4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15022

Exploring Redditors' Topics with Natural Language Processing

Yilang Zhao
Department of Curriculum and Instruction, University of Wisconsin–Madison, USA

Abstract

This paper examines how people in Reddit develop topics across threads in a given subreddit and how discussions concentrate on the topic in given threads with natural language processing (NLP) methods. By implementing an LDA topic model and TF-IDF models, this paper discovers people's aggregated concerns are related to real-world issues and their discussions are concentrative considering the topics they discuss.

Keywords: *Reddit, online discussion community, online discussion topics, natural language processing, LDA, TF-IDF*

1. Introduction and Theoretical Backgrounds

Reddit is a popular online discussion community among the young generation today, in which a lot of online discussions take place. Since topics of those discussions significantly vary, the Reddit community is divided into user-created subreddits that only allow posts with certain themes. Subreddit is regarded as real communities in virtual space in which people have their rules, languages, rules, etc. (Medvedev et al., 2017). Thus, it is worth investigating how people build up their community by exploring what they care about beyond the topic of a specific thread in a subreddit channel.

Since people's discourses are likely to focus on topic-relevant content in a given thread, scrutinizing the topics that come from a couple of threads can be constructive to understand the core interest of a subreddit community. In addition, as subreddits are self-organized spaces, its information spread yields to a more dynamic way compared to direct follower social media such as Facebook or Twitter (Medvedev et al., 2017). That is to say, the posts on Reddit may vary a lot, even in one single thread with a fixed topic. Fang et al. (2016) discover that jokes and funny comments are much more underpredicted than controversial comments, which they think can to some extent develop the discussion. Most jokes and funny comments, however, only entertain the discussion participants and may not be helpful with the discussion, yet more concentrated discussions may promote more constructive conversations on Reddit. Therefore, we are also interested in the level of concentration in a specific subreddit channel. The contribution of this research is from two aspects: one is about the community itself, and the other is about its practical application. Better understanding the topics in a subreddit channel can improve the user experience and further elevate the engagement level of participating in the discussion in online communities. As for the potential practical contribution, Park and Conway (2017) discover that public health relevant discussion on Reddit can predict the trending public interest in some public health issues and serve as an information source for certain user groups. Thus, enhancing the quality of the discussion will make some Reddit discussions more reliable information providers. Therefore, in this study, our research questions are RQ 1. what do people concern about across topics in a subreddit? and RQ 2. how do discussions in a subreddit concentrate on topics?

2. Methods

2.1. Data Collection

To answer the research questions, we collected comments from a subreddit, r/science, which is the eighth most popular subreddit channel and has over 25 million subscribers. It is an online community where people can share the latest science news and discuss. The threads we target are the top 10 hot threads of the year 2020 (see Appendix), which can be ranked by

Reddit built-in filter. These 10 topics have 20,273 comments in total when we used PRAW (Python Reddit API Wrapper) to retrieve on December 10, 2020. We stored retrieved data in a pandas data frame in which each comment is a row and has a tag of whether it is a submission (the first comment of a thread) or a following comment.

2.2. Data Analysis

As for our data analysis methods, we have two ways to analyze the data collected. In response to RQ 1, we used LDA (Latent Dirichlet Allocation) modeling. LDA is a probabilistic model that uses the Bayesian model to infer topics with their underlying probabilities and provides a representation of the document (Blei et al., 2003). We mixed all the comments and inputted them into the LDA model to discover the top 10 most probabilistic topics across the threads. To answer RQ2, we have done a two-step analysis. The first step is to implement TF-IDF modeling. TF-IDF measures the term-frequency that is the times that a term occurs in the given document and the inversed document frequency which indicates how common and rare that term is across all documents (Luhn, 1957; Jones, 1972). TF-IDF ensures that common words such as “this” are filtered out when we are looking at key information of a document as their IDF value is 0 which makes their TF-IDF value is 0. A high TF-IDF value indicates that the term is important to the given document and possibly represents key information of that document. We did TF-IDF modeling for each thread in a tri-gram way (three words as a term) to inspect if the key information extracted by TF-IDF modeling aligned with the topic. The second step is to investigate the relevance between the keywords of a thread and the title of that thread. A uni-gram (one word as a term) TF-IDF model was utilized to produce each thread's keywords. We then count the keywords that appeared in the title of a post and implement a regression model to examine the relationship between the number of keywords (explanatory variable) and the counts of its occurrences in the title (response variable).

