Learning Programming in Social Media: An NLP-powered Reddit Study

Conference Paper · September 2022

DOI: 10.1109/TransAI54797.2022.00015

2 authors:

Yang Liu
North Carolina Agricultural and Technical State University
14 PUBLICATIONS **62** CITATIONS
SEE PROFILE

Mohd Anwar
North Carolina Agricultural and Technical State University
126 PUBLICATIONS **1,640** CITATIONS
SEE PROFILE

PREPRINT Accepted by 2022 IEEE International Conference on Transdisciplinary AI

Learning Programming in Social Media : An NLP-powered Reddit Study

Yang Liu
Computer Science
North Carolina A&T State University
Greensboro, NC, USA
yliu2@aggies.ncat.edu

Mohd Anwar
Computer Science
North Carolina A&T State University
Greensboro, NC, USA
manwar@ncat.edu

***Abstract*—Reddit is a popular social media platform that is used for various purposes. In this study, we explore how Reddit has been utilized by online learners to learn programming. Our research questions are: 1) how is Reddit generally used as an alternative resource for learning programming? 2) what programming topics are popularly discussed in Reddit communities? 3) what are the concerns and experiences of these learners in subreddit communities? Using Natural Language Processing techniques (e.g., topic modeling), we have studied the four largest programming-related subreddit communities, namely *r/learnprogramming*, *r/javascript*, *r/Python*, and *r/learnpython* during January 1st, 2019, through April 30th, 2022. Our study finds that learners discussed various programming constructs, participated in code sharing activities, and expressed their feelings/emotions about learning programming and writing working codes. Generally, we have observed that learners' participations during summertime are focused on mastering general programming concepts. However, during fall and spring, the learners seek help on very specific programming problems and get engaged in lot more code sharing. Outside of the learning-related discussions, at the beginning of the pandemic, the learners' concerns on the pandemic were expressed. Overall, Reddit platform is used to supplement textbooks for learning programming.**

Keywords—Reddit; Social Media; Programming; Natural Language Processing; Online Learning

I. Introduction

With the rapid development of information technology, all kinds of social media platforms emerge in an endless stream. People take to social media to share their life experiences, gain new experiences, and express their concerns on many issues, including public health [1, 2], politics [3], society [4], education [5], etc. Today, billions of people around the world use social media to communicate their opinions on various topics, products, and services, which provide rich datasets for text mining and social network analysis.

Traditional education and learning approaches are teacher-centered, with face-to-face instructions in schools or institutions. While traditional learning offers good communication platform with in-person interactions, the conventional one-size-fits-all teaching method lacks personalization, and learners cannot initiate their learning. More and more learners are taking advantage of alternative resources for learning. Online learning has been attracting learners' attention in the age of social media. The number of learners attending online courses continues to grow, especially during the lockdown period of the COVID-19 pandemic. According to the *Admissionly.com*, over 100 million students have enrolled in open online courses in the world and 63% of US high school students use e-learning tools every day [6]. As of April 2022, Reddit ranks as the No. 6 most visited website in America and No. 15 in global internet engagement, according to Reddit statistics [7]. There are 430 million people use Reddit monthly.

Online learning systems have become very popular for STEM education, especially for learning computer programs (e.g., Coursera online learning platform). However, it is unclear how popular social media such as Reddit is used for learning programming languages, specifically during the pandemic. Therefore, we explore the four largest subreddit communities to understand the use of social media for learning programming. In this exploratory study, the following questions were explored on how online learners have utilized Reddit to learn programming.

1) How is Reddit generally used as an alternative resource for learning programming?

2) What programming topics are popularly discussed in subreddit communities?

3) What are the concerns and experiences of these learners in subreddit communities?

The organization of the rest of the paper is as follows: Section 2 presents the related work section that surveys literature related to this study. Section 3 introduces techniques of data collection, data preprocessing, and topic modeling. Section 4 presents the results and discussions. Finally, Section 5 provides the limitations and future work of this study.

