Exploring Crowdsourced Content Moderation Through Lens of Reddit during COVID-19

Waleed Iqbal
w.iqbal@qmul.ac.uk
Queen Mary University of London
United Kingdom

Muhammad Haseeb Arshad
marshad4@ualberta.ca
University of Alberta
Canada

Gareth Tyson
gtyson@ust.hk
Hong Kong University of Science and Technology
China

Ignacio Castro
i.castro@qmul.ac.uk
Queen Mary University of London
United Kingdom

ABSTRACT

In 2020, when COVID-19 struck, social media gained even more influence in people's lives due to increased online activity. This event led to a surge of false information and cyberbullying, making content moderation harder than ever. Given this challenge, exploring opportunities to explore content moderation solutions to reduce hate speech and fake news on social media is vital. In this paper, we examine if existing content moderation systems are enough during global pandemics and, if not, where gaps may lie. Due to its intriguing Decentralized Content Management System (DCMS), we chose Reddit as the key social networking platform for our hypothesis testing. We used 1.8 million Reddit posts from COVID-19-related subreddits from January 2020 to April 2021. Our findings reveal several significant trends regarding the effect of a worldwide event on content moderation methods designed to lessen the prevalence of hazardous content and fake news. In light of these considerations, we provide the results of comprehensive research conducted with particular attention paid to the user-generated material and the DCMS of Reddit.

CCS CONCEPTS

• **Information systems** → **Social networks**; **Data analytics**.

KEYWORDS

content moderation, reddit, hate speech, fake news, COVID-19

ACM Reference Format:
Waleed Iqbal, Muhammad Haseeb Arshad, Gareth Tyson, and Ignacio Castro. 2022. Exploring Crowdsourced Content Moderation Through Lens of Reddit during COVID-19. In *The 17th Asian Internet Engineering Conference (AINTEC'22), December 19–21, 2022, Hiroshima, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3570748.3570753

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AINTEC'22, December 19–21, 2022, Hiroshima, Japan
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9981-4/22/12…$15.00
https://doi.org/10.1145/3570748.3570753

1 INTRODUCTION

On March 11, 2020, WHO declared COVID-19 a global epidemic, which acted as a cherry on top of the massive wave of misinformation on online social platform [3, 21, 51]. During this pandemic, misleading news and hate speech caused many incidents resulting in the loss of property and life. Some notable cases include the lynching of people in Palghar Village, India, over allegations of kidnapping children and organ theft [34], and the destruction of 5G cell phone equipment because of claims that it spreads the COVID-19 virus [5, 47].

Fake news and hate speech are common problems for social media moderators [33, 36]. Malicious actors can easily post fake news and harmful content faster on social media than on conventional media sources. **Fake news** refers to completely fabricated reports. It threatens press freedom due to the public's diminishing faith in legitimate news outlets. Similarly, **toxic content** is another form of harmful data. Hate speech, misogynist, abusive, and violent material are all examples of the toxic content, as defined by [17]. Different social networks typically employ one of the following three main methods for regulating user-generated content [1]:

- **Automation** refers to pattern matching, such as text-based Regular Expressions (RegEx) or machine learning-based classifiers to filter user-generated content on social networks.
- **Crowdsourcing** where social media users flags a post for moderators by up/down voting a post.
- **Professional Human Review** where companies employ a team of human moderators to verify and filter user-generated content.

Facebook and Twitter use sophisticated machine learning-based algorithms and human moderators to police the information posted on their platforms. They also use hired workers for their content moderation systems, as cited in [44]. Facebook, however, also allows group managers to serve as volunteer moderators to police the group's content. On the other hand, Reddit is unlike social media giants like Facebook or Twitter. It uses all three of the aforementioned content moderation strategies as part of a DCMS. This DCMS saves Reddit a lot of time and resources and gives the user more power to decide about content on Reddit [7].

Reddit is self-referred as the *Front Page of Internet*, which consists of small communities based on different topics called *subreddits*. A user-generated post on a subreddit is referred as a *submission*. Reddit

has 430 million members and 2.4 million subreddits as of April 2022 [2], out of which 3172 discussed COVID-19. Anyone can view a subreddit's content, but only members can post and comment. Each subreddit has its own rules and volunteer moderators. Considering user-generated content and user volume, the DCMS makes Reddit a perfect platform for investigating the effectiveness of content moderation systems during COVID-19.

This study's key contributions include analyzing the impact of global events (in our case, COVID-19) on content moderation methods and figuring out the opportunities to improve these to mitigate fake news and hate speech propagation. We observe the efficiency of Reddit's content moderation system in light of four research questions:

RQ1: Is toxic content allowed across subreddits, and how strictly are subreddit rules enforced?

RQ2: Does the subreddit's regional focus impact toxicity and fake news during COVID-19?

RQ3: How can we improve the content moderation systems of Reddit for Detecting Fake News?

To answer these research questions, we use 1.8 million submissions from 3172 COVID-19-related subreddits during January 2020–April 2021 collected using the *Pushshift API*[1].

2 RELATED WORK

Fake news on social media posses fundamental challenges, which, even to this date, social networking companies have failed to eliminate completely. Several authors have worked towards studying the effect of fake news on various scenarios. Grinberg et al. [18] conducted an exploratory analysis of the tweets during the 2016 US elections to show the extent of social media content manipulation. Zollo et al. [54] highlight the spread of false information about healthcare information in Italy and the US on Facebook. A very insightful work on political interactions and opinion influencing through Instagram is conducted in [45]. A comprehensive analysis providing a detailed description of disinformation dissemination through Wikipedia and the potential procedure of reporting the disinformation is provided in [38]. More details on how social media applications are abused for misinformation/disinformation and fake news dissemination are reviewed in [41–43, 52, 53]. Some works also focus on fake news and rumours dissemination over Reddit [8, 37, 40].

