealth launch with a working release of AntiRAID.AI. We will invite a small group of investors to our launch group, and shill on several community pages thereafter. After attracting initial attention at launch, we will then apply for Top 5 trending status on BuyBotTech's trending, a bot channel for the widely used BuyBotTech, instantly opening up a vast audience. A basic website consisting of a landing page and important links will be live at launch, with an improved version released no more than 2 days later. We will then pursue a multi-faceted marketing campaign during the first week including coordinated shilling and advertising on 4chan, Reddit, Twitter, coinchan.xyz etc and also purchasing calls where possible from reputable influencers with reasonable prices.

During this phase we aim to build as much momentum as possible and establish a brand name, as more and more groups add the bot to test its features.

Stage 2 - ChatGPT analysis, customisation and upgrading the UI: release major update with improved algorithm and customisability for admins:

The bot will be upgraded to allow each message to be queried by OpenAI/ChatGPT to analyse whether it contains spam, solicitation, FUD or requests to check DM - vastly improving its ability to detect unwanted messages. As discussed previously, this will involve directly querying the OpenAI text completion endpoint and asking it to analysis messages for FUD, solicitation etc and return a Python object determining whether it contains these elements and reasoning for a decision. When any action is taken after a positive result, the bot will give an AI-generated reason for its decision in the group.

Customisable settings will be added allowing the user to adjust the sensitivity of the bot over 5 different levels, that is, the strength of a similarity required before it is regarded as a match. We will also be introducing settings allowing moderators to choose whether the bot mutes or bans, and allowing a user-defined number of warnings to be given beforehand, as well as allowing moderators to more easily manage the blocklist by removing individual items. We’ll also add settings allowing you to choose whether the blocklist for your group remains accessible for users, or private.

Stage 3 – Smart Raid detection, official portal, dynamic security settings

In stage 3 we’ll significantly improve the anti-raid features of the bot by adding functionality that automatically detects potential raids, notifies admins, and takes according action. Particular attention will be paid to new users, and the vectors representing the semantic content of their messages will be saved to a database and regularly compared. If it is found that 3 or more messages from new users in a short period of time have a high semantic similarity, RAID MODE will be activated. Slow mode will be automatically enabled in the chat, media will be locked down, and the messages that triggered the alarm will be added to the blocklist. The sensitivity levels will be temporarily increased, making messages with semantic similarity to the blocklist more likely than usual to return a positive match.

If the admin has chosen that users normally receive warnings before being muted/banned , these will temporarily be removed or reduced, and if users were normally muted, the moderators would have the option of temporarily banning on the first offence while RAID MODE is activated.

RAID MODE will also be able to be activated manually by admins, and, as outlined above, all features will be fully customisable. We'll also add an 'Antiflood' feature in stage 3 akin to the one found in Miss Rose, which temporarily mutes users who post more than X number of messages in a short period of time.

We hope that by Stage 3 our token will have been enough of a success to achieve listings on CoinMarketCap and CoinGecko, and to have purchased at least one major caller. By the end of stage 3 we expect mass exposure for the project as it becomes recognisable to large numbers of people in the crypto and wider Telegram space.

We'll also add an official portal during stage 3 to allow the bot to directly compete with the likes of Safeguard Portal, MEVFree or Calsibot. These portals allow a private group to be directly linked to a public channel, whereby users on the public channel can access the private group after 'proving' they are human by completing a picture-based CAPTCHA on the AntiRAID.AI website. While these CAPTCHAs are not foolproof and can be defeated by AI, they do limit the amount of spam bots that can access the group, the portal will bring huge exposure for the bot, and as mentioned this will give us the direct ability to compete with other group protection bots offering this feature.

Stage 4 – Custom-Trained Models, Google AI, smart monitoring and global blocklists, premium features and token integration

In stage 4 we aim to release a major update to the bot which uses language models specifically trained on cryptocurrency Telegram groups, and, if the bot achieves popularity in subcultures on Telegram outside of crypto, from those groups as well. As anyone native to this space will know, there is a large amount of internet slang present in these groups ("jeets", "bogged", "rugs", "Chads", "sendor") and significant amounts of technical jargon such that an AI trained on text from the general internet may have trouble understanding its meaning and context.

By releasing our own pre-trained model, this will improve even further the bot's ability to recognise the semantic relation between pieces of text and thus become more accurate in detecting similarity.

We will also monitor the content of phrases across all groups and channels being added to the blocklist, and create a “Global blocklist” for all groups. In short, if we detect identical phrases being added to the blocklist in multiple groups, these will be temporarily added to a “global blocklist” applying to all groups. This ensures that, when there are particular pests running riot in multiple groups, AntiRAID.AI will be on the lookout for them automatically across all groups smart enough to be running our bot. All without the admins having to do anything. We will also manually review terms being added to groups ourselves, and anything particularly undesirable we may add to the global blocklist.