II. Related Work

Both the rise of social media and e-learning communities have significantly impacted online learning. To investigate the impact of using social media to learn English during the COVID-19 pandemic, Muftah [8] surveyed 166 university undergraduates. More than 84% students were interested in using social media as an educational tool and felt it helped them get more useful information. Learning through social media is becoming a cornerstone of lifelong learning [9]. Akcaoglu [10] applied digital methods to analyze tweets of Georgia K-12 school personnel discussing educational technology. On Reddit platform, a subreddit is a focused community in which

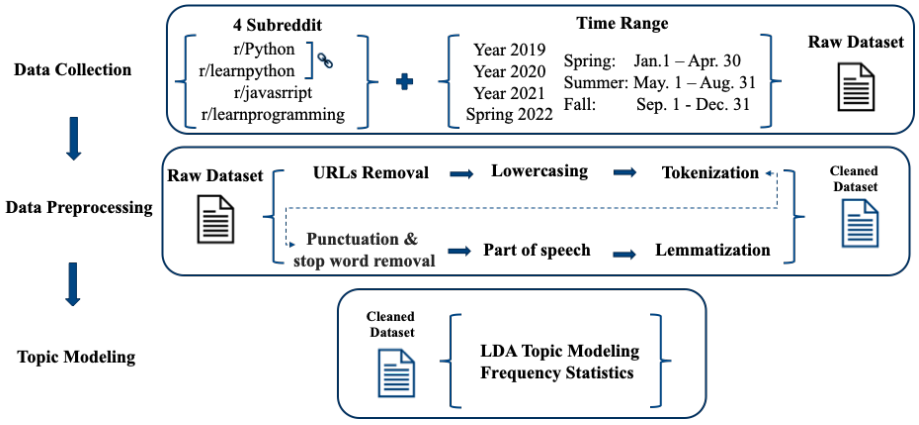

Fig. 1. Methodology Workflow.

discussions are organized into user-created areas of interest. Learners are increasingly using Reddit as an online learning environment. Exploring how to share knowledge, ideas, and resources in open online learning forums, Haythornthwaite et al. [9] studied learning practices in four "Ask" subreddit communities (AskHistorians, Ask_Politics, Askscience, Askacademia) to develop a coding schema for informal learning. Valle et al. [11] examined two online communities on Reddit to study how learning ties are formed and sustained among users.

Using Natural language processing (NLP) techniques to analyze social media data is becoming increasingly widespread. NLP methods are very useful to extract information from multitudinous social media data. NLP-based approaches to mining social media text include sentiment classification [12], topic modeling [1], and named-entity recognition [2]. This paper aims to explore how Reddit is used as an alternative resource to learn programming and what programming related topics are most popular in subreddit communities.

III. Data and Methodology

The methodological workflow of our study is shown in Figure 1. The workflow has three steps: (i) Reddit text collection (two criteria: most popular subreddits for programming and within time range), (ii) data preprocessing (removal of URLs, lowercasing, tokenization, punctuation and stop word removal, part-of-speech tagging, and lemmatization), (iii) topic modeling and frequency statistics.

The Reddit dataset contains all posts and related comments published in the four largest programming-related subreddit communities, namely *r/learnprogramming*, *r/javascript*, *r/Python*, and *r/learnpython* from January 1st, 2019 through April 30th, 2022. Users post the programming-related news and questions and ask for general advice about learning programming in these subreddits. The extensive Reddit Application Program Interface (API) allows direct access to posts in subreddits. We use PRAW [13], a Python Reddit API Wrapper, to collect text from subreddits. We combine the data from r/Python and r/learnpyton as r/python for data preprocessing and topic modeling. The basic information of largest four programming-related subreddits are as shown in Table 1.

Data preprocessing is the process of converting raw data into a useful and efficient format [2]. We performed the preprocessing steps as shown in Figure 1: URLs removal, lowercasing, tokenization, removal of punctuation and stop word, part of speech tagging, and lemmatization.

Topic modeling is an unsupervised learning method used for grouping documents into different topics in which the topics are comprised of high probability keywords. This study chose the most common modeling approach, Latent Dirichlet Allocation (LDA) topic modeling. LDA topic modeling is an unsupervised learning method to analyze and cluster data into similar groups. The two main assumptions that guide LDA are that each document is a mixture of topics, and each topic is a mixture of words. LDA randomly assigns each word to a topic and then computes two probabilities to update the words in each topic over multiple iterations. Then, the documents are grouped into different topics, and the topics are comprised of high-probability keywords. LDA topic modeling helps qualitatively identify a number of topics, and we use them to quantitatively investigate our corpus of text. In this study, we are using *gensim.models.ldamodel.LdaModel* to perform LDA, and we chose 5 clusters of topics for the convenience of preliminary analysis, each topic contains the top 10 keywords. For reproducibility of results, we run the algorithm a fix number of times by setting *random_state* to 40.

TABLE I. Distribution of the number of posts from Spring 2019 to Spring 2022 for each subreddits.