Similarly, the problem of hate speech inspired a growing body of work for various social media platforms like *Facebook* [12, 35], *Twitter* [4, 11, 15, 31, 49, 50], *Reddit* [9, 16, 28], *WhatsApp* [22, 23], *Whisper* [29], and *Yahoo News and Finance* [13, 30, 48]. Mathew et al. also studied the effects of hate speech in YouTube comments [27]. Some authors have worked on over-generalized and directed hate speech as well. El-Sherief et al. [14] have studied and observed that directed hate speech is more focused on personal anger, whereas generalized hate is more focused on religious affairs and types.

There is a lot of research on fake news and harmful content on social networks with centralised content moderation but relatively little on platforms with DCMS, such as Reddit. On top of that, it is still unclear how successful pattern matching and crowdsourcing content moderation can effectively reduce toxic content and fake news. Furthermore, previous studies on the effectiveness of content moderation in suppressing fake news and hate speech in decentralised systems failed to account for the effects of community-defined rules and norms. This work is focused on analysing the efficiency of DCMS in mitigating hate speech and fake news, especially in the times of COVID-19 and observing where it can be improved.

3 DATA COLLECTION AND PREPROCESSING

3.1 Dataset Preparation

3.1.1 Subreddit Selection. To conduct this research, we started with a list of over 2.4 million subreddits. Using the method discussed in [26] on the list of existing subreddits ranked by *frontpagemetrics*,[2] a subreddit ranking website, we infer if the discussion in these subreddits were connected to COVID-19.

3.1.2 Data Collection. We extracted 3172 COVID-19-related subreddits out of all the subreddits. This set of COVID-19-related subreddits covers various genres, e.g., news and discussions, humour, testimonials, and skill development. We further filter our data out of these 3172 COVID-19-related subreddits from January 2020–April 2021 using the *Pushshift API*.[3] Pushshift API allows searching submissions and comments of subreddits based on several parameters, including subreddit name, search string, range of dates, and author name. Table 1 shows the main characteristics of our dataset.

Table 1: Features of dataset gathered from Reddit using Pushshift API

Attributes	With Rules Mentioned	Without Rules Mentioned
Number of all Posts	1201451	604521
Number of subreddits	1012	2160
Number of moderated posts	92192	9196
Number of non-moderated posts	1007901	396674
Number of toxic posts	34609	10903
Subreddits with geographical focus	465	144
Subreddits without geographical focus	1695	868
Posts with geographical focus	230219	52147
Posts without geographical focus	593760	177356

The dataset includes title, text of a post, name of subreddit, date it was published, name of the submitter, "flair" (also called a "tag" on Reddit), number of comments on the submission, and any URLs mentioned in the submission, the number of subscribers, and, if the post was moderated, who moderated it.

3.2 Preprocessing

3.2.1 Subreddit Details Collection. We extracted the details of the "About" section of the subreddit and the rules mentioned for these subreddits. We first use web crawling available through the Selenium Web Driver package for this task[4]. We open each subreddit's URLs one by one and incorporate some loading time so that all contents of the page load correctly. After that, we use requests[5], and beautifulsoap[6] packages to scrape and process the contents from those web pages.

[1] https://github.com/pushshift/api

[2] https://frontpagemetrics.com/

[3] https://github.com/pushshift/api

[4] https://www.selenium.dev/projects/

[5] https://pypi.org/project/requests/

[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

3.2.2 Toxicity Score Computation. We use Google's Perspective API to compute the toxicity of text posts in the dataset. The API accepts the post's text as input and produces a toxicity value from 0 to 1. (0 means no toxicity and 1 means extremely toxic). The toxicity score threshold for a post depends on data and application. Each platform assign threshold based on the amount of data it receives. Studies [10, 19] have found 0.7 or higher as the optimal toxicity threshold. In contrast, some studies [32, 39] have used a threshold of 0.5 toxicity score to have a balanced dataset labelling.

For our research, we swept the toxicity threshold between 0.4 and 0.9 on our dataset and checked the resulting posts to see how many toxic posts were categorised as harmful. We check the Cohen's Kappa score [46] and got a value of 0.97 on our final selected threshold of 0.5. By assigning a threshold of 0.5, we obtained 45.5K harmful posts out of 1.8 Million posts in our data. 120K posts were moderated during the selected time duration from 3172 subreddits. Unmoderated 45.5K harmful posts are 38% of overall moderated posts from COVID-19-related subreddits.

3.3 Topic Modeling

Investigating popular topics is considered one of the best ways of studying the theme of the discussion on any platform [6, 20]. We use Latent Dirichlet Allocation (LDA) [6] to observe latent topics in subreddit content. LDA takes raw text, the number of topics, and a dictionary of words as the input and outputs the most significant topics with words from the raw data. We kept the number of latent output topics to 10 and iterated our algorithms 400 times on our dataset to achieve converged results.

The rest of the paper will analyse our dataset from COVID-19-related subreddits in the light of three research questions above-mentioned in the section 1.

4 RQ1: IS TOXIC CONTENT ALLOWED ACROSS SUBREDDITS, AND HOW STRICTLY ARE SUBREDDIT RULES ENFORCED?

4.1 Rules for User Posting on Subreddits.

Before observing the toxicity levels in different subreddits, we examine the rules and norms for posting in subreddits. We extract the rules from 3172 COVID-19-related subreddits. Of 3172 subreddits, 1012 have specific user activity rules, and 2160 subreddits mention no rules. We also explore which rules are commonly mentioned in the subreddit's "About" sections. Figure 1 presents rules mentioned in subreddit's "About" section and their co-occurrence with each other.

