

Community-Rated Misinformation Analysis

Mark Glasgow, Erin Connolly, Valerie Kwan, Hugh Merrell

We present Dolos, an extendable bot built for the Discord platform which aims to dispel Coronavirus misinformation in online communities. This bot can be added to community servers, giving moderators the utilities to better safe-guard their members. Dolos encourages user interaction via a credibility leaderboard and several ways to interact and with the responses. Given the recent explosion in Discord’s growth and the nature of the audiences that congregate there, Dolos provides a solid framework which is easily expanded with additional data-sources which is crucially needed as misinformation consumes our online communities.

Index Terms—Discord, Disinformation, Fake News, Coronavirus

I. INTRODUCTION

We were tasked with developing an interactive system that allows people to monitor, understand or support population health. In February, the director of the World Health Organisation announced that COVID was not the only public health emergency we were facing as a society, but that an “infodemic”(who) of misinformation had been spawned in response, exploiting public fear and uncertainty and plaguing governments and scientific communities. Posing a serious risk of strong negative impacts on individual users and the broader society. A study published in the American Journal of Tropical Medicine and Hygiene approximates that 800 people have died and 5876 were hospitalised following a myth which perpetuated that drinking cleaning products would prevent coronavirus(md), a survey by YouGov found that almost one-third of Americans and half of Fox News viewers thought that vaccines were a ploy to insert microchips into people(sanders2020). To combat these challenges effectively, technology must be developed within so-

cial media and instant messengers themselves.

II. DESIGN

To tackle the plethora of issues caused by the spread of misinformation, we devised three potential products that could aid in this effort, thus encouraging good population health. These consisted of a chat bot, a browser plugin and a bot for the popular messaging service, Discord.

1) Chat Bot

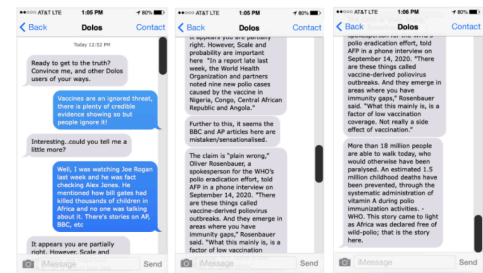


Figure 1: Survey

The bot would respond to users messages regarding Coronavirus, allowing users to interact with a bot on a conversational level, providing information in response to a question or analysing the legitimacy of

statements. The wire frames shown in 1 illustrate the proposed functionality.

This product was desirable for numerous reasons. It could drive engagement through the novelty of machine learning and provide a tailored experience unique to each individual, helping dispel any misinformation they have acquired.

2) Browser Plugin

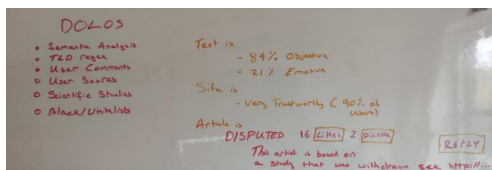


Figure 2: Survey

The second idea was a browser plugin that would allow users to determine the nature of a web page. The plugin would perform an analysis, showing the percentage of the emotive and objective language used, whether or not the site is trustworthy and the other plugin users’ feedback on a given page.

The plugin had the perk of being relatively simple to implement whilst still being effective in allowing a user to determine the reliability of a given web page quickly. However, there were several drawbacks to this particular design that made us hesitant to follow through with it. Firstly, on Google Chrome, which has the most significant browser market share, most extensions have less than 1000 active users, and about 46% have a 0-star rating (**Chromestats**). Furthermore, it seems probable that individuals who are most likely to engage with sites promoting disinformation would be unlikely to engage with the plugin. It also has the problem of not integrating well with social media, where most disinformation spreads. If a plugin user were to encounter a link on their social media feed promoting fake news, the

plugin would not pick up on this.

A. Discord Bot

Our first two designs, whilst having their drawbacks, possessed certain advantages that we wanted to carry over to a new product. The chatbot was engaging with its use of automated responses; however, the vast scope of the problem made it infeasible. The browser plugin was simplistic and effective in its use of links, but the tangible real-world use was questionable, and it did not effectively combat social media posts. With this in mind, we came up with a new design that used a discord bot to highlight disinformation. When a user posted a link on a discord chat, the bot would analyse it, similar to the browser plugin, showing the percentage of the emotive and objective language used and whether or not the site is trustworthy. There is a basic proof-of-concept in the chatbot implemented into Dolos utilising a BERT Classifier which answers basic factual questions regarding Coronavirus. This could be further expanded to provide a more cohesive experience.

dunce points are awarded to the posters of queried links which return a corpus hit. Users are ranked by the level of disinformation they have posted which is then displayed on a leaderboard.