Subreddit	Year 2019			Year 2020			Year 2021			Spring 2022	Member
	Spring 2019	Summer 2019	Fall 2019	Spring 2020	Summer 2020	Fall 2020	Spring 2021	Summer 2021	Fall 2021	Spring 2022	
r/learnprogramming	14,472	14,242	15,465	16,250	18,892	16,444	17,200	22,354	22,890	8,735	3,043,099
r/javascript	6,918	7,690	6,301	6,894	6,869	5,935	5,896	13,114	5,775	2,056	2,041,911
r/Python	7,559	9,527	9,406	11,010	11,175	11,913	9,935	10,799	9,150	3,694	954,103
r/learnpython	11,534	12,318	11,035	11,955	13,093	12,270	11,290	13,114	13,804	4,594	616,818
Total	**443,567**										6,655,931

TABLE II. TOP THREE WORDS FOR EACH TOPIC FROM THE WORD CLOUDS.

	r/javascript					r/learnprogramming					python (r/Python + r/learnpython)				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Spring 2019	function, variable, type	react, website, build	code, language, write	link, jquery, code	const, shortcut,canvas	run, linux, window	learn, program, code	go, get, job	number, game, player	c, program, course	code, python, file	learn, book, program	print, result, return	version, install, package	object, data, class
Summer 2019	standard, support, browser	try, learn, something	code, function, example	language, library, project	array, loop, function	pointer, list, memory	learn, job, degree	function, code, example	language, course, book	computer, science, math	instance, test, code	advice, range, comment	class, list, function	package, version, edit	learn, python, program
Fall 2019	key, need, object	code, write, type	component, browser, react	function, return, language	catch, problem, question	learn, start, job	code, program, language	table, row, number	editor, code, comment	compant job, degree	web, data, automate	code, file, function	range, return, true	learn, python, code	class, print, recommend
Spring 2020	great, people, year	run, code, request	learn, language, javascript	point, change, package	function, react, code	program, **corona,** language,	learn, program, course	file, git, database	company, job, experience	write, code, problem	code, line, return	python, code, learn	data, email, version	class, loop, function	panda, range, thread
Summer 2020	module, const, dependency	error, test, code	function, object, variable	people, great, support	look, test, code	run, code, call	learn, program, project	program, course, language	school, degree, job	tell, help, people	function, class, list	learn, python, project	print, math, tutorial	beginner, code, game	run, module, file
Fall 2020	user, support, browser	people, react, project	window, test, file	code, type, function	look, page, github	learn, python, course	web, software, development	learn, code, project	save, check, math	learn, year, job	python, code, list	data, course, file	learn, python, start	question, ask, answer	print, loop, statement
Spring 2021	need, code, react	user, support, comment	look, code, state	function, value, method	google, game, backend	school, college, degree	learn, course, project	take, start, program	math, code, git	question, need, help	import, output, attribute	python, code, need	start, learn, course	data, file, code	function, object, list
Summer 2021	function, object, example	write, code, language	look, google, server	react, browser, project	user, make, test	game, question, answer	learn, job, year	student, day, hour	computer, science, comment	learn, code, program	function, code, name	data, need, run	print, list, loop	learn, python, code	panda, lambda, class
Fall 2021	function, object, array	tool, code, version	user, load, image	const, return, comment	react, learn, javascript	get, job, company	question, interview, ask	write, code, language	project, game, github	learn, program, start	data, file, code	print, score, video	project, book, tutorial	function, class, number	learn, program, code
Spring 2022	project, code, community	node, library, framework	code, need, know	record, type, object	code, write, change	school, money, skill	job, software, engineering	make, video, channel	data, job, teach	learn, program, language	run, code, file	ask, question, answer	function, list, loop	input, print, type	learn, problem, program