In order to ensure that the value of our token is linked to the utility the bot provides, during Stage 4 we will require token holding for premium features of the bot. Please see the section below regarding the MOD.AI() token for a full list. There will be a requirement of holding 100,000 in order to enable these premium features. *(N.B. as the price of the token increases we will reduce this amount to avoid usage of the bot being prohibitively expensive, however groups which already purchased will be required to maintain their balance to continue using)*

By the release of Stage 4 we also expect Google to have made public access to its LaMDA language model, with the recent announcement of its soon to be released BARD chatbot. We will, as a premium feature, give token holders the ability to choose between using GPT-3 or LaMDA to analyse the chat. Moreover, we will add a special feature whereby administrators will have the option of using GPT-3 and LaMDA in tandem with one another. You will have the option of requiring BOTH LaMDA and GPT-3 to return a match for action to be taken, thus virtually eliminating false positives. You will also have the option of requiring that action is taken when EITHER GPT-3 or LaMDA detect a match, thus maximising the chance that spammers and bad actors will be detected by the bot.

Stage 5 – Developing into fully fledged community bot, GPT-4, AI interaction, and release of Discord bot

In stage 5 we will aim to develop AntiRAID.AI to having all the features of a fully fledged community assistant such as Rose, Shieldy or MEVFree. We will add greetings, support for /filters (allowing pre-written text to be posted in the chat when a user types a certain filter), manual warns, mutes and bans, among other features yet to be announced.

We aim to become the superior of Telegram community moderator bots and eliminate the need for using any others. On top of our smart moderation features and integration of general community moderation tools, in stage 5 we will also aim to enable AntiRAID.AI to interact with the community using the power of AI. Using OpenAI’s text generation features, and with the help of the custom-trained Telegram models released in stage 4, we will give AntiRAID.AI the ability to randomly join in conversations and give its thoughts on any topic the community may be discussing. Or perhaps we might program AntiRAID to start sharing its dankest memes…….

Once the bot is perfected, we will develop a Discord bot with all of the same features in order to capture an even bigger audience.

With GPT-4 expected to release in 2023, we anticipate that it might be available to the public at or around the same time Stage 5 is released (there have been estimates as early as Q1-Q2). GPT-4 is expected to be substantially more powerful than GPT-3 and have an even better understanding of context, and is rumoured to have dramatically more parameters. Therefore, once GPT-4 is released we will as a premium feature allow token holders to use GPT-4 to analyse the chat.

This is not an exhaustive list of things that AntiRAID.AI will be able to eventually do. I have no doubt we will have some smart folk joining our Telegram and suggesting features I would never even have thought of, so this will be very much a community driven project. If you can think of a cool new way for AntiRAID.AI to utilise OpenAI, join the Telegram and let us know.

THE TOKEN

The cryptocurrency accompanying the release of AntiRAID.AI will be an ERC-20 token on the Ethereum network trading on Uniswap. It will have the following tokenomics:

- TICKER: MOD.AI()
- 100 million total supply
- 2 million maximum hold per wallet
- 1 million maximum transaction amount

- 1000 USDC initial liqudity (we are deliberately choosing a stablecoin as our base pair to avoid being subject to harsh movements in the rest of the crypto market)
- 5.75% tax (on buys/sells/transfers) :
 . 3% goes towards marketing and hiring external developers
 . 1.5% goes to paying project team salaries
 . 1.25% goes to auto-liquidity

- 2 month initial liquidity lock – will be increased to 6 months when we hit $100K market cap. Increased to 2 years when we hit $420K.

Note: the contract we are using has a dynamic liquidity tax. Until the price of the token is very high and the market cap has diverged from the LP value greatly, the tokens from the liquidity tax will be sold for marketing. Therefore in the early days the marketing tax will be 4.25%.

As previously discussed, holding the token will be a requirement for using premium features after Stage 4 is released. **Holding the token will be a requirement for the following features:**

- ‘ChatGPT’ analysis to detect FUD, spam, solicitation and DM requests (each group will be granted a 2 week trial after adding the bot)
- Keeping AntiRAID.AI ad free in your group
- Premium embedding models when released (v3 of Davinci for text embedding)
- Ability to use Google’s BARD AI (when released) as an alternative to or in conjunction with OpenAI/GPT-3
- Eventual utilisation of GPT-4 with AntiRAID.AI

That sums up our vision for AntiRAID.AI. I’m sure you will agree there is much to be excited about in these early days, so please, join the Telegram, add the bot to any groups you moderate, and let’s get AntiRAID.AI in every TG group and channel imaginable.

Website: antiraid.ai
Twitter: twitter.com/AntiRAIDAI
Telegram bot: @AntiRAIDAI_bot
Telegram community: @antiraidai