The implementation of a discord bot that responded to links allowed us to take the positive automation based aspects of the chatbot without having to implement all the associated complexity. Furthermore, the Discord communities vital utilisation of bots as guardians over their communities would provide a real-world use scenario for Dolos. Large Community Server Administrators could easily add Dolos, mitigating the risk of exposure to fake news by their member base.

1) Survey

We created a survey as part of our initial design phase to solidify what people would want in a Discord bot. We set it up through Google forms and sent it to a few friends and classmates, in order to get our target user group, which was people who often use Discord to chat with their friends. The main aim of the survey was to gather interest in our idea and inform some of our design decisions; for example, how much human involvement should be involved in decision making. Nine responses were received. As part of the ethics procedure, we made sure that all respondents gave their full informed consent. They were assured that all information stored was anonymous and they could edit or withdraw their response in accordance with GDPR (Figure 15).

Question 1 asked if subjects felt that Discord might be a platform where fake news could spread easily (Figure 17). This question was vital as we needed to find out if people agreed with us that Discord could be used to spread fake news. As our goal is to encourage good population health by eliminating COVID-19 related fake news, we needed to ensure that people thought that fake news could spread through Discord. If people did not expect to find fake news on Discord, then they would probably not install our bot on their server. However, everyone thought that Discord could definitely or maybe be used to spread fake news, so we knew that Discord would be a good platform for discouraging fake news.

Question 2 asked if users thought the link rating process should be entirely automatic or if it should have more human involvement (Figure 18). With our bot's link rating process, we could have entirely automated link rating through corpus matching, logistic regression and sentiment analysis. We could also

have a user rating system through reactions, and we were not sure which method was better. No respondents wanted an entirely automated rating, with most wanting a mix of the two. A mix of approaches allows for a more robust system with less room for manipulation.

Question 3 asked if users thought the bot should only respond to COVID-19 related links (Figure 19). With many links being posted in a Discord server, the bot may produce many messages which could frustrate users and also slow down the server. We decided to ask if users would want the bot to only respond to COVID-19 related links in order to limit some of the spam, as well as ensure it does not respond to links that are not supposed to be news of any sort. Most survey participants agreed that the bot should only respond to links containing COVID-19 news.

We also had an open response section for respondents to suggest any features that may be useful to them, that we may not have considered (Figure 20). We only received one response, who said that the only link responding was enough, to keep the bot simple.

Overall, feedback for our bot was positive, with most respondents saying they would add this bot to their server. This showed us that our idea was popular, and could be used by many people in their Discord server to discourage fake news, thus improving population health.

III. SOCIAL MEDIA

While fake news was undoubtedly present in years past, the rising popularity of social media has made it far more prevalent. As of August 2017, around 67% of Americans got their news from social media. Features such as retweets, comments and likes help drive engagement and disseminate information

rapidly. Paired with the echo chamber effect that frequently occurs on individuals’ social media feeds, communication with a particular bias or viewpoint is often amplified. As a result, a user who has an inclination towards less mainstream health information, such as anti-vax, has their online experience continuously tweaked until they may find that their feed is filled with nothing but fake news and misinformation. Even those outside of these circles can quickly come into contact with misinformation without being aware of the untrustworthy source, mistaking false claims as credible, and sharing them with others.

IV. DOLOS

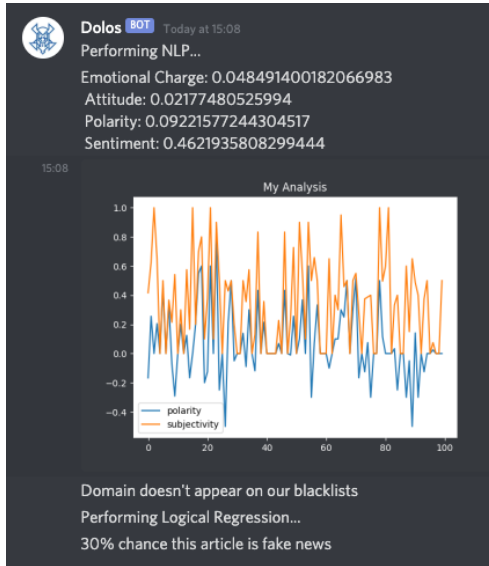


Figure 3: Misinformation analysis

A. Discord

“These kids are not calling, texting or Skyping each other anymore. They’re all just Discording,”

