Text Mining of COVID-19 Discussions on Reddit

Syed Saad Naseem, Dhruv Kumar, Mohammad S. Parsa and Lukasz Golab
University of Waterloo, Ontario, Canada
{ssnaseem,dhruv.kumar,mohammad.parsa,lgolab}@uwaterloo.ca

***Abstract*—We identify discussion communities on the Reddit social curation platform that frequently mention COVID-19-related terms, and we apply the Non-negative Matrix Factorization (NMF) topic modelling algorithm to extract topics discussed by these communities. In addition to forums dedicated to COVID-19, we find such discussions on general question-answering forums, on a forum for teenagers, and on advice forums, specifically medical, mental health, relationship, and legal advice. Topic modelling results reveal the hardships of life during the pandemic and the effects of social distancing: fear of contracting the virus, job safety and security, loss of motivation and productivity, troubles with online activities, and dealing with those who do not follow social distancing rules.**

***Index Terms*—Topic modelling, Social media mining, COVID-19**

I. Introduction

The COVID-19 pandemic has become a public health emergency and a critical socioeconomic issue worldwide. It is changing the way we live and do business. These changes have happened quickly, as a result of social anxiety about the pandemic and public policies designed to limit its spread, and are likely to be long-lasting if not irreversible. Thus, it is crucial to study public reaction to the COVID-19 crisis.

Social media have been recognized as rich sources of data about public opinion on a variety of topics, including, recently, the COVID-19 crisis. For example, there has been recent work on analyzing mental health discussions on social media during the pandemic [1], sentiment and emotion analysis of pandemic-related tweets [2], and extracting questions related to the pandemic from social media and search engine logs [3].

We contribute to this effort with a text mining study of COVID-19 discussions on the Reddit (reddit.com) social curation platform. Reddit consists of over 128,000 user-created discussion communities called *subreddits*. Each subreddit has a name that starts with "r/" and describes the corresponding discussion community; e.g., r/mentalhealth is a subreddit to discuss and share information about mental health. Within a subreddit, a user may write a new post to initiate a discussion, or write a comment in response to an existing post or an existing comment.

We leverage the organization of Reddit into interest-based communities with descriptive names to answer two questions:

- *Which online communities have been impacted by the COVID-19 pandemic?* To answer this question, we identify subreddits whose posts frequently mention words related to the pandemic.
- *How are these communities impacted?* To answer this question, we apply the Non-negative Matrix Factorization (NMF) [4] topic-modelling algorithm to extract common discussion topics.

II. Data and Methods

We used the publicly-available pushshift.io API (accessed via the psaw library) to search posts and comments on Reddit. We downloaded all posts, across all subreddits, created between January 1, 2020 and May 31, 2020, that mentioned at least one of the following four terms related to the pandemic: "corona", "covid", "quarantine", or "pandemic". We performed case-insensitive substring search, meaning that the term "corona" also matches "coronavirus", while the term "covid" also matches "COVID19" and "covid-19". Our search terms are a subset of terms used in previous work on mining social media discussions related to the virus [5], with Twitter-specific terms such as "#stayhomechallenge" removed since hashtags are not used frequently on Reddit.

The total number of downloaded posts matching at least one search term is 729,597. Fig. 1 shows the number of such posts per day, indicating a peak at the end of March. We also downloaded all the comments made on these posts, for a total of 6,753,269 comments.

To answer the first question, which is to identify communities impacted by the pandemic, we sorted the subreddits by the number of posts matching at least one of the four search terms. We report the top-50 subreddits in Table I.

To answer the second question, which is to characterize the discussions on the subreddits affected by the pandemic, we applied topic modelling to selected subreddits from the top-50 list identified above. We only used posts for topic modelling, not comments. Posts are meant to initiate discussions and thus they are good indicators of the topics being discussed.

Before topic modelling, we performed standard text pre-processing. Using the NLTK[1] library in Python, we tokenized the text, removed stopwords, converted all letters to lower case, and lemmatized the words. Next, we vectorized the tokens by converting each post to its TF-IDF vector.

