On the Congruence Between Online Social Content and Future IT Skill Demand

JALEHSADAT MAHDAVIMOGHADDAM, Ryerson University, Canada
NIRANJAN KRISHNASWAMY, Ryerson University, Canada
EBRAHIM BAGHERI, Ryerson University, Canada

The speed of digital transformation has resulted in new challenges for job seekers to become lifelong learners and to develop new skills faster than before. In this paper, our main objective is to examine how online content can serve as indicators for changes to the Information Technology (IT) industry and its in-demand skills. To study this relationship, we collect Reddit posts to represent social media content and job postings to reflect the IT industry based on which we explore possible correlations between them. Further, we propose a methodology to quantitatively estimate the predictive power of social media content for future in-demand skills. Our results show that the frequency of skill-related conversations on Reddit correlates with the popularity of skills in job posting data. Additionally, our findings indicate that the number of social posts dedicated to a specific skill can be a strong indicator for future job requirements. This is an important finding because identifying what skills the labor force should acquire will assist job seekers to plan their lifelong learning objectives to (a) maximize their employability, (b) continuously update their skills to remain in demand, and (c) be informed and actively engaged in defining knowledge trends, rather than reactively becoming informed of the latest information.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Future of Work; Lifelong Learning; Reddit; Social Media; Granger Causality

ACM Reference Format:
Jalehsadat Mahdavimoghaddam, Niranjan Krishnaswamy, and Ebrahim Bagheri. 2021. On the Congruence Between Online Social Content and Future IT Skill Demand. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 367 (October 2021), 27 pages. https://doi.org/10.1145/3479511

1 INTRODUCTION

While the evolution of technology, such as automation, artificial intelligence, and big data, has led to increased productivity and higher economic growth [43], it has also had a notable impact on the labor market, in-demand skills, and jobs. For example, today fewer people work in manufacturing compared to 1997 and many low-skill jobs have disappeared in, for example, the post office, book-keeping, and customer service sectors [76]. Even high-skilled workers are pushed down to perform jobs that were traditionally done by low-skilled workers [30]. A study by Frey and Osborne [30], who trained a Gaussian process on O*Net occupations for predicting their automatability, showed that about half of the US occupations are at the risk of being automated. Although this study does not indicate a time horizon for technological job-loss, it estimates this mass unemployment happening in the next 10 to 20 years.

Authors' addresses: Jalehsadat Mahdavimoghaddam, Jmahdavi@ryerson.ca, Ryerson University, Toronto, Ontario, Canada; Niranjan Krishnaswamy, Ryerson University, Toronto, Ontario, Canada, nkrishnaswamy@ryerson.ca; Ebrahim Bagheri, Ryerson University, Toronto, Ontario, Canada, bagheri@ryerson.ca.

2573-0142/2021/10-ART367 $15.00
https://doi.org/10.1145/3479511

Although it is not feasible to predict how exactly digitalization will impact the labor force [59, 76], previous studies show that technology is in favor of more skilled workers [43]. In addition to the importance of education and skill acquisition in the knowledge economy [30], the speed of digital transformation has remodeled existing jobs and has created new occupations that require new/multidisciplinary skills [17]. These changes are so fast that McKinsey Global Institute (MGI) reported that by 2030, up to 14 percent (75 M to 375 M) of the global workforce should change their job category and acquire new skills to be able to survive in the new job market [57]. Therefore, learning becomes a lifelong process and lifelong learners are required to develop skills faster than before. However, this rapid evolution of work and skills can lead to a lack of agreement on the disciplinary knowledge required for work that makes the process of job seeking and acquiring relevant skills challenging for both job seekers and learners [49]. Moreover, the ambiguity of the required knowledge and the speed with which skill demands change can affect the workforce's mental health and lead to anxiety [78, 84]. Greenspan [36] claims that these psychological effects will remain high on the workforce and learners in the knowledge economy, in which learning is a constant activity. Therefore, the need to continuously inform the workers on the changes in the job market and in-demand skills to help them upgrade and obtain new skills is paramount.