3. Results

3.1. LDA Modeling

The below figure (see Figure 1) demonstrates the results of our LDA modeling. Each line indicates a probable topic that is consisted of some words and their probabilities. For example, the first line is the topic that contains the words “people,” “money,” “government,” “better,” “language,” “federal,” “asset,” and “month” and their probabilities, 0.019, 0.016, 0.011, 0.011, 0.011, 0.008, 0.008, and 0.008. As all the words are tokenized in the LDA topic model, the results of LDA topic modeling yield to the researcher’s interpretations. In other words, LDA, as a probabilistic model, will not provide a certain conclusion of what the exact topics are in the document but offer terms with probabilities for inference.

```
(0, '0.019*"people" + 0.016*"money" + 0.011*"government" + 0.011*"better" + 0.011*"language" + 0.008*"federal" + 0.008*"asset" + 0.008*"month"')
(1, '0.012*"virus" + 0.012*"people" + 0.009*"think" + 0.009*"marijuana" + 0.009*"neuron" + 0.006*"right" + 0.006*"better" + 0.006*"decision"')
(2, '0.045*"remove" + 0.024*"percent" + 0.010*"people" + 0.010*"adult" + 0.010*"medical" + 0.008*"problem" + 0.008*"american" + 0.008*"deductible"')
(3, '0.021*"would" + 0.013*"people" + 0.013*"economy" + 0.013*"cheap" + 0.009*"government" + 0.009*"money" + 0.009*"could" + 0.009*"billion"')
(4, '0.027*"would" + 0.011*"something" + 0.008*"placebo" + 0.006*"testing" + 0.006*"antibody" + 0.006*"people" + 0.006*"pretty" + 0.006*"assume"')
(5, '0.014*"people" + 0.013*"would" + 0.011*"state" + 0.010*"immune" + 0.010*"pathogen" + 0.008*"large" + 0.008*"school" + 0.008*"teacher"')
(6, '0.021*"neuron" + 0.017*"people" + 0.017*"leadership" + 0.012*"brain" + 0.010*"elephant" + 0.007*"involve" + 0.007*"years" + 0.007*"intelligence"')
(7, '0.016*"pressure" + 0.010*"would" + 0.010*"think" + 0.007*"people" + 0.007*"brain" + 0.007*"something" + 0.007*"human" + 0.007*"create"')
(8, '0.019*"study" + 0.010*"save" + 0.010*"patient" + 0.010*"someone" + 0.010*"better" + 0.007*"private" + 0.007*"emergency" + 0.007*"design"')
(9, '0.024*"people" + 0.009*"still" + 0.009*"vaccine" + 0.009*"community" + 0.009*"delete" + 0.009*"physical" + 0.007*"voice" + 0.007*"country"')
```

Figure 1. LDA Topic Modeling Results

3.2. Thread TF-IDF

The below figures (see Figure 2 and Figure 3) show the results of tri-gram TF-IDF modeling for each thread. From the model for each topic, we can find some key information of that thread based on the tri-grams. The TF-IDF value indicates the relevance of the term to a document. In this context, a higher TF-IDF score means that the given term is important to the comments in that thread. Therefore, the main topic of a thread can be inferred from the TF-IDF model. However, in each model, the amount of key tri-grams varies so that some topics are easier to infer based on those key terms, e.g., topic 1, whereas others like topic 7 are much more difficult to make an inference.