Overall, subreddits discourage the user from posting hate speech and malicious content. Figure 1 shows that each subreddit may have different rules according to the nature of a subreddit, but commonly, subreddits set rules to avoid malicious content. Around 80% of subreddits mention "Be Kind/Civil" and "Verified Information" in their rules which shows their desire to encourage their users to publish lower-toxicity and higher-quality content. "No medical advice", "Keep to the topic", and "No Repost" rules are also listed together on numerous subreddits, showing that moderators seek relevant, lawful content with minimal repetition.

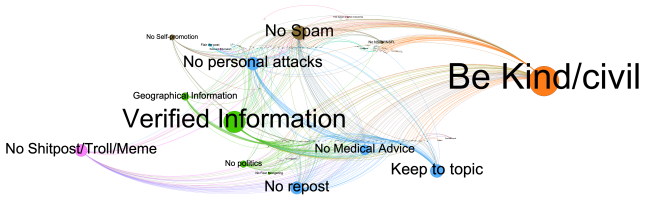


Figure 1: Co-occurrence of different rules over COVID-19 related subreddits. *Rules with larger font sizes are more commonly mentioned on different subreddits alongside other rules.*

These rules suggest that subreddit moderators are interested in following Reddit's content policy[7] while maintaining the high information quality for users.

4.2 User Posting Behavior over Subreddits

After analysing the rules for posting in different subreddits, we analyse the behaviour of users' posting over those subreddits in terms of their submissions. Figure 2 shows the distribution of different kinds of posts in 25 COVID-19-related subreddits with the highest user submission activity.

In Figure 2, bar lines represent the distribution of percentages of non-moderated non-toxic, toxic, and moderation posts, for each subreddit, given on the left y-axis, and the no. of submissions per subreddits are presented using the black line with counts on the right y-axis. We observe that generally, toxic content does not depend on the size of user activity *e.g. China_Flu* subreddit (COVID-19-related subreddit with highest user activity) contains 85.31% non-moderated non-toxic posts, 8.73% non-moderated toxic posts, and 5.95% moderated posts. Similarly, 25th subreddit with highest user activity, *CovidVaccinated*, also contains 83.17% non-moderated non-toxic posts, 7.95% non-moderated toxic posts, and 8.8% moderated posts.

This shows that toxicity in user content on subreddits is not dependent on user activity size. Also half of the toxic content is not being properly moderated for Reddit, which further highlight the issues with Reddit's current content moderation systems.

4.3 Topics of discussions in COVID-19-related subreddits

Figure 3 highlights the topics discussed in COVID-19-related subreddits in toxic and non-toxic content. These topics are manually tagged based on expert recommendations utilising Latent Dirichlet allocation findings (LDA). The LDA results are detailed in Appendix A. We discovered that toxic content in COVID-19-related subreddits is not restricted to vanilla categories like "Hate Speech", "Rumors", and "Fake News", but is also embedded throughout broader COVID-19-related topics of discussion. We obtained topics from more than 50% of the data in three major broad categories: "Precautions", "Case Reports", and "Symptoms", which include non-toxic and toxic themes.

[7] https://www.redditinc.com/policies/content-policy

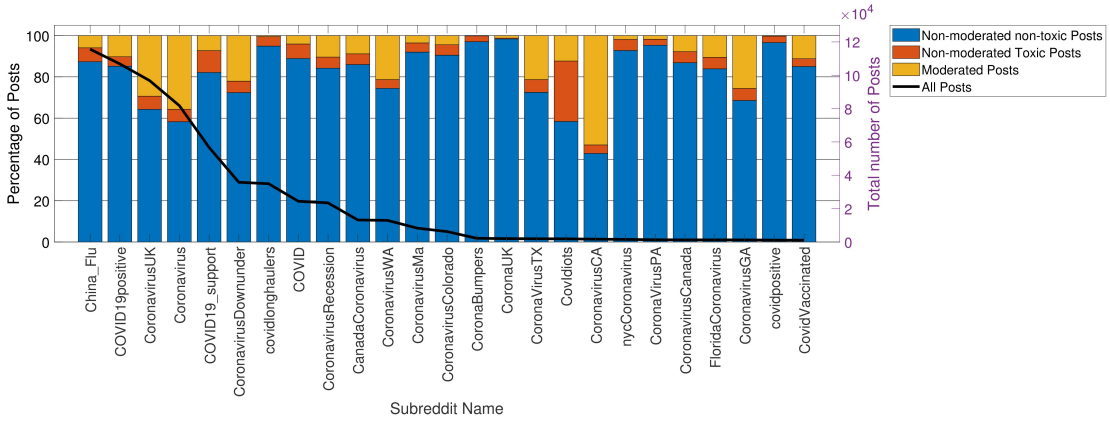


Figure 2: Distribution of Toxic, moderated and overall posts in COVID-19-related subreddits. ***In this figure, bar lines represent the Distribution of percentages of non-moderated non-toxic, toxic, and moderation posts in each subreddit. The black line represents the total number of posts in each subreddit, and its y-axis scale is mentioned on the right side of the figure.***