The computer-supported cooperative work (CSCW) community has a rich tradition of exploring and offering a deeper understanding of new forms of work in the evolving economic models [29, 51]. For example, CSCW researchers have already proposed a variety of computational tools that recommend new skills to lifelong learners and job seekers to support them in new skills acquirement for dealing with the rapidly evolving in-demand skills and the associated ambiguity [41, 63]. Recently, there has been a growing interest within the CSCW community to explore the impact of social media on professional development and the understanding of evolving skill requirements [48, 58]. However, to the best of our best knowledge, the existing literature does not provide any insight on the possible relationship between social content and skill demand change in the job market. Given the important role of social influence on individuals' skill development [16], there is a clear methodological gap in estimating the impact of online social content on skill demand shift.

In this paper, we examine whether social content can serve as an indication for the future of IT skills. In other words, we are interested to discover whether the content provided by online social communities can be considered as a source of information to obtain insight into the future in-demand IT skills. To this end, we propose a quantitative methodology to measure the correlation between social content around skills and job requirements reflected on Reddit and job postings, respectively. We systematically curate job postings and content posted on Reddit related to technology, career development, and learning that reflect the skills in the "Skills and Competencies" concepts from the ESCO [1] ontology [86]. Next, we study how each skill mentioned on Reddit correlates with skill representations within the job postings. Finally, we develop a Granger causality-based framework [35] to identify how social content are indicators of future job requirements.

Our work provides the following contributions:

(1) First, we empirically show the presence of a meaningful correlation between social content posted on online social platforms and future in-demand IT skills represented in job postings.
(2) Second, we show that by using historical social and job posting data, we can estimate the shift time at which the maximum correlation between online content and in-demand IT skills occurs.
(3) Third, we suggest that social content is an appropriate factor for gaining insight into the future skill requirements of IT jobs. A clear application of our study is providing insight to

[1] https://ec.europa.eu/esco/portal/skill

lifelong learners to help them systematically decide how and where to invest their time and resources to continue to be in demand, stay relevant, and reduce the risk of unemployment.

We hope this work inspires a new research direction that gives considerable consideration to the role online content can play on determining the future of in-demand skills. We believe that our findings enable future research to weigh social content in designing more insightful career and skill progression tools.

2 RELATED WORK

2.1 Historical Context of Labor Market Research

In previous studies, CSCW and HCI scholars have investigated the impact of technological advancements, such as Artificial Intelligence (AI), on the labor market and the future of work [33, 53, 64]. According to a literature review conducted by Lima and Souza [22], the interest in obtaining insight into the future of work dates back to 1956 with a growing amount of research in the 2000s. For example, Wang et al. [89] and Kalimeri and Tjostheim [42] studied the effect of Artificial Intelligence (AI) on employment and society, respectively. Moore [61] showed that AI enables human resource managers to more effectively acquire new talents and make decisions about their existing employees. Also, using internal and external data and utilizing machine learning algorithms, Moore argued that organizations can better understand their workforce performance, behavior, and the chance of attrition. However, despite the productivity enhancements offered by AI, the workforce might feel higher levels of anxiety and stress due to the constant feeling of being tracked [61]. Kluge et al. [46] explored the psychological impact of technology on the workers in different age groups. They showed that younger workers are significantly more impacted in terms of stress compared to both experienced and older workers. Other aspects of the future of the work include, but are not limited to, crowdsourcing [7], emergent skill relationships [37], professional development [49, 58], future workplace [9, 13, 38], and cognitive-based workstations [81]. Although the future of work has already been studied from different perspectives, reviewing all such content is beyond the scope of this paper. Therefore, in this work, we only review the literature related to skills and professional development.