For each selected subreddit (or a group of related subreddits), we then ran the Non-negative Matrix Factorization (NMF) topic modeling algorithm on the TF-IDF matrix. NMF clusters the posts into topics and returns a list of terms that represent each topic. Each such term has a score indicating how well the term represents the corresponding topic. We used three pieces of information to interpret each topic. First, we selected the top-10 highest-scoring representative terms for

[1] https://www.nltk.org/

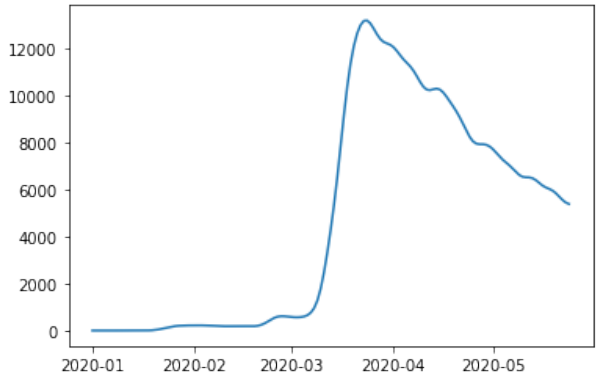

Fig. 1. Daily number of posts matching at least one COVID-19-related term.

each topic. Second, we extracted the five most frequent word bigrams and trigrams occurring within the posts assigned to each topic. Third, for each topic, we manually examined a sample of 5% of posts with the most comments.

NMF requires the number of topics as input. To find an appropriate number of topics, we ran NMF to produce between 5 to 100 topics and calculated the coherence [6] of topic descriptors for each topic. We chose the number of topics with the highest coherence score, ranging from 5 to 30 topics for the selected subreddits.

III. Results

A. Which Communities Discuss COVID19?

Table I shows the top 50 subreddits in terms of the number of posts matching at least one search term. We also report the total number of substring matches within these posts, and the number of matches for each search term. We categorize notable top-50 subreddits into the following groups:

- Six communities dedicated to the pandemic: r/Coronavirus, r/China_Flu, r/CoronavirusUS, r/CoronavirusUK, r/COVID19positive, and r/COVID19.
- "Automated" subreddits such as r/autotldr (which summarizes long posts from various subreddits), r/removalbot (comments removed by users or moderators from various subreddits), and r/TalkativePeople (copies of posts made by the most active Reddit users). These subreddits will not be analyzed further.
- General question-answering communities such as r/AskReddit, r/Advice, and r/NoStupidQuestions.
- Topic-specific question-answering communities for medical (r/AskDocs), legal, and relationship advice.
- Topic-specific discussion communities dedicated to teenagers, relationships (including r/r4r - a personal ad forum), mental health, politics, unemployment, and computer gaming.
- Communities discussing conspiracy theories, and communities allowing users to speak their mind and express alternative viewpoints (r/unpopularopinion).

We note that variations of the terms 'corona' and 'covid' are frequently used by communities dedicated to the pandemic. On the other hand, the term 'quarantine' is often used in other communities, suggesting discussions related to the consequences of social distancing.

B. Topic Modeling

We now present the results of topic modelling of large subreddits from the above categories in Tables II through VIII. We do not analyze mental health subreddits as these have been studied in prior work [1]. For brevity, we report the five largest topics per subreddit. For each such topic, we show the percentage of posts belonging to it, the representative terms, and frequent word n-grams.

Six subreddits dedicated to COVID-19 (Table II): The largest topic describes experiences of people who were infected and have recovered. The next two topics discuss outbreaks in various places as well as affected celebrities and politicians (such as Boris Johnson). The last two topics share information about policies such as lockdowns and travel bans, and symptoms, respectively.

AskReddit (Table III): In this general question-answering subreddit, the two most common topics are related to changes in people's personal and professional lives, such as what to do to stay positive and productive, or how to maintain a daily routine. Topic 3 includes questions about online gaming during the pandemic. Topics 4 and 5 include questions about things to do during the quarantine, such as learning new skills, picking up new hobbies, watching TV, and other ways to spend time. We found similar topics on r/showerthoughts, r/Advice and r/NoStupidQuestions.

Relationship advice (Table IV): The largest topic includes various relationship troubles amplified by the pandemic. The next two topics discuss problems with starting or maintaining relationships during the lockdown. Topic 4 includes discussions about dealing with partners' or family members' attitudes towards social distancing. Topic 5 includes discussions of people wishing to reunite with their ex-partners.