Discord is a quickly growing messenger application that launched in 2015—gaining 87 million members in its first three years, growing at twice the rate Twitter did during its launch. This growth was predominantly on the back of video-game communities

before it gained notoriety in 2017 when the New York Times reported that it was becoming a “favourite chat app” of the alt-right.(**NYT**)

Since then Discord successfully managed to launch itself into the mainstream, gaining \$100million in investment and 250 million registered users.(**WSJ**)

What sets Discord apart from most mainstream social media sites and messenger services is that it allows almost complete anonymity. There have been some attempts to reign this in recently, with the recent launch of Community Servers - which prompt server owners to enable additional security configurations.

While communities are relatively autonomous, unless strict care is taken to lock-down access, inevitably others will find their way into the server.

B. RedBot

RedBot is a fully modular self-hosted bot framework distributed under the MIT License.(**Red**) We selected RedBot as it allowed us to rapidly deploy a bot with a wide variety of code examples and tutorials to speed up the process. This ensures a robust interface and modular capacity for any server administrators who utilise Dolos.

C. Functionality

We trained a rudimentary logistical regression model on a COVID-19 misinformation data set. The article text is then fitted to this model, and a value returned. At the moment, this is merely calculating the probability of each sentence being fake-news, and returning that as a percentage.

The domain is compared against several ‘corpora, allowing us to make a clear call and penalise the user. Most notably we are using a list of 366 domains

which are known sources of disinformation specific to the coronavirus, curated by NewsGuard. We also performed a manual sweep of sites with known conspiracy sites such as *Parler* and manually added obvious sources of misinformation.

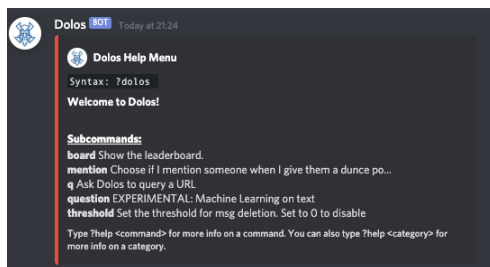


Figure 4: Misinformation analysis

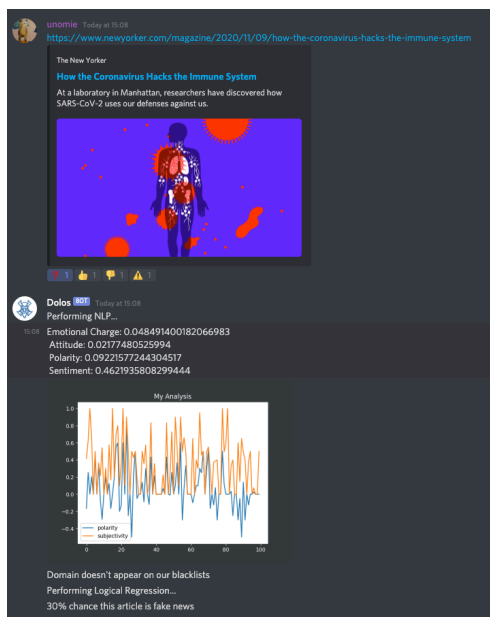


Figure 5: For sites that aren't obvious misinformation, detailed sentiment analysis and machine learning is used

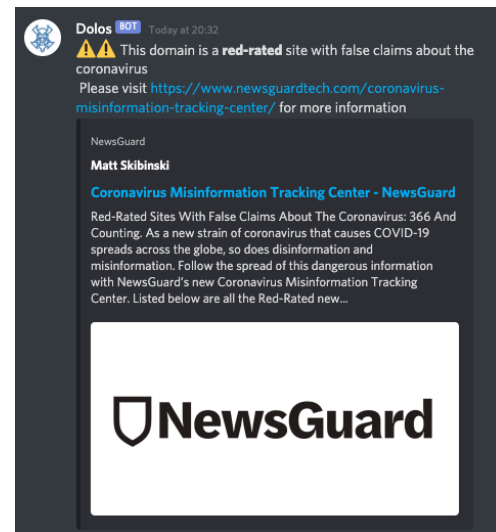


Figure 6: Domain is checked against COVID misinformation corpus

D. User Interaction

Users can interact with Dolos in several different ways. For the most basic use, checking if an article contains reliable information, users post the link into the chat and anyone can react with the question mark emoji to prompt Dolos to analyse it. Dolos' analysis will be posted in response, allowing all users to see the predicted legitimacy of the post.