The literature shows that in addition to new trends in the labor market, the very nature of work has also undergone considerable changes over the years. Advancement of technology has altered the world economy with new economical models, such as the gig economy [87] and the sharing economy [52, 72], which have reshaped the concept of employment and job. CSCW and HCI scholars are among the communities to investigate the technology-driven transformation of work(places) and new forms of work [29, 51]. These new forms include, but are not limited to, crowdwork [7, 10] and on-demand mobile work [5, 39], which refer to a series of tasks that are conducted by digital labor through a web-based platform, such as Amazon Mechanical Turk and clickworker, or software applications, such as Uber and Lyft [23]. According to the report presented by the international labor organization [5], the most important reasons for conducting on-demand work through digital labor platforms include the ability to work from home, having time flexibility, and receiving a complimentary source of income. However, the majority of the tasks on the digital platforms are currently simple tasks that do not require high education or a wide range of skills which further exacerbate the lack of skill development and career advancement.

Previous studies have shown that technological advances have shifted the labor market towards highly skilled workers [43] with an increased polarization between job opportunities and unemployment [15]. Also, technological change has caused the creation of new occupation types and dramatic change within existing jobs by requiring new skills and knowledge that were not required in the past [17]. Therefore, these occupational changes have led to a lack of agreement between

the required skills and the workforce knowledge [49]. For this reason, a growing body of literature is dedicated to understanding such rapid changes and the rise of new job types. We review these works in the context of our research in the next subsection.

2.2 Skill Development and Skill Demand Analysis

There have been efforts to provide insight and build tools within the CSCW and HCI communities that help learners and job seekers to transit to new jobs and acquire new skill sets as work becomes increasingly knowledge-intensive. As argued by Garn [32], some of the key success factors for learners in 2030 are the abilities to predict the skills that need to be learned and continuously gain new knowledge to keep up with industrial change. To support the skill development process, Broos et al. [8] have proposed a framework that provides Science, Technology, Engineering, and Mathematics (STEM) students feedback on their learning skills. Similarly, Kaewkiriya [41] introduced a tool in which information technology skills are recommended to students based on criteria, such as interpersonal and logical-mathematical intelligence. There have also been other works that offer a variety of tools to help job seekers and employees gain the required skills in the job market [25, 63, 75]. In addition to the works that primarily focus on helping students and job seekers gain industry skills, CSCW researchers recognize other factors that can impact skill development. For example, Marlow and Dabbish [58] reported the positive impact of the interactions that users have on an online community-based network named Dribbble on their professional development. Kou and Gray [48], studied Reddit and specifically, a subreddit called "r/userexperience", where users share knowledge and information about UX design. They investigated what UX professionals reveal about their knowledge and how self-disclosure supports professional communication. Finally, they showed that there is a connection between self-disclosure and professional development.

Besides skill development, it is crucial to systematically examine the shift in skill demand and continuously inform job seekers and lifelong learners about the skills that they should learn. Despite the importance of identifying the evolution of skills demand, only in recent years, quantitative analysis of the labor market is maturing [69]. There is a growing body of research utilizing historical data to analyze the trendiness of skills and career opportunities [85, 86, 92] for recommending new skills [18] and jobs [1, 31, 73]. As an example, Zue et al. [94] and Xu et al. [93] used job postings from an online recruitment platform to capture the change in recruitment demands. They used unsupervised learning algorithms to find the topic of each job posting for analyzing how the trends are changing. Aiming to investigate how the workforce adapts to the labor market changes, Chancellor and Counts [11] studied employment demand by analyzing the number of searches in different job categories on a search engine. Similarly, Jhaver et al. [40] used search queries to investigate the evolving trend of skills within the labor market.