Topic 1	TF-IDF
juvenile incarceration place	0.198829
determine sentence natural	0.198829
state typically pay	0.198829
typically pay prison	0.198829
county led stark	0.198829
natural experiment whereby	0.198829
drop incarceration suggest	0.198829
stark drop incarceration	0.198829
experiment whereby cost	0.198829
whereby cost burden	0.198829
led stark drop	0.198829
prison county determine	0.198829
us state typically	0.198829
incarceration place county	0.198829
pay prison county	0.198829
place county led	0.198829
sentence natural experiment	0.198829
incarceration suggest mass	0.198829
county determine sentence	0.188275
cost burden juvenile	0.188275
burden juvenile incarceration	0.188275
suggest mass incarceration	0.180786
incarceration us part	0.180786
us part due	0.174978
mass incarceration us	0.174978
part due misalign	0.170232
due misalign incentive	0.162744

Topic 2	TF-IDF
immune response these	0.262191
these result represent	0.262191
find safe welltolerated	0.262191
result represent important	0.262191
vaccine find safe	0.262191
induce rapid immune	0.262191
safe welltolerated induce	0.262191
represent important milestone	0.262191
welltolerated induce rapid	0.262191
covid19 vaccine find	0.262191
trial covid19 vaccine	0.262191
response these result	0.262191
human trial covid19	0.252087
first human trial	0.244249
rapid immune response	0.227742

Topic 3	TF-IDF
lancet team yale	0.266959
team yale epidemiologist	0.266959
find medicare would	0.266959
yale epidemiologist find	0.266959
new study lancet	0.266959
would save 68000	0.266959
study lancet team	0.266959
epidemiologist find medicare	0.266959
annually well 450	0.254312
well 450 billion	0.254312
medicare would save	0.254312
life annually well	0.254312
68000 life annually	0.245339
save 68000 life	0.238379
450 billion cost	0.232692

Topic 4	TF-IDF
content mind manifestation	0.242472
manifestation higher intelligence	0.242472
know know ponder	0.242472
humans higher mammal	0.242472
find crow know	0.242472
thought long believe	0.242472
research find crow	0.242472
long believe sole	0.242472
know ponder content	0.242472
analytical thought long	0.242472
mind manifestation higher	0.242472
province humans higher	0.242472
intelligence analytical thought	0.230166
believe sole province	0.230166
higher intelligence analytical	0.230166
ponder content mind	0.221435
sole province humans	0.221435
crow know know	0.193626

Topic 5	TF-IDF
cancer cell find	0.201752
find new australian	0.201752
tumour growth mouse	0.201752
cell find new	0.201752
also found venom	0.190307
component combine exist	0.190307
new australian research	0.190307
study also found	0.190307
exist chemotherapy drug	0.190307
drug extremely efficient	0.190307
reducing tumour growth	0.190307
chemotherapy drug extremely	0.190307
main component combine	0.190307
found venom main	0.190307
venom main component	0.190307
combine exist chemotherapy	0.190307
research study also	0.190307
australian research study	0.190307
efficient reducing tumour	0.190307
extremely efficient reducing	0.190307
kill aggressive hardtotreat	0.182186
found rapidly kill	0.182186
hardtotreat breast cancer	0.182186
rapidly kill aggressive	0.182186
venom honeybee found	0.182186
honeybee found rapidly	0.182186
aggressive hardtotreat breast	0.182186
breast cancer cell	0.159296

Figure 2. TF-IDF Tri-gram Keywords (topics 1-5)

Topic 6	TF-IDF
higher legalization however	0.210659
even higher legalization	0.210659
state legalize recreational	0.200088
recreational marijuana use	0.200088
jump even higher	0.200088
use among college	0.200088
trend upward years	0.200088
years state legalize	0.200088
use jump even	0.200088
college student trend	0.200088
marijuana use jump	0.200088
marijuana use among	0.200088
upward years state	0.200088
student trend upward	0.200088
state marijuana legal	0.200088
however student show	0.192588
legalization however student	0.192588
among college student	0.192588
legalize recreational marijuana	0.192588
greater drop binge	0.186770
drinking peer state	0.186770
student show greater	0.186770
show greater drop	0.186770
peer state marijuana	0.186770
binge drinking peer	0.186770
drop binge drinking	0.182017

Topic 7	TF-IDF
70 percent painting	0.453996
percent painting wind	0.429847
death 70 percent	0.412713
bird death 70	0.399423
painting wind turbine	0.388565
wind turbine blade	0.358141