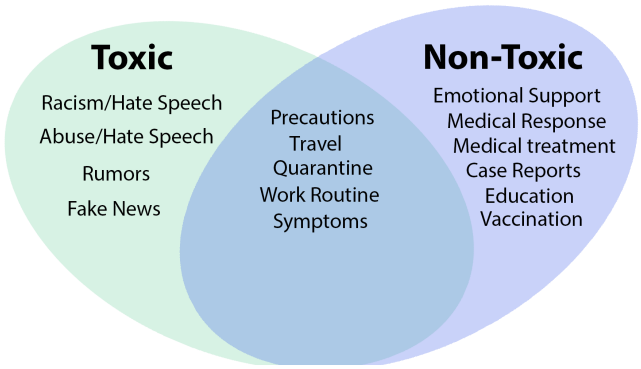


Figure 3: Top categories discussed in COVID-19-related subreddits and their overlap w.r.t. toxicity and non-toxicity. ***These categories are derived from latent topics extracted from COVID-19-related subreddits.***

These findings reveal that toxic data is absorbed throughout the discussion from general and vanilla toxic categories. and is widespread across all types of COVID-19-related subreddits.

4.4 Performance of Content moderation System on User Submissions

In earlier analysis, we observe that the content moderation system of Reddit is partially efficient in detecting and mitigating toxic content. In Figure 4, we observe the distribution of moderated content by different content moderation mechanisms in 25 COVID-19-related subreddits with the highest user submission activity. Bars show the contribution of moderated posts by each moderation mechanism in subreddits, whereas the black line shows the overall number of moderated posts in each subreddit.

By comparing results from Figure 2 and 4, We observe that content moderation activity on subreddits is dependent on the size of user activity *e.g.* subreddits with higher user activity have higher content moderation activity. In contrast, subreddits with lower user activity have lower content moderation activity. Secondly, human moderators seem more active in subreddits with higher user submission activity, whereas Automatic content moderation was predominant in subreddit with lower user activity. Reddit Moderation team was least active in overall moderation activity, which is understandable as automatic and volunteer moderators are the work horse of DCMS of reddit.

Typically, Auto-Moderator moderates the majority of the data. However, in the worst circumstances, Auto-Moderator may miss the harmful information, leaving it on the main page of the subreddit until the subreddit moderators can remove it.

The Automation part of the subreddit's content moderation system needs to be improved to prevent such issues. There are no built-in machine learning-based algorithmic content moderation tools available on Reddit. In order to flag or censor user-generated content before it is posted to the subreddit, subreddit moderators might employ tools based on third-party machine learning models.

5 RQ2: DOES THE SUBREDDIT'S REGIONAL FOCUS IMPACT TOXICITY AND FAKE NEWS DURING COVID-19?

We do a geographical breakdown of discussion on a subreddit dedicated to the topic of COVID-19 to provide an answer to this researh question. By reviewing each subreddit's "About Page," we were able to categorise the data from all 3,172 COVID-19-related subreddit according to whether or not they focus on localised content. Approximately 23% of the data in our dataset comes from subreddits with a geographical focus, while 77% of the data comes from non-geographic subreddits.

In Figure 5, we also observe the amount of user activity, moderation, and toxicity based on subreddits from different countries. It can be seen from the Figure 5 that USA, Canada, UK, Brazil, and

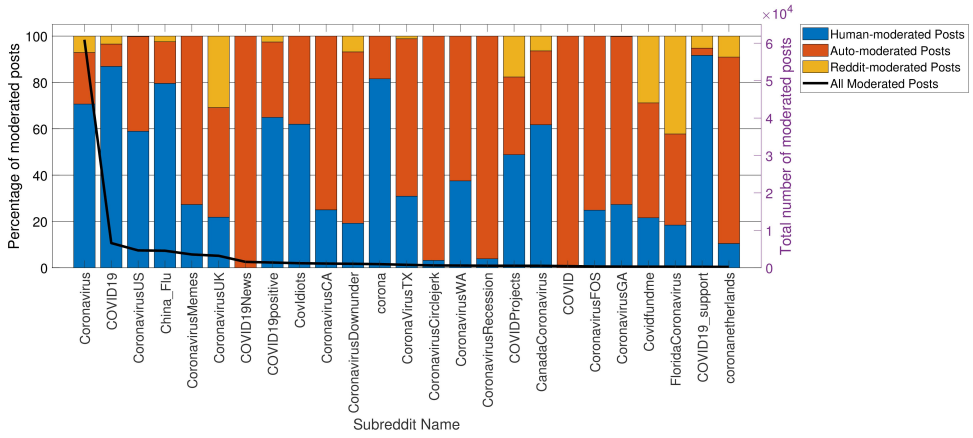


Figure 4: Distribution of moderated posts w.r.t. to different content moderation mechanisms available in COVID-19-related subreddits. ***Bars show the percentage of moderated posts by any method w.r.t. all moderated posts in each subreddit, whereas the black line shows the total number of moderated posts in each subreddit***