The ranking system encourages user engagement and provides an opportunity to improve our dataset. When Dolos judges a link to be misinformation, the poster receives a 'dunce point', whereas the person who queried the link receives a reputation point as a reward.

The reason why we designed that users have to react to the link for analysis is due to the prevention of spam. Users might like to share different sorts of links, including funny cats videos or Dropbox of study notes. If Dolos responds to all these irrelevant links, this would quickly overrun the channel and damage the user experience.

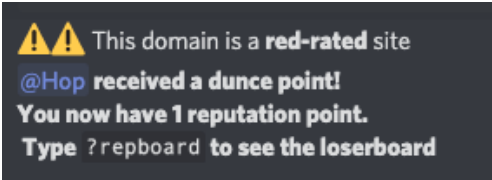


Figure 7: Users can be shamed with dunce points

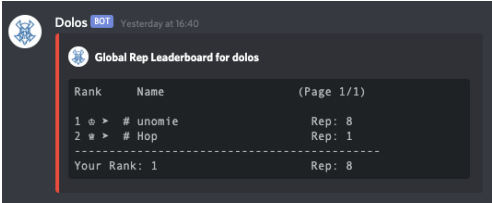


Figure 8: Users can see their ranking on a repboard

We have an alternative design that Dolos can review if the link contains any COVID keywords, and only analyse the ones that do include the keywords. The advantage of this design is that the users do not have to worry about getting in time to react to an emoji and can benefit from the automatic progress. However, due to the time consuming of implementing, our version of Dolos has not included this function yet.

A Dolos response example is shown below in figure 11. For each analysis response, three reactions are attached to the comment. Users can react with ‘thumbs up’ if they agree with the result, ‘thumbs down’ if they do not agree with the result, or ‘warning’ emojis if they would like to challenge the Dolos analysis result. These results are then recorded against a weighted value for the domain. These weighted values are used to verify Dolos’ judgement calls, forwarding any discrepancies to a moderator for approval.

Emojis are inserted dynamically to gauge sentiment and the accuracy of our models. When a user reacts to ‘challenge’ the results, Dolos then prompts them to enter supporting information.

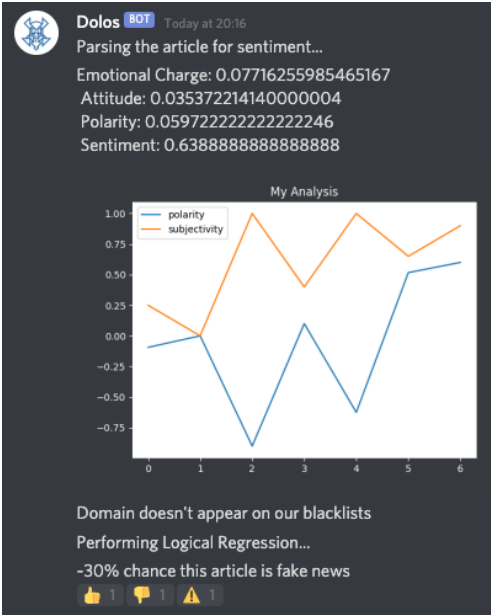


Figure 10: User is prompted to agree / disagree or challenge

Dolos will then analyse the backing sources, verifying they do not match any of our predefined corpora, then the challenge reasoning will be voted by other Dolos users by emojis reacting - ‘thumbs up’ for agreeing and ‘thumbs down’ for not. If the sources seem reliable to Dolos and the feedback gains enough agreement, the ‘challenge’ would be accepted, and the user who provided the challenging feedback will get reward points for the ranking system; otherwise, the ‘challenge’ would not be accepted, and the user will gain a ‘dunce point’.