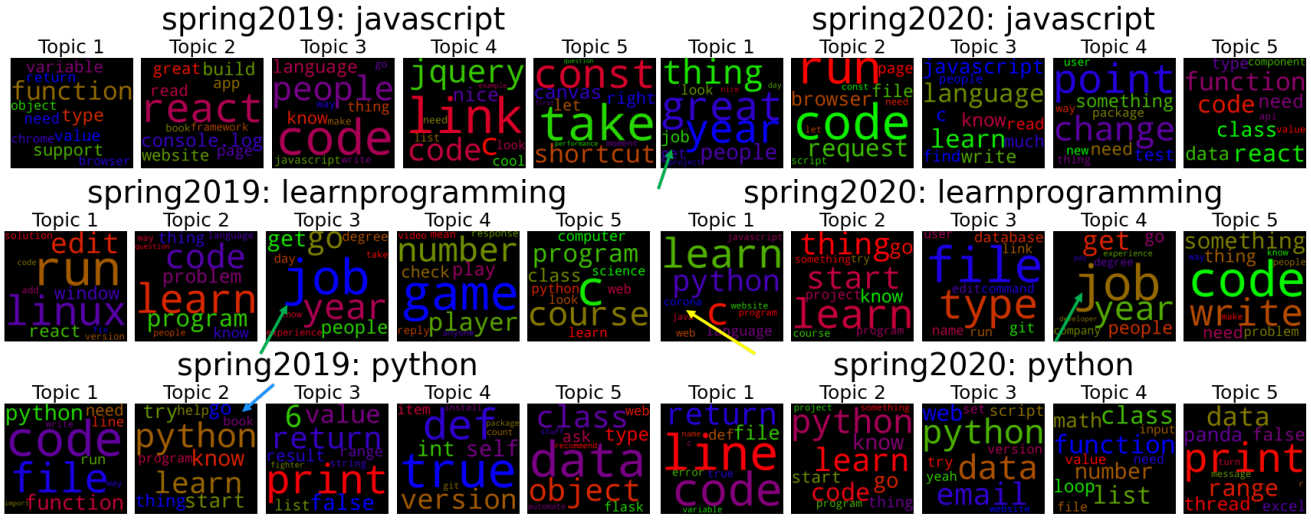

Fig. 3. Word clouds representing each topic (spring 2019 vs spring 2020).

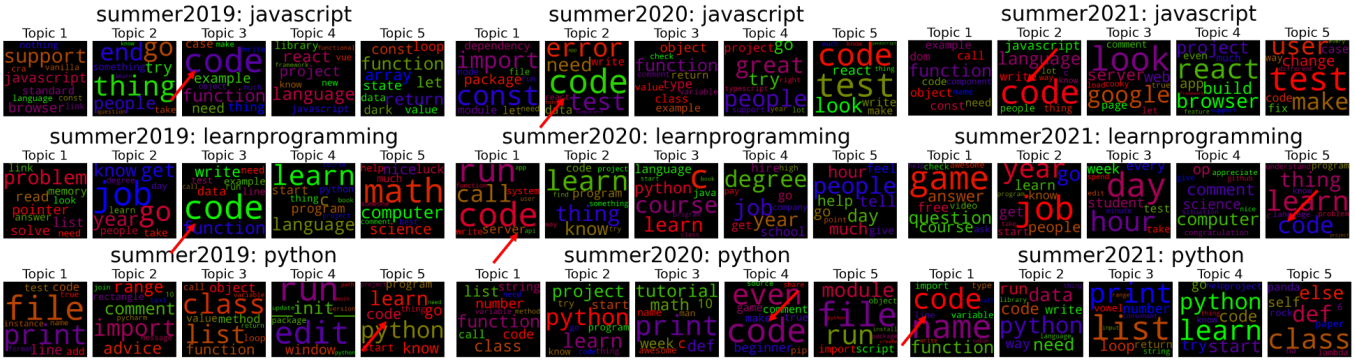

Fig. 4. Word clouds representing each topic (summer 2019 vs summer 2020 vs summer 2021).

IV. Results and Discussion

By harnessing LDA topic modeling, this study uses forty (40) months of 443,567 programming-related Reddit posts and comments to survey online learners using Reddit as an alternative resource to learn programming. The distribution of the number of posts for each subreddits from spring 2019 to spring 2022 is shown in Figure 2. Our study found that programming-related subreddits are popular, professional, and learning-oriented as an alternative resource for learning programming. To the best of our knowledge, this is the first NLP-powered study (using LDA topic modeling) to explore how online learners have utilized Reddit to learn programming.

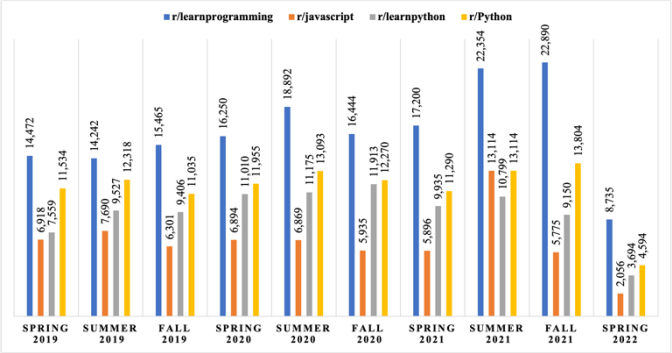

Fig. 2. Distribution of the number of posts from Spring 2019 to Spring 2022.