Teenagers (Table V): frequent topics among teenagers include how to spend their time during the quarantine (Topic 1), loneliness (Topic 2), online education (Topic 3), missing one's friends (Topic 4), and reaching out to others to talk due to boredom (Topic 5).

Legal advice (Table VI): The most frequent topic is about filing claims for medical bills or unpaid salaries. The next topic includes people inquiring about the validity of rental leases and lease cancellation procedures. Topic 3 includes questions about unemployment benefits, whereas Topic 4 includes concerns about losing one's job after testing positive for the virus. Finally, topic 5 includes discussions about protecting oneself in the workplace.

AskDocs (Table VII): The largest topic contains posts describing symptoms and asking for advice. Topic 2 includes questions about one's risk of catching the virus. Topics 3 and 4 includes people having other medical problems but being scared to go to the hospital during the pandemic. Topic 5 includes advice sought by those who tested positive for the virus.

TABLE I
TOP-50 SUBREDDITS WITH THE MOST COVID-19 RELATED POSTS.

Subreddit	Matching posts	Total matches	corona	covid	quarantine	pandemic
Coronavirus	41679	53139	35450	11291	3078	3320
AskReddit	31749	37352	2237	2645	20828	11642
TalkativePeople	20854	27041	3534	2318	11774	9415
China_Flu	17041	39524	21981	10760	3909	2874
relationship_advice	11635	19285	1599	2424	11661	3601
teenagers	10981	13889	899	479	11194	1317
Showerthoughts	10337	12342	944	778	5492	5128
CoronavirusUS	7392	13843	6867	4520	1218	1238
Advice	5622	9228	901	1269	4598	2460
gonewild	5603	5633	6	9	5531	87
r4r	4291	5208	211	203	4149	645
autotldr	4180	32625	18608	4671	2411	6935
unpopularopinion	4018	5305	519	481	1965	2340
legaladvice	3455	6497	579	1749	1564	2605
funny	3302	3637	155	123	2989	370
CoronavirusUK	3233	7383	3415	3180	353	435
MakeNewFriendsHere	3090	3598	110	59	3126	303
relationships	2939	4439	294	461	2802	882
conspiracy	2932	13216	5399	2927	1370	3520
NoStupidQuestions	2929	3712	315	271	1681	1445
depression	2910	4227	298	310	2651	968
COVID19	2868	3854	1380	2125	105	244
COVID19positive	2674	4805	618	3704	391	92
CasualConversation	2583	3424	188	255	2148	833
memes	2489	2586	59	25	2192	310
TrueOffMyChest	2456	4698	679	786	1653	1580
dating_advice	2263	3507	299	267	2348	593
NoFap	2223	2991	137	95	2533	226
removalbot	2195	6477	1617	655	1027	3178
snapchat	2142	2203	11	4	2149	39
gaming	2059	2379	95	66	1949	269
PokemonGoFriends	1891	2176	41	24	2022	89
videos	1876	2194	182	98	1557	357
Needafriend	1772	2103	60	62	1778	203
AskDocs	1745	2799	225	664	999	911
raisedbynarcissists	1724	2947	294	373	1420	860
pics	1488	1603	52	40	1165	346
mentalhealth	1395	2293	200	272	1220	601
personalfinance	1369	2022	177	352	260	1233
selfie	1330	1348	4	3	1310	31
Vent	1315	2169	199	291	1062	617
phr4r	1314	1504	11	42	1358	93
Anxiety	1311	2080	230	285	876	689
politics	1304	2180	739	94	146	1201
rant	1282	2366	314	360	868	824
Unemployment	1244	2413	158	685	244	1326
wallstreetbets	1196	3124	1081	529	619	895
trees	1152	1504	92	149	1004	259
BreakUps	1134	1747	138	161	1090	358
ForeverAloneDating	1132	1424	78	84	1057	205

Unpopular opinion (Table VIII): The largest topic includes discussions arguing that people who refuse to follow social distancing guidelines should catch the virus to understand its seriousness. Topic 2 is related to high school education during the pandemic, specifically that all classes should be online, which is believed to be an unpopular opinion. Topic 3 represents criticisms of U.S. quarantine policies, with many poster arguing for stricter rules. In topic 4, users are discussing enjoyable aspects of the quarantine, such as working from home, and therefore not having to commute to work and spending more time with their children. Notably, this was the only topic encountered in our analysis that had an overall positive sentiment. Finally, topic 5 mainly discusses foreign governments' roles in the pandemic.