One of the limitations of existing approaches is that they overlook the relationship between the variables that impact the labor market in favor of building predictive models. Additionally, existing literature on the future of work primarily relies on historical recruitment data and disregards the value of social media in analyzing future skills demand [66]. A benefit of using social media content over historical job postings is having access to near real-time data, which is valuable for fast-changing conditions. For example, as a result of the most recent global pandemic, most organizations across the world had to quickly turn to remote work. This new working situation required workers to obtain new IT skills to be able to work from home and survive in the changing economy. However, job postings took several months to reflect this new demand. In contrast, users on social media immediately started discussing the changes that the pandemic had brought about. For instance, a new sub-community named "COVIDProjects" was created in March 2020 on Reddit for users to discuss job demands related to COVID 19. The quick reflection of changes in social

networks before they become prevalent in job postings provides valuable insight for skill training, job replacement, and career planning.

The goal of our work in this paper is to study the impact of social content in this context. While there has already been a rich body of work on how social media can be used to understand different phenomena, such as mental health [28, 79], physical activities [2], social behaviors [56], food choices [21, 82], and social jet lag [54], among others, how online social content can be used to model the labor market remains unexplored. In this work, we study how social content can be indicative of the skill demand shift.

3 RESEARCH QUESTIONS AND HYPOTHESIS DEVELOPMENT

The main hypothesis of our research is to investigate whether online communities can indicate the changes in the IT job market. Although this hypothesis is not studied in prior work, there is a growing body of literature that suggests that social media is an effective source for understanding skill evolution in emerging jobs [49]. Occupations, such as cloud architects and network security engineers, are evolving fast without any consistent indication of the essential skills [45]. This rapid change introduces many uncertainties and challenges for universities to reflect occupational skill requirements in their curriculum and for learners and job seekers to follow a clear learning path for acquiring the required skills and entering the job market. Existing research in CSCW shows the content from online social communities has the potential to garner an understanding of in-demand skills [48, 58]. In community-based social media, such as Twitter and Reddit, users form a community with common goals and therefore they will interact more willingly even with strangers [90]. Kou and Gray have shown that people are comfortable with revealing information about their academic and professional backgrounds, future occupation plans, and skill development experiences within online social communities [48]. Constant et al. have shown that the technical advice obtained from strangers in situations where the information seeker has access to experts can be even more helpful compared to those coming from her strong ties. As Reddit is a virtual community formed based on weak ties between users from different countries, diverse levels of societal power, and knowledge expertise, individuals have the chance to connect to others with superior knowledge, resources, and expertise. Such weak ties allow the users to freely share their personal experiences and knowledge on a variety of topics including technological trends, potentially in-demand skills, and the evolution of the needs of certain market segments [68].

Despite the growth in the interest toward utilizing social content to answer a variety of societal questions, such as mental health [26], nutrition [82], and hate speech [65], to name a few, to the best of our knowledge, there has not been much work that takes advantage of social content to obtain actionable insight related to the job market. However, some of the existing studies in other domains, such as health and finance, are close to our research, from a methodology perspective. In the rest of this section, we discuss our research questions and review some methodologically-relevant related works, summarized in Table 1, that have motivated our work.

To use social media to model the effect of user-generated content on social phenomena, existing works often adopt two *strategies*, outlined in Table 1: (S1) identifying the variables that represent the social data and the phenomenon under study and finding any relationships between them, and (S2) developing quantitative models to understand the impact of the social content on the future of the phenomenon under study.

As shown in the first half of Table 1, in the first strategy (S1), we note that it is a common practice to represent data with its volume. Phillips and Gorse [71] highlight the importance of volume to understand the relationship between social media and financial market movement. Similarly, Sun et al. [88] used word counts to find the correlation between the stock market and Twitter user activities. From the labor market perspective, Yiming et al. [67] represented user data on Twitter

and Linkedin with word counts and investigated the relationship between personality traits and career progression. We adopt a similar strategy in our first Research Question (RQ):

RQ1: Is there any relationship between the frequency of social content and future IT skill demands expressed in job postings?