From the topic modeling of Spring 2019 versus Spring 2020 (see. Fig. 3.), we found that: the keyword of "corona" appeared in the topic 1 of *r/learnprogramming* spring 2020. COVID-19 pandemic affects every aspect of society, including economy, education, housing, transport, etc. In reviewing our corpus, we found that learners share in-person learning experience or things affecting their learning in these subreddits, for example, COVID-19 impacting their lives or work. The keyword of "book" (book suggestions) appears in the topic 2 of *python* in spring 2019. During pre-pandemic period, there were fewer topics related to "job", more "job" was mentioned during the pandemic (after 2019). Common words such as "learn", "program", "code", "python", and "c" all appeared in *r/learnprogramming* spring 2019 and 2020. From the comparison of topics of summer versus non-summer (see. Fig. 4.), we found that: during summer, there are more general *skills upgrade* related posts. Form non-summer, the posts are seeking help on more specific programming problems and about code sharing activities. The most common word was "code" (seeking coding help) every summer. Comparing these five topics of *r/javascript, r/learnprogramming*, and *python*, we found that except for spring 2021, at least one topic in *r/learnprogramming* contains the keywords "job". In all topics in *r/javascript* and *python*, there is no "job" in the top three words. Table 2 shows the top three words for each topic from the word clouds.

Our study finds that learners discussed various programming constructs, participated in code sharing activities, and expressed their feelings/emotions about learning programming and writing working codes. Even though the discussions in these subreddit communities are always focused on programming, at the beginning of the pandemic, the learners participated in the topic of coronavirus. Generally, we have observed that learners' participations during summertime are focused on mastering general programming concepts. However, during fall and spring, the learners seek help on very specific programming problems and get engaged in lot more code sharing.

Social media presents a quick and easy way for learners to connect with like-minded learners. Online learning communities make learning learner-centric, question-centric, and not teacher-centric. Our work will inform online learning communities about the prospect of utilizing social media for learning online. We suggest that integrating and incorporating social media into the teaching and learning process can lead to better learning outcomes.

V. Limitations and Future Work

There are a few limitations of our study. First, our Reddit dataset only contains the posts written in English; the results only capture the concerns and experiences of English-speaking users. Second, we collected the data only from the largest four programming-related subreddits communities, which are *r/learnprogramming*, *r/javascript*, *r/learnpython*, and *r/Python*. Future research should include more programming-related subreddit communities, and more social media platforms also need to be considered.

References

[1] Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T., & Anwar, M. (2021). Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, *9*(1), 1-16.

[2] Whitfield, C., Liu, Y., & Anwar, M. (2021, August). Surveillance of COVID-19 pandemic using social media: a reddit study in North Carolina. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 1-8).

[3] Kruse, L. M., Norris, D. R., & Flinchum, J. R. (2018). Social media as a public sphere? Politics on social media. *The Sociological Quarterly*, *59*(1), 62-84.

[4] Hemsley, J., Jacobson, J., Gruzd, A., & Mai, P. (2018). Social media for social good or evil: An introduction. *Social Media+ Society*, *4*(3), 2056305118786719.

[5] Chan, T. M., Dzara, K., Dimeo, S. P., Bhalerao, A., & Maggio, L. A. (2020). Social media in knowledge translation and education for physicians and trainees: a scoping review. *Perspectives on medical education*, *9*(1), 20-30.

[6] Online Education Statistics – How COVID-19 Changed the Way We learn? https://admissionsly.com/online-education-statistics/, last accessed 24 May 2022.

[7] Campbell, S (2022). Reddit Statistics 2022: How Many Reddit Users Are There? https://thesmallbusinessblog.net/reddit-statistics/, last accessed 11 June 2022.

[8] Muftah, M. (2022). Impact of social media on learning English language during the COVID-19 pandemic. *PSU Research Review*.

[9] Haythornthwaite, C., Kumar, P., Gruzd, A., Gilbert, S., Esteve del Valle, M., & Paulin, D. (2018). Learning in the wild: coding for learning and practice on Reddit. *Learning, media and technology*, *43*(3), 219-235.

[10] Akcaoglu, M., Hodges, C. B., & Jensen, L. J. (2022). Using Twitter to Form Professional Learning Communities: An Analysis of Georgia K-12 School Personnel Discussing Educational Technology on Twitter. In *Handbook of Research on Advanced Research Methodologies for a Digital Society* (pp. 510-525). IGI Global.

[11] Valle, M. E. D., Gruzd, A., Kumar, P., & Gilbert, S. (2020). Learning in the wild: Understanding networked ties in Reddit. In *Mobility, data and learner agency in networked learning* (pp. 51-68). Springer, Cham.

[12] Liu, Y. (2020). *A Comparative Study of Vector Space Language Models for Sentiment Analysis Using Reddit Data*(Doctoral dissertation, North Carolina Agricultural and Technical State University).

[13] PRAW: The Python Reddit API Wrapper, https://praw.readthedocs.io/en/stable/, last accessed 15 May 2022.