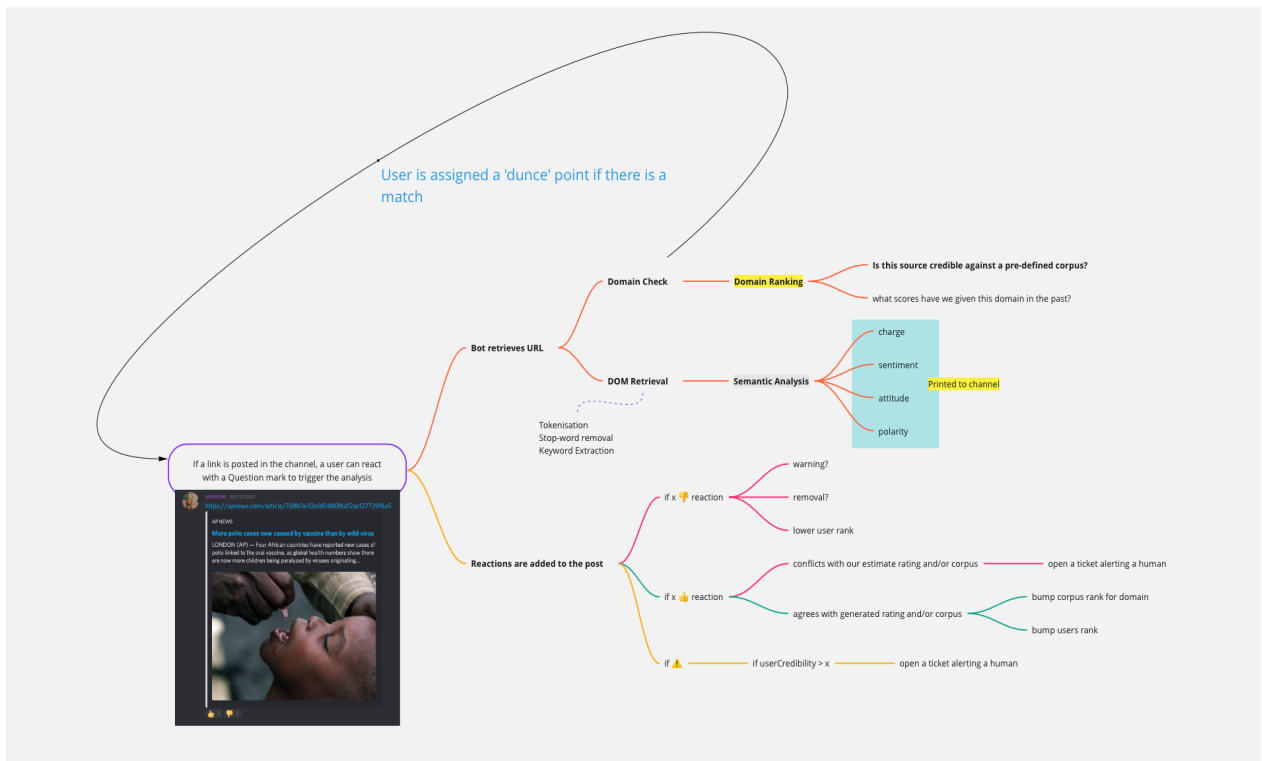


Figure 9: Mindmap outlining Dolos' features

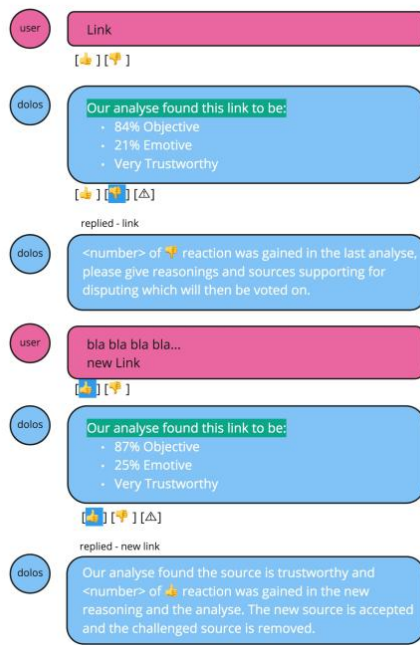


Figure 11: An example of users' 'challenge' Dolos results

With the ranking system, users can check their ranking by typing a command into the chat. Dolos would pick it up automatically. This is shown in figure 12.