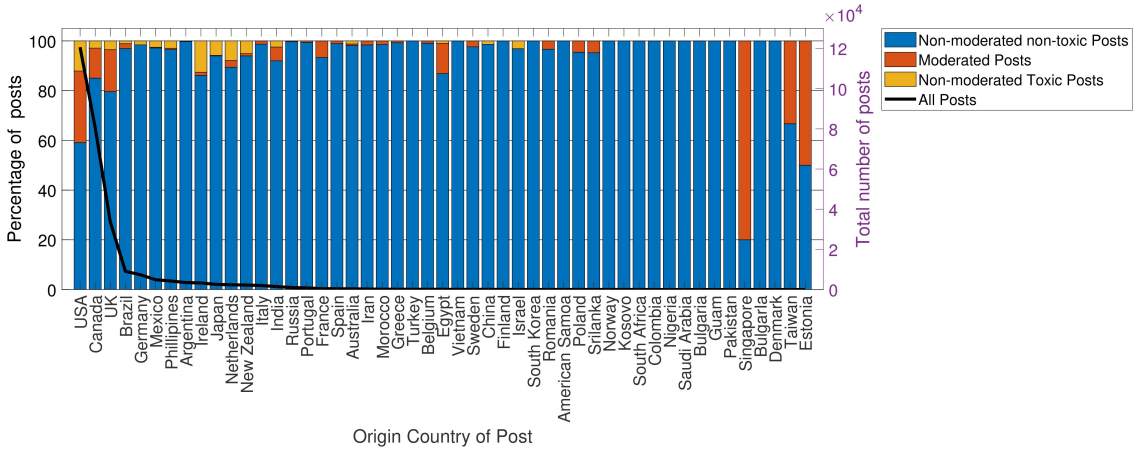


Figure 5: Distribution of non-moderated, moderated, and non-moderated toxic posts from COVID-19-related subreddits with geographical focus w.r.t. countries. ***Bars represents the percentage of non-moderated, moderated, and non-moderated toxic posts w.r.t. to overall data associated with a country whereas black line represents the total number of posts associated with a country.***

Germany were the top 5 countries on the list. It is worthy to note that these countries contribute the highest Reddit network traffic[8]. This implies that the amount of content coming from a country (whether toxic or non-toxic) does not depend upon the severity of COVID-19 in a country but the amount of overall Reddit network traffic. We also observe that countries like Singapore, Egypt, Taiwan, and Estonia show more content moderation over their subreddits. We assume this may result from a higher level of moderation which may lead users to move towards those subreddits with a higher user base and moderate moderation level.

Furthermore, we analyse the dissemination of non-fake and fake news in subreddits from different countries. In Figure 6, we observe a few fake news posts with a geographical focus on subreddits. We see a comparatively higher number of fake news posts in some countries, such as the Netherlands, Sri Lanka, Saudi Arabia, and Taiwan. However, these countries contributed less than 1% of data in the dataset, and there was not much user activity and content in those subreddits.

Instead of geographically specific subreddits, toxicity and fake news are mainly concentrated on COVID-19-related generic subreddits such as r/COVID, r/China_Flu, and r/CoronaVirus. Access to users from all over the world and the rapid dissemination of content is a primary factors for this effect.

[8]https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

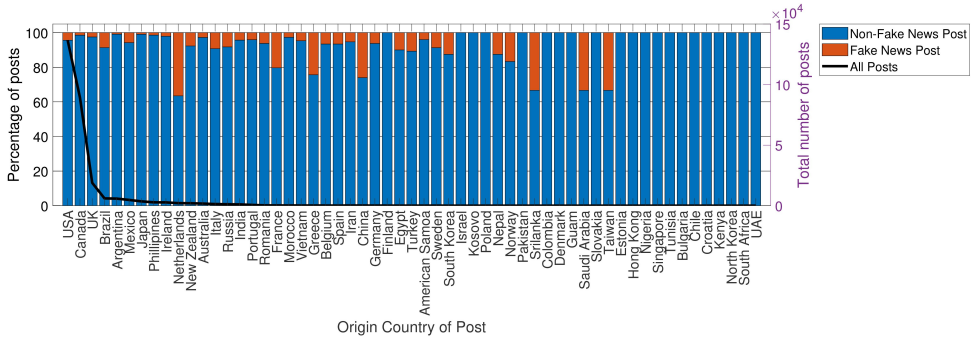


Figure 6: Distribution of Non-fake news posts and fake news posts from COVID-19 related subreddits with geographical focus w.r.t. countries.

6 RQ3: HOW CAN WE IMPROVE THE CONTENT MODERATION SYSTEMS OF REDDIT FOR DETECTING FAKE NEWS?

To answer this question, firstly, we detect URLs of fake news media outlets from all the URLs mentioned with user submissions.

6.1 Categorisation of News-related URLs

To perform this analysis, we annotate the URLs based on their categories using VirusTotal API[9]. After obtaining categories of URLs, we extract the URLs with categories of "News and Media". We get 36,123 unique URL categories as "News and Media". Finally, we check the authenticity of the URLs using two major fact-checking websites; Snopes[10], and PolitiFact[11] to tag URLs as Fake or True. These are two of the most comprehensive databases for categorising fake news outlets. We observe that subreddits with higher user activity are more prone to fake news propagation. Similarly, some subreddits which allow sarcasm, and advocate free speech, such as "r/CoronaVirusCircleJerk" and "r/CoronaVirusmeme" are also very prone to fake news.

Consequently, we explore the distribution of fake and non-fake URLs in user submissions. Figure 7 shows the distribution of news-related URLs across 25 COVID-19-related subreddits with the highest number of URLs.

We also analyse the distribution of URLs categorised as fake and non-fake in Figure 8. We observe that most non-fake news is contributed by famous mainstream media outlets such as CNN, BBC, and Reuters. Nevertheless, the highest percentage of fake news is spread by the imposter sites of famous mainstream media outlets. We find approximately 10,000 URLs impersonating different media outlets and presenting blogs spreading fake news.