IV. DISCUSSION AND RELATED WORK

Previous work on mining COVID-19 discussion on Reddit focused on specific topics such as mental health [1], [7] or gender differences in word use [8]. Most of the previous work used Twitter instead. Specifically, the work on topic modelling [2], [9], [10] revealed that Twitter users tend to discuss news (such as the number of new cases and deaths), and government decisions and preventive measures related to the pandemic (such as physical distancing and hand washing).

While previous work focused on specific topics of interest, we provide a holistic view of pandemic-related discussions

TABLE II
TOP 5 TOPICS ON COVID-19 SUBREDDITS (TOTAL 25 TOPICS).

1 : covid19, daily, discussion, help, fight, treatment, disease, study, cure, response (9.9 percent) 'daily discussion', 'covid19 daily', 'covid19 daily discussion', 'due covid19', 'covid19 case'
2 : test, positive, tested, negative, antibody, testing, result, johnson, boris, kit (5.1 percent) 'test positive', 'positive coronavirus', 'tested positive', 'test positive coronavirus', 'test kit'
3 : new, york, total, city, raising, case, zealand, cuomo, dead, jersey (4.9 percent) 'new case', 'new york', 'new coronavirus', 'case coronavirus', 'new death'
4 : u, cdc, testing, million, response, military, government, citizen, travel, tell (3.9 percent) 'coronavirus case', 'coronavirus death', 'u government', 'u citizen', 'confirmed case"
5 : symptom, day, fever, cough, tested, got, chest, feel, throat, mild (3.5 percent) 'sore throat', 'feel like', 'shortness breath', 'tested positive', 'dry cough'

TABLE III
TOP 5 TOPICS ON ASKREDDIT (TOTAL 5 TOPICS).

1 : pandemic, covid19, coronavirus, life, change, world, end, positive, current, like (56.3 percent) 'covid19 pandemic', 'coronavirus pandemic', 'pandemic end', 'due pandemic', 'pandemic affected'
2 : people, reddit, home, like, live, work, affected, relationship, worker, dealing (18.1 percent) 'people reddit', 'people quarantine', 'reddit quarantine', 'worker reddit', 'coronavirus pandemic'
3 : game, play, video, fun, friend, online, playing, free, gamers, mobile (11.7 percent) 'game play', 'video game', 'play quarantine', 'friend quarantine', 'game play quarantine'
4 : learn, skill, new, hobby, picked, learning, home, pick, useful, cool (9.3 percent) 'learn quarantine', 'skill learn', 'new hobby', 'new skill', 'hobby picked'
5 : best, worst, happened, movie, watch, birthday, far, home, spend, make (4.6 percent) 'best quarantine', 'quarantine best', 'best worst', 'time quarantine', 'watch quarantine'

TABLE IV
TOP 5 RELATIONSHIP ADVICE TOPICS (TOTAL 10 TOPICS).

1 : like, time, feel, want, really, know, day, said, told, say (15.8 percent) 'feel like', 'make feel', 'felt like', 'even though', 'spend time'
2 : boyfriend, pandemic, social, covid, want, upset, distancing, covid19, 19f, 20f (13.6 percent) 'feel like', 'long distance', 'best friend', 'even though', 'social distancing'
3 : relationship, distance, long, year, love, feeling, end, month, feel, want (11.5 percent) 'long distance', 'feel like', 'distance relationship', 'long distance relationship', 'month ago'
4 : mom, dad, sister, brother, family, parent, mother, father, house, know (10.4 percent) 'feel like', 'even though', 'since quarantine', 'take care', 'mental health'
5 : ex, broke, gf, year, contact, ago, breakup, new, month, 20m (9.5 percent) 'feel like', 'due quarantine', 'since quarantine', 'quarantine making', 'break quarantine'

TABLE V
TOP 5 TOPICS ON TEENAGERS (TOTAL 20 TOPICS).