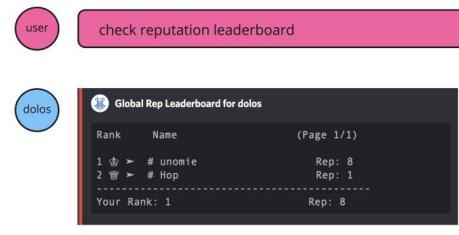


Figure 12: Leaderboard

E. Evaluation

The reason why we designed that users have to react to the link for analysis is due to the prevention of spam. Users might like to share different sorts of links, including funny cats videos or dropbox of study

notes. If Dolos responds to all these irrelevant links, this would quickly overrun the channel and damage the user experience.

We have an alternative design that Dolos can review if the link contains any COVID keywords, and only analyse the ones that do include the keywords. The advantage of this design is that the users do not have to worry about getting in time to react emoji and can benefit from the automatic progress. However, due to the time consuming of implementing, our version of Dolos has not included this function yet.

1) Think Aloud

After finishing our initial implementation of Dolos, we performed three think-aloud experiments in conjunction with an interview in order to establish how effective Dolos was and any design choices that should be reconsidered. One particular think-aloud took place on two subjects who were added to a server with Dolos, informed of its general functionality and asked to have a conversation about Coronavirus which referenced news sources. One individual was asked to role-play a ‘COVID sceptic’ who was less likely to subscribe to mainstream media. The two others took place on individuals who were asked to post links of their choosing and analyse them with Dolos. After the think aloud, subjects were asked what they thought of Dolos and how it could be improved.

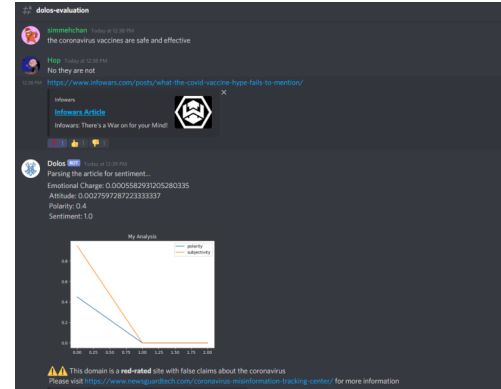


Figure 13: Think aloud with two users

The think-aloud and interview resulted in mostly positive feedback. Users quickly established how to use the question mark emoji to prompt Dolos and found the thumbs up and thumbs down functionality intuitive. However, we found that users were concerned that the ranking system could potentially facilitate the spread of fake news. The implementation at the time meant that the user who posted the most misinformation appeared top of the ‘loserboard’, highlighting particularly unbelievable commenters. Our test subjects suggested that this would likely encourage trolling in larger servers in which users would attempt to maximise their points. As a result, we redesigned this particular feature to score users who posted the most reliable content to be ranked the highest.

V. REFLECTION

Our current iteration of the bot is still a prototype and as such there are many aspects that could be improved. One thing that could be improved is the rating system. Users who call out fake news could be awarded points, to produce a more positive leaderboard. This could encourage people to call out fake news more often, leading to high engagement and competition. Users could also be encouraged

to share links that they know are trustworthy, seeing what result they get from Dolos. In return, they get a point and Dolos gets new data to use in further analyses, improving reliability as the dataset expands. Another feature we considered including was giving Dolos the ability to give daily news links from a trustworthy source, instead of only verifying what it is given by users. It would also be useful to consider ways to expand the audience to people who may be reluctant to install the bot on their server, for example COVID skeptics.

VI. CONCLUSION

Our interactive system for supporting population health is Dolos, a bot created for the Discord messaging platform. It does this by aiming to decrease the amount of fake news surrounding COVID-19. The bot responds to user messages containing links that mention Coronavirus, analysing the content of that link for disinformation. It does this with a mix of logical regression, sentiment analysis, and pre-existing corpuses containing blacklisted sites. Along with this, users can challenge the bot's decision. With this, we hope to decrease the number of users who spread fake news, on purpose or not. We considered using a chat bot or browser extension to achieve this, however settled on a Discord bot. Through the use of surveys and think aloud evaluations, we have focused on what users want from a Discord bot of this kind, and what they thought of our existing designs. We also presented some ideas for future improvement, for example changing the rating system. Overall, we believe this system is helpful for encouraging population health as disinformation can be harmful to it, and Dolos

aims to prevent that.