Topic 8	TF-IDF
disappearance coronavirus swiftly	0.215333
coronavirus swiftly serum	0.215333
level leading significant	0.215333
ace2 hrsace2 disappearance	0.215333
hrsace2 disappearance coronavirus	0.215333
cytokine level leading	0.215333
serum nasal cavity	0.203471
inflammatory cytokine level	0.203471
treat human recombinant	0.203471
lung reduction inflammatory	0.203471
cavity lung reduction	0.203471
successfully treat human	0.203471
severe covid19 patient	0.203471
first severe covid19	0.203471
leading significant clinical	0.203471
patient successfully treat	0.203471
nasal cavity lung	0.203471
covid19 patient successfully	0.203471
swiftly serum nasal	0.203471
reduction inflammatory cytokine	0.203471
soluble ace2 hrsace2	0.195055
significant clinical improvement	0.195055
recombinant soluble ace2	0.188527
human recombinant soluble	0.174777

Topic 9	TF-IDF
like center disease	0.265973
response coronavirus pandemic	0.265973
organization like center	0.265973
control prevention rather	0.265973
rather president lead	0.265973
country response coronavirus	0.265973
adult look scientific	0.265973
prevention rather president	0.265973
scientific organization like	0.265973
president lead country	0.255860
look scientific organization	0.255860
disease control prevention	0.248016
lead country response	0.248016
center disease control	0.236188
us adult look	0.231494

Topic 10	TF-IDF
elimination within week	0.197347
goldstandard could lead	0.197347
inexpensive rapid covid19	0.197347
within week even	0.197347
sensitive goldstandard could	0.197347
rapid covid19 test	0.197347
even test le	0.197347
week even test	0.197347
weekly inexpensive rapid	0.197347
bars retail school	0.187607
orders without shutting	0.187607
without shutting restaurant	0.187607
shutting restaurant bars	0.187607
could lead personalized	0.187607
stayathome orders without	0.187607
lead personalized stayathome	0.187607
test le sensitive	0.180696
toward elimination within	0.180696
population weekly inexpensive	0.180696
restaurant bars retail	0.180696
test would drive	0.180696
covid19 test would	0.180696
personalized stayathome orders	0.180696
le sensitive goldstandard	0.175335
drive virus toward	0.170955
half population weekly	0.170955
would drive virus	0.170955
virus toward elimination	0.170955
testing half population	0.164044

Figure 3. TF-IDF Tri-gram Keywords (topics 6 –10)

3.3. Keywords and title relevance

The below table (see Table 1) shows the keywords extracted from the uni-gram TF-IDF modeling of each topic. The column *TF-IDF (>0)* indicates the number of keywords that have a TF-IDF value greater than 0 in a given thread. *InTitleCount* column refers to the counts of those keywords that are also in the submission (the first comment of the thread). *Percentage* is the ratio of those keywords in the title over the total keywords.

Our regression model tests whether the number of keywords from TF-IDF modeling in a thread predicts its occurrences in the title. The results of the regression indicates that the predictor, the number of keywords, explains 62.2% of the variance ($R^2 = .622$, $F(1,8) = 13.14$, $p < 0.01$). The result reveals that the number of keywords in a given thread statistically significantly predict the occurrences of the keywords in the title of that thread.

Table 1. Keywords and Title Analysis Results

Topic	TF-IDF (>0)	InTitleCount	NotInTitleCount	Percentage
1	25	8	17	32%
2	17	9	8	53%
3	17	8	9	47%
4	18	6	12	33%
5	28	11	17	39%
6	23	12	11	52%
7	8	3	5	38%
8	26	8	18	31%
9	17	9	8	53%
10	30	13	17	43%

4. Discussion

There are two findings from the LDA modeling analysis. The first finding is that although the overarching topics of the selected threads are different, there are aggregated concerns across those threads. In our LDA modeling results, terms that are related to people, government, and public health indicate that across these 10 threads, Redditors concern about the impact brought by the COVID-19 pandemic. This can also be inferred from the probabilistic topics 2,3,4,5,8 and 9 (see Figure 1). Redditors mention the terms of virus, vaccine, testing, placebo, economy, etc., which is reasonable as the pandemic affected everyone's life in the year 2020. The second finding is LDA generated topics may not cover all the threads. For instance, topic 7 has a theme of the bird's death and contains 1,503 rows, which takes up 7.4% of the total comments. However, the top 10 LDA topics have no evidence of this topic. Therefore, although probability-based LDA topics have the limitation of failing to represent all the concerns, it can be inferred that people's overarching concern in the subreddit channel r/science is the COVID-19 pandemic and the impacts on the public health system. Thus, in a subreddit channel that distributes the latest news, Redditors plausibly concern about what is trending in the real world.