1 : time, people, know, really, help, need, going, make, good, want (32.9 percent) 'since quarantine', 'feel like', 'video game', 'since quarantine', 'need help'
2 : like, feel, got, really, feeling, girl, lonely, shit, horny, know (10.4 percent) 'feel like', 'quarantine got', 'quarantine like','since quarantine', 'even though'
3 : school, work, online, class, year, teacher, going, pandemic, grade, high (9.1 percent) 'high school', 'miss school', 'online school', 'online class', 'school closed'
4 : friend, best, new, really, group, make, looking, want, miss, need (7.4 percent) 'best friend', 'friend quarantine', 'nothing special', 'feel like', 'nothing special'
5 : bored, chat, got, quarantine, af, looking, need, discord, talk, dm (5.9 percent) 'bored quarantine', 'got bored', 'quarantine bored', 'quarantine got', 'really bored'

TABLE VI
TOP 5 LEGAL ADVICE TOPICS (TOTAL 10 TOPICS).

1 : house, car, know, u, time, told, said, like, money, going (32.1 percent) 'even though', 'due pandemic', 'year old', 'feel like','legal action'
2 : lease, landlord, rent, apartment, roommate, month, tenant, pandemic, pay, signed (13.9 percent) 'pay rent', 'security deposit', 'month rent', 'due pandemic', 'signed lease'
3 : unemployment, job, benefit, file, eligible, quit, pandemic, claim, laid, filed (8.1 percent) 'file unemployment', 'unemployment benefit', 'due pandemic', 'eligible unemployment', 'back work'
4 : covid19, employer, wedding, testing, pandemic, exposed, venue, tested, policy, positive (7.1 percent) 'due covid19', 'covid19 pandemic', 'social distancing', 'positive covid19', 'tested positive'
5 : quarantine, self, job, travel, symptom, doctor, home, sick, workplace, force (6.5 percent) 'self quarantine', 'stay home', 'day quarantine', 'sick leave', 'work home'

TABLE VII
TOP 5 TOPICS ON ASK DOCS (TOTAL 30 TOPICS).

1 : throat, sore, tonsil, strep, nose, allergy, swallow, white, mouth, swallowing (8.4 percent) 'sore throat', 'feel like', 'grade fever', 'back throat', 'strep throat'
2 : risk, virus, people, work, coronavirus, know, home, live, case, mask (5.8 percent) 'year old', 'immune system', 'high risk', 'corona virus', 'stay home'
3 : pain, sharp, right, left, lower, area, dull, chest, hurt, shoulder (5.1 percent) 'feel like', 'back pain', 'lower back', 'sharp pain', 'left side'
4 : foot, toe, ankle, walk, nail, shoe, walking, heel, leg, running (4 percent) 'right foot', 'year old', 'feel like', 'go doctor', 'look like'
5 : symptom, mild, headache, tested, experiencing, covid19, fatigue, feeling, severe, anxiety (2.4 percent) 'year old', 'get tested', 'feel like', 'came back', 'sore throat'

on Reddit. At a high level, Table I reveals that Reddit users, including teenagers, are seeking general advice, as well as specific relationship, medical (including mental health) and

TABLE VIII
TOP 5 UNPOPULAR OPINION TOPICS (TOTAL 20 TOPICS).

1 : people, stupid, making, make, want, protesting, sick, complain, die, allowed (12.5 percent) 'many people', 'social medium', 'quarantine people', 'see people', 'people say'
2 : time, going, life, day, better, year, make, school, know, kid (12.4 percent) 'many people', 'feel like', 'spend time', 'high school', 'long time'
3 : pandemic, global, current, covid, care, trump, selfish, real, hope, worse (10.5 percent) 'global pandemic', 'current pandemic', 'covid pandemic', 'pandemic selfish', 'using pandemic',
4 : quarantine, enjoy, bored, protesting, awesome, shouldnt, loving, introvert, birthday, protest (9.2 percent) 'quarantine measure', 'birthday quarantine', 'protesting quarantine', 'enjoy quarantine', 'whole quarantine',
5 : china, world, government, country, u, chinese, responsible, current, blame, global (6.9 percent) 'global pandemic', 'united state', 'around world', 'chinese government', 'climate change'

legal advice.