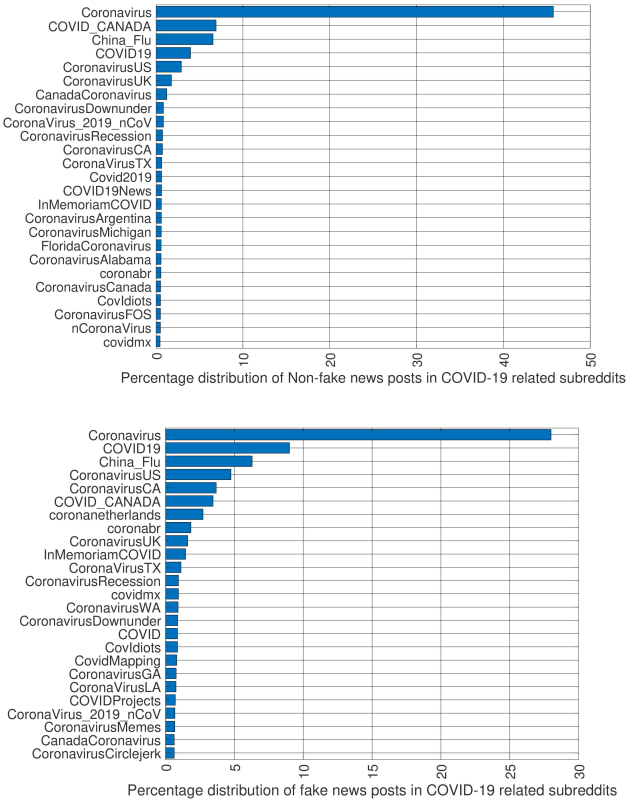


Figure 7: Distribution of percentage of Non-fake and fake news posts in COVID-19-related subreddits.

[9]https://developers.virustotal.com/reference/overview
[10]https://www.snopes.com/
[11]https://www.politifact.com/

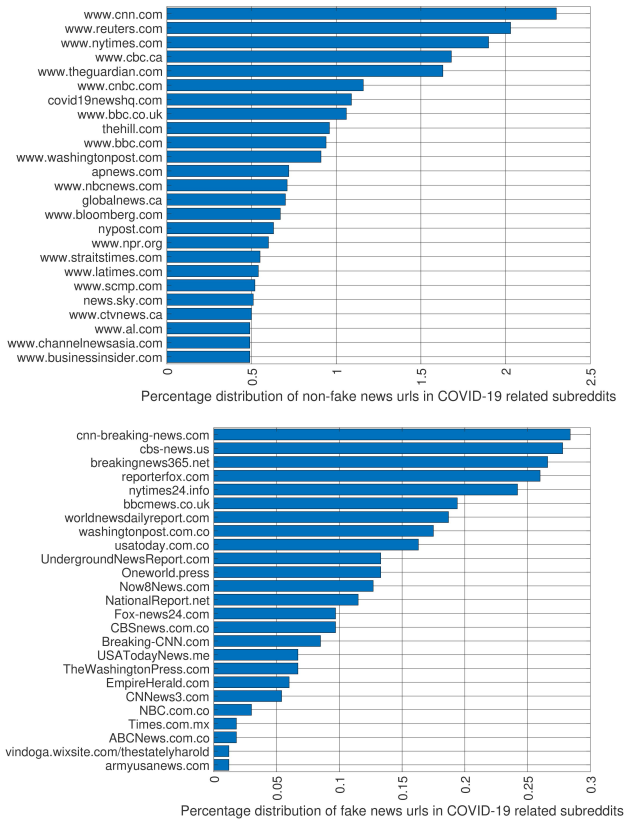


Figure 8: Distribution of percentage of Non-fake and fake news URLs in COVID-19-related subreddits. *In fake news related URLs, there is a high number of imposter websites of mainstream news outlets*

6.2 Improving the DCMS of Reddit using the Transfer Learning based Automatic Content Moderation System.

Automated content moderation methods reduce manual content moderation workload from human moderators. To contribute towards this goal, we use the Recurrent Convolutional Neural Network (RCNN) architecture proposed in [25] to screen fake news in user submissions on Reddit. The proposed RCNN is cable of capturing the contextual information with the recurrent structure and reconstruct text representation using the Convolutional Neural Network (CNN). This architecture leverages the advantages of both Recurrent Neural Network (RNN) and CNN techniques in a model [24].

To get started with this model, we initially train this model from Kaggle FakeNews Challenge dataset[12]. After the pre-training, we freeze the base model and retrain the classifier head using the 70% of out preprocessed dataset. The model achieves a test F1-score of 0.8725 on the final testset made from our dataset. Figure 9 shows the overall framework of transfer learning approach used in this regard.

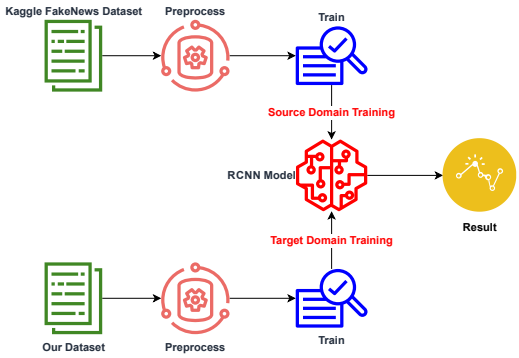


Figure 9: Proposed Transfer Learning based framework for Automatic Content Moderation Improvement for Detecting Fake news and Topic Content

Further details of model performance metrics are presented in Table 2.

Table 2: Performance metrics score for RCNN FakeNews Detection Model

	Training	Validation	Test
Precision	0.961	0.912	0.8831
Recall	0.975	0.9012	0.8623
F1-Score	0.967	0.9065	0.8725

We find that the RCNN-based algorithm is remarkably accurate in predicting fake news. One of the reasons for such a high number of fake news on Reddit is the unavailability of a more sophisticated AI power detection mechanism for fake news. Although, each subreddit has its primary content moderation mechanism, but it lags at detecting fake news automatically. As a result, it is incredibly challenging to identify fake news on these subreddits and requires extensive human monitoring.