VII. APPENDIX

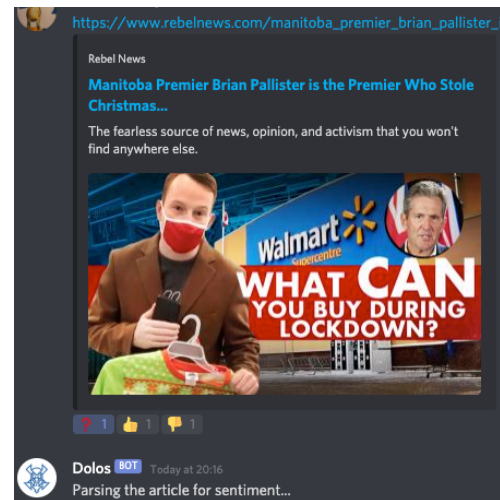


Figure 14: Sentiment Analysis based on user-trigger

Before beginning, please confirm:

- ☐ You agree to take part in this survey, and consent to the data you provide being used for the developme...
- ☐ You are 16 or older.
- ☐ You understand that you can withdraw from the survey at any point.
- ☐ You do not have an impairment that may limit your understanding or communication.
- ☐ You are not in a position where any of the group members hold a position of power over you.

Figure 15: Survey

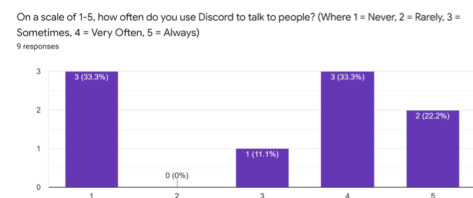


Figure 16: Survey

Do you feel that Discord may be a platform where fake news could spread easily?
9 responses

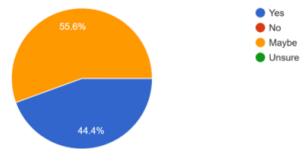


Figure 17: Survey

Would you like the link rating process to be entirely automatic or would you like more human involvement? (With 1 being entirely automated (only sentiment analysis etc) and 5 being only human involvement (user ratings of links))
9 responses

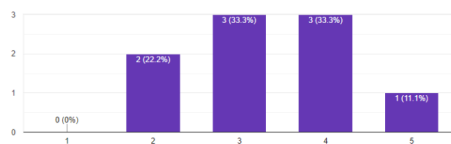


Figure 18: Survey

Would you like the bot only to respond to COVID-19 related links (for example through keyword matching) or any links posted?
9 responses

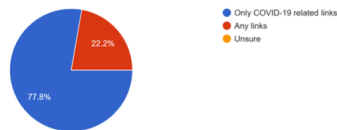


Figure 19: Survey

Are there any additional features you would like to see added to the bot?
1 response

only link responding is good enough, to keep it simple

Figure 20: Survey

How likely would you be to add this bot to your Discord server? Please rate between 1-5 (Where 1 = extremely unlikely, 2 = unlikely, 3 = neutral, 4 = likely, 5 = extremely likely)
9 responses

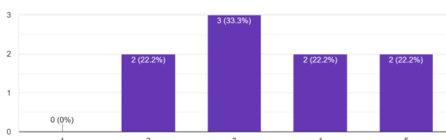


Figure 21: Likelihood of adding

REFERENCES

- . [S.l.]: New York Times, 2017.
- BROWN, Adam. **Discord Was Once The Alt-Right's Favorite Chat App. Now It's Gone Mainstream And Scored A New \$3.5 Billion Valuation.** [S.l.]: Forbes, 2020.
- REDBOT - OVERVIEW. [S.l.: s.n.].
- 8 INTERESTING STATS YOU DIDN'T KNOW ABOUT CHROME WEB STORE . [S.l.]: Google, 2020.
- MIT TECHNOLOGY REVIEW: THE CORONAVIRUS IS THE FIRST TRUE SOCIAL-MEDIA "INFODEMIC". [S.l.]: MIT Technology Review, 2020.
- ISLAM SAIFUL, Md. **COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis.** [S.l.]: The American Society of Tropical Medicine and Hygiene, 2020.
- JARGON, Julie. **The Dark Side of Discord, Your Teen's Favorite Chat App.** [S.l.]: The Wall Street Journal, 2020.
- MUSTAFARAJ, Eni. **Discord: The Next Big Thing? A First Look at Discord's Growth, Servers, and Users.** [S.l.]: ethanchiu, 2020.
- PIERCE, David. **How Discord (somewhat accidentally) invented the future of the internet.** [S.l.]: Protocol, 2020.
- SANDERS, L. **The Difference between What Republicans and Democrats Believe to be True about COVID-19.** YouGov. [S.l.]: MIT Technology Review, 2020.