To better understand every thread, it is necessary to inspect each of them. From Figure 2 and Figure 3, key phrases produced by tri-gram TF-IDF models in each topic make it feasible to infer what the main topic of a given thread is. For example, from the TF-IDF key terms of topic 2 in Figure 2, it can be deduced that this thread is about the COVID-19 vaccine and its trial on humans because terms like safety, immune response are key to people's discussion in this thread based on the TF-IDF values. Therefore, to answer RQ2, Redditors' discussion in the subreddit r/science is relatively concentrated, which can also be validated in our third analysis.

The third analysis is consisted of uni-gram TF-IDF key terms generating and a follow-up linear regression modeling process. As demonstrated in the previous section, the counts of TF-IDF of each topic can predict the number of matched terms in the title of the tread in a statistically significant way. This result indicates that Redditors' comments, which are the data source of those TF-IDF terms, are relevant to the discussion topic as they frequently refer to the terms in the title of each post. This might be because the subreddit channel selected is somewhat serious so that there are fewer distracting comments. In some other subreddits such as r/todayilearned, although people are learning fun facts, they tend to respond to others in a more entertaining way and engage in discussion with more memes and jokes.

Our study has two limitations. The first is that the discussion style and language features may significantly vary across subreddits. There might even not be any subreddit-wise concerns, and an extreme example can even be that a subreddit channel might be created for amusement, and people never concentrate on any specific topics there. Therefore, whether we should care about what people are concerning depends on the purposes that people create and join a subreddit. Another is about the sample size of our data. The PRAW only allows us to send 1 request per second, which prevents us from retrieving a large volume of posts. Therefore, the generalizability of our results needs further study to test.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Fang, H., Cheng, H., & Ostendorf, M. (2016, November). Learning latent local conversation modes for predicting comment endorsement in online discussions. In Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media (pp. 55-64).

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

Medvedev, A. N., Lambiotte, R., & Delvenne, J. C. (2017, June). The anatomy of Reddit: An overview of academic research. In Dynamics on and of Complex Networks (pp. 183-204). Springer, Cham.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21. https://dx.doi.org/10.1108/eb026526

Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. In AMIA Annual Symposium Proceedings (Vol. 2017, p. 1362). American Medical Informatics Association.

Appendix

2020 top 10 threads in r/science:

1.(social science)
https://www.reddit.com/r/science/comments/k1ofcu/in_the_us_states_typically_pay_for_prison_while/

2.(medicine)
https://www.reddit.com/r/science/comments/gp2hdt/the_first_human_trial_of_a_covid19_vaccine_finds/

3.(health)
https://www.reddit.com/r/science/comments/f4998k/a_new_study_in_the_lancet_by_a_team_of_yale/

4.(psychology)
https://www.reddit.com/r/science/comments/izbj3r/research_finds_that_crows_know_what_they_know_and/

5.(cancer)
https://www.reddit.com/r/science/comments/ikivxq/venom_from_honeybees_has_been_found_to_rapidly/

6.(health)
https://www.reddit.com/r/science/comments/eoomwz/marijuana_use_among_college_students_has_been/

7.(environment)
https://www.reddit.com/r/science/comments/igmtvw/bird_deaths_down_70_percent_after_painting_wind/

8.(medicine)
https://www.reddit.com/r/science/comments/jp3w7w/the_first_severe_covid19_patient_successfully/

9. (social science)
https://www.reddit.com/r/science/comments/gl1nvf/us_adults_look_to_scientific_organizations_like/

10. (epidemiology)
https://www.reddit.com/r/science/comments/jy8knh/testing_half_the_population_weekly_with/