The proposed RCNN based automatic content moderation requires computing resources which are not yet available on Reddit, therefore, moderators may need to use third-party tools to implement this solution in their subreddit. Deploying Machine Learning and Deep Learning automated models can reduce human labour significantly and mitigate the spread of fake news to a much greater extent.

7 CONCLUSION

In this paper, we studied 3172 COVID-19-related subreddits to learn more about the efficacy of Reddit's content moderation system during this pandemic. To further understand user posting behaviour and the usefulness of Reddit's DCMS worldwide and in particular locations, we analysed 1.8 million posts across these subreddits. We made an effort, mindful of Reddit's decentralised nature, to assess the content and efficacy of its moderation system. We shed light on the impact of subreddit-defined rules and norms on fake news

[12]https://www.kaggle.com/c/fake-news/

and toxicity across the platforms and the influence of subreddits' geographic focus on the efficacy of Reddit's content moderation mechanisms in reducing fake news and toxic posts. We also constructed an RCNN-based solution to showcase the effectiveness of the Deep Learning-based model for improving the automatic detection of fake news and harmful content, which can minimise human moderation efforts. Our research provides a window into content moderation research and suggests fundamental ways to improve the content moderation system for online social media platforms during global crises.

ACKNOWLEDGMENTS

This work is supported by EPSRC REPHRAIN "Moderation in Decentralised Social Networks" (DSNMod), AP4L: Adaptive PETs to Protect & emPower People during Life Transitions (EP/W032473/1), and Streamlining Social Decision Making for Improved Internet Standards (EP/S033564/1).

REFERENCES

[1] 2021. Content Moderation Techniques. (2021). https://vsd.ccs.neu.edu/content_moderation/techniques/
[2] 2022. 10 Reddit Stats You Should Know in 2022. (2022). https://www.oberlo.com/blog/reddit-statistics
[3] Jon Agley and Yunyu Xiao. 2021. Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health* 21, 1 (2021), 1–12.
[4] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*. 759–760.
[5] BBC. 2020. Coronavirus: The human cost of virus misinformation. (2020). https://www.bbc.com/news/stories-52731624
[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
[7] Robyn Caplan. 2018. Content or context moderation? (2018).
[8] Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. 2018. Manipulating visibility of political and apolitical threads on reddit via score boosting. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 184–190.
[9] Eshwar Chandrasekharan and Eric Gilbert. 2019. Hybrid approaches to detect comments violating macro norms on reddit. *arXiv preprint arXiv:1904.03596* (2019).
[10] Corel. 2021. Corel: Toxic Comments|Talk Documentation. (2021). https://docs.coralproject.net/talk/toxic-comments/
[11] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
[12] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.
[13] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.
[14] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[15] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. 2020. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e313.
[16] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*. 87–96.
[17] Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*. 6786–6794.
[18] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.
[19] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify. (2020).
[20] Waleed Iqbal, Junaid Qadir, Gareth Tyson, Adnan Noor Mian, Saeed-ul Hassan, and Jon Crowcroft. 2019. A bibliometric analysis of publications in computer networking research. *Scientometrics* 119, 2 (2019), 1121–1155.
[21] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene* 103, 4 (2020), 1621.
[22] R Tallal Javed, Mirza Elaaf Shuja, Muhammad Usama, Junaid Qadir, Waleed Iqbal, Gareth Tyson, Ignacio Castro, and Kiran Garimella. 2020. A First Look at COVID-19 Messages on WhatsAppin Pakistan. *arXiv preprint arXiv:2011.09145* (2020).
[23] R Tallal Javed, Muhammad Usama, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, and Kiran Garimella. 2022. A deep dive into COVID-19-related messages on WhatsApp in Pakistan. *Social Network Analysis and Mining* 12, 1 (2022), 1–16.
[24] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information* 10, 4 (2019), 150.
[25] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.
[26] Siddique Latif, Muhammad Usman, Sanaullah Manzoor, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, Adeel Razi, Maged N Kamel Boulos, Adrian Weller, et al. 2020. Leveraging Data Science To Combat COVID-19: A Comprehensive Review. (2020).
[27] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 369–380.
[28] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*. Springer, 51–56.
[29] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th acm conference on hypertext and social media*. 85–94.
[30] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.
[31] Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for fine-grained hate speech classification. *arXiv preprint arXiv:1809.00088* (2018).
[32] Elizabeth Reichert, Helen Qiu, and Jasmine Bayrooti. 2020. Reading between the demographic lines: Resolving sources of bias in toxicity classifiers. *arXiv preprint arXiv:2006.16402* (2020).
[33] Julio CS Reis, Philipe Melo, Kiran Garimella, and Fabrício Benevenuto. 2020. Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review* (2020).
[34] Reuters. 2020. Police arrest 110 people over lynching in western India. (2020). https://www.reuters.com/article/us-india-crime/police-arrest-110-people-over-lynching-in-western-india-idUSKBN2222HT?il=0
[35] Axel Rodríguez, Carlos Argueta, and Yi-Ling Chen. 2019. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 169–174.
[36] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.
[37] Lee K Royster. 2017. Fake News: Political Solutions to the Online Epidemic. *NCL Rev.* 96 (2017), 270.
[38] Diego Saez-Trumper. 2019. Online Disinformation and the Role of Wikipedia. *arXiv preprint arXiv:1910.12596* (2019).
[39] Jaydeb Sarker, Asif Kamal Turzo, and Amiangshu Bosu. 2020. A Benchmark Study of the Contemporary Toxicity Detectors on Software Engineering Interactions. In *2020 27th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 218–227.
[40] Vinay Setty and Erlend Rekve. 2020. Truth be Told: Fake News Detection Using User Reactions on Reddit. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3325–3328.
[41] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 1–42.