Furthermore, while Twitter topic modeling results show that most pandemic-related discussions are related to news, government policies, and public health interventions, we find that COVID-19 discussions and questions on Reddit focus on the impact of the pandemic on people's lives. This allows us to report new insight into public well-being and coping strategies, summarized below.

As seen on r/AskReddit and r/teenagers, there is a need for coping strategies such as maintaining a daily routine to stay positive, productive and healthy, both at home and remotely at school and at work. Some people report picking up new hobbies, while others spend their free time on online gaming (note the presence of pandemic-related terms on r/gaming). We also observed conflicting opinions about in-person education during the pandemic, with some posters believing that making all high school classes online is an unpopular opinion (Topic 2 on r/unpopularopinion).

Relationship advice is sought for starting and maintaining relationships during the pandemic (as expected, this topic is popular with the young Reddit user base), but also for dealing with friends and family members who do not follow social distancing rules (r/relationship_advice and, more broadly, Topic 1 on r/unpopularopinion). This finding should be of interest to counsellors as they may want to prepare solutions to this problem.

In addition to living arrangements and social benefits, legal discussions show concern for one's job security if diagnosed with COVID-19 and job safety to avoid being infected at work (r/legaladvice). This highlights the importance of workplace policies during the pandemic and into the near future.

Medical advice (r/AskDocs) is sought for potential symptoms and treatment; there appears to be more public information about preventive measures such as wearing a mask and washing hands, but less on what to do if infected. This suggests the need for clear public health guidelines on both fronts. Some people report fear of visiting a doctor or a hospital for other medical problems, which underscores the importance of being able to speak with a doctor online.

Finally, we note that about 10 percent of posts on r/unpopularopinion are positive, and point out the advantages of working from home.

V. CONCLUSIONS

In this paper, we studied the Reddit social curation platform using text mining tools to understand discussions related to the COVID-19 pandemic. We exploited the organization of Reddit into interest-based communities with descriptive names to identify those which have been affected by the virus and extract their discussion topics.

One limitation of this study is that that the findings represent the views of Reddit users, 49.7 percent of whom are from the United Stated[2] and 58 percent are between the ages of 18 and 29[3]. An interesting direction for future work is to study COVID-19 discussions generated by other populations on other platforms.

REFERENCES

[1] L. Biester, K. Matton, and J. Rajendran, "Quantifying the effects of COVID-19 on mental health support forums," in *NLP COVID-19 Workshop at ACL Conf.*, 2020. [Online]. Available: https://openreview.net/pdf?id=DAiyXps5q1T

[2] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, X. Liu, and T. Zhu, "Twitter discussions and emotions about COVID-19 pandemic: a machine learning approach," 2020. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/2005/2005.12830.pdf

[3] J. Wei, C. Huang, S. Vosoughi, and J. Wei, "What Are People Asking About COVID-19? A Question Classification Dataset." in *NLP COVID-19 Workshop at ACL Conf.*, 2020. [Online]. Available: https://openreview.net/pdf?id=qd51R0JNLl

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[5] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set," *JMIR Public Health Surveill*, vol. 6, no. 2, p. e19273, May 2020. [Online]. Available: http://publichealth.jmir.org/2020/2/e19273/

[6] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[7] J. T. Wolohan, "Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic," in *NLP COVID-19 Workshop at ACL Conf.*, 2020. [Online]. Available: https://openreview.net/pdf?id=2f70OXlGQMd

[8] J. Aggarwal, E. Rabinovich, and S. Stevenson, "Exploration of gender differences in COVID-19 discourse on reddit," in *NLP COVID-19 Workshop at ACL Conf.*, 2020. [Online]. Available: https://openreview.net/pdf?id=mlmwkAdIeK

[9] H. Jang, E. Rempel, G. Carenini, and N. Janjua, "Exploratory analysis of COVID-19 related tweets in north america to inform public health institutes," 2020. [Online]. Available: https://arxiv.org/pdf/2007.02452.pdf

[10] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," 2020. [Online]. Available: https://arxiv.org/pdf/2007.02304.pdf

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

[3]https://www.statista.com/statistics/517218/reddit-user-distribution-usa-age/