[42] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.
[43] Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. *arXiv preprint arXiv:2001.00623* (2020).
[44] Spandana Singh. 2019. Everything in moderation: An analysis of how Internet platforms are using artificial intelligence to moderate user-generated content. *New America* 22 (2019).
[45] Martino Trevisan, Luca Vassio, Idilio Drago, Marco Mellia, Fabricio Murai, Flavio Figueiredo, Ana Paula Couto da Silva, and Jussara M Almeida. 2019. Towards Understanding Political Interactions on Instagram. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 247–251.
[46] Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*. IEEE, 1–8.
[47] Vox. 2020. How the 5G coronavirus conspiracy theory went from fringe to mainstream. (2020). https://www.vox.com/recode/2020/4/24/21231085/coronavirus-5g-conspiracy-theory-covid-facebook-youtube
[48] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.
[49] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
[50] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666* (2019).
[51] Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226* (2020).
[52] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* 57, 2 (2020), 102025.
[53] Xinyi Zhou and Reza Zafarani. 2018. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315* (2018).
[54] Fabiana Zollo and Walter Quattrociocchi. 2018. Misinformation spreading on Facebook. In *Complex Spreading Phenomena in Social Systems*. Springer, 177–196.

A APPENDIX

Table 3: LDA based topic modeling to determine popular topics in toxic and non-toxic content in COVID-19-related subreddits. ***Content Percentage column shows the percentage of content available in specific category.***

Category	Latent Topics	Content percentage
Precautions	("mask", "wear", "hand", "wash", "touch", "sanitize", "protect", air", "disinfect", "glove")**(Non-Toxic)** ("mask", "hand", "cough", "walk", "store", "face", "wear", "touch", "wash", "toilet")**(Non-Toxic)** ("mask", "wear", "protect", "spread", shit, "distance", "travel", "dumb","transmission", "social")**(Toxic)**	Non-Toxic(94.07%) Toxic(5.93%)
Fake News	("people", "virus", "online", "help", "corona", "savior", "die", "know", "fake", "direct")**(Toxic)**	Non-Toxic(0%) Toxic(100%)
Racism/Hate Speech	("f**k", "people", "chinese", "racist", "hate", "shit", "lie", "kill", "coronavirus", "china")**(Toxic)**	Non-Toxic(0%) Toxic(100%)
Medical treatment	("hospital", "doctor", "heart", "blood", "patient", "breath", "lung", "help", "medical", "oxygen")**(Non-Toxic)** ("oxygen", "ventilator", "patient", "machine", "pressure", "bed", "lung", "provide", "supplies", "airway")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)
Vaccination	("vaccine", "effect", "dose", "first", "shot", "second", "arm", "develop", "immune", "inject")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)
Travel	("travel", "quarantine", "people", "close", "home", "flight", "health", "coronavirus", "countries", "restrict")**(Non-Toxic)** ("people", "situation", "countries", "government", "news", "tourist", "shit", "transmission", "travel", "blood")**(Toxic)**	Non-Toxic(65.25%) Toxic(34.75%)
Abuse/Hate Speech	("bullshit", "shit", "see", "f**k", "news", "ban", "people", "remove", "inform", "discuss")**(Toxic)**	Non-Toxic(0%) Toxic(100%)
Quarantine	("smell", "stay", "taste", "home", "weeks", "cough", "chest", "fever", "quarantine", "throat")**(Non-Toxic)** ("test", "positive", "home", "work", "sick", "go", "hospital", "f**k", "quarantine", "symptom")**(Toxic)**	Non-Toxic(78.58%) Toxic(21.42%)
Work Routine	("work", "job", "pay", "time", "back", "call", "home", "downsize", "companies", "month")**(Non-Toxic)** ("job", "business", "pay", "money", "die", "worker", "f**k", "companies", "home", "essential")**(Toxic)**	Non-Toxic(71.6%) Toxic(28.36%)
Symptoms	("smell", "pain", "taste", "breath", "felt", "cough", "chest", "fever", "headache", "throat")**(Non-Toxic)** ("test", "positive", "cough", "week", "fever", "result", "start", "feel", "sick", "covid")**(Non-Toxic)** ("feel", "symptom", "f**k", "test", "pain", "smell", "breath", "cough", "fever", "bitch")**(Toxic)**	Non-Toxic(80.59%) Toxic(19.41%)
Rumors	("people", "report", "spread", "death", "t**ts", "f**k", "less", "infect", "number", "government")**(Toxic)**	Non-Toxic(0%) Toxic(100%)
Case Reports	("case", "death", "new", "report", "total", "positive", "rate", "test", "hospital", "confirm", "increase")**(Non-Toxic)** ("test", "case", "positive", "patient", "result", "hospital", "symptom", "health", "coronavirus", "infect")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)
Medical Respose	("government", "medical", "health", "pandemic", "care", "hospital", "public", "support", "crisis", "response")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)
Education	("school", "student", "study", "online", class", "participate", "teacher", "research", "university", "health")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)
Emotional Support	("people", "feel", "better", "live", "work", "home", "families", "support", "alone", "friend")**(Non-Toxic)**	Non-Toxic(100%) Toxic(0%)