As it is shown in the second half of Table 1, in the second strategy (S2), insights on future social phenomena are inferred from online social content. For instance, an increasing number of researchers study the future and evolution of health-related issues with the aid of online social content. Notable examples include the work by De Choudhury et al. [19] who employed a Logistic Regression based model using word counts and the volume of Reddit posts and comments to get insight into future suicides. Dutta et al. [27] applied Granger causality and developed VAR models to understand the changes in online interaction from the changes in users' anxiety levels. Similarly, Shen et al. [83] developed VAR models and conducted linear and nonlinear Granger causality tests to estimate the future volume of traded Bitcoins. In a recent work conducted by Wu et al. [91], an occupational phenomenon named job burnout was studied using a social medium called Weibo. In their work, the authors identified job burnout posts with representative keywords, such as "work burnout" and "job burnout" and introduced different sets of variables by using posts words volume, posts sentiments, and the number of likes, posts, comments, and reposts. Then, the relationship between these variables and burnout was identified using the T-Test and Chi-Squared Test to understand whether these variables can be the right indicators. Motivated from this line of research, we put forward our second research question as follows:

RQ2: How can social content indicate the future of in-demand IT skills? In other words, can social content be used as an indicator for determining future job requirements of the IT industry?

By studying the correlation between social content and in-demand IT skills and the impact that social content has on the IT job market, we can understand whether social media is a valuable source of information to analyze future in-demand skills. It is important to note that the goal of this study is not to add evidence to the causes and effects of the labor demand and supply, but rather to sketch a broader picture of their equilibrium and try to find a source that helps us to understand in-demand IT skills. By finding a relationship between these time series, meaningful information on the required skills can be provided to job seekers, job providers, and educational institutes to systematically prepare them for the changes in the job market in the future.

4 DATA

4.1 Gathering IT Skills Data

In this work, to study the relation between social content and future IT job requirements, we utilized two data sources, namely Reddit and job postings. To curate our datasets, first, we collected a seed list of skills from the European Skills, Competences, Qualifications, and Occupations website (ESCO) that has been used in prior work in the context of labor market skills and occupations demand [14, 45, 86]. At the time of this study, the ESCO classification consists of 13,485 skills/competences and 2,942 occupations, mapped to exactly one ISCO-08 code in International Standard Classification of Occupations [2] (ISCO), which is a tool for international labor market reporting [4]. This connection between ESCO and ISCO enables us to use ESCO for studying the job postings.

We collected a list of 273 skills and competences from ESCO to query Reddit posts and job postings. These skills were gathered from all occupations under two broader categories namely, "information and communications technology professionals" and "business and administration professionals". Two researchers manually went through the obtained list and removed the homonyms or general skills, such as "Eclipse", "Perl", "Scratch", "Statistics", and "APL", to name a few. The reason for

[2]https://www.ilo.org/public/english/bureau/stat/isco/

Table 1. A brief summary of several methodologically-relevant related work.

Strategy	Paper	Venue	Source	Data Rep.	Domain	Framework
Strategy 1 (S1)	Yiming et al., 2017 [67]	Big Data	Twitter + Linkedin	Word counts (Reciptiviti)	Human Resource (career development)	Support Vector Regression (SVR) + ensemble learning
	Phillips and Gorse, 2017 [71]	SSCI	Reddit	Volume of posts per term + volume of new subscribers and authors in each subreddit + trading volume	Finance (cryptocurrency movement)	Hidden Markov models (HMM)
	Sun et al., 2016 [88]	IRFA	StockTwits	Word counts	Finance (stock market movement)	Sparse Matrix Factorization model
Strategy 2 (S2)	De Choudhury et al., 2016 [21]	CSCW	Instagram	Topic models	Public Health (food desert status)	Support Vector Machine (SVM)
	De Choudhury et al., 2016 [19]	CHI	Reddit	Word counts + volume of posts and comments	Public Health (mental health/suicide)	Logistic Regression (LR)
	Dutta et al., 2018 [27]	ICWSM	Twitter	Word counts + supervised classifiers	Public Health (social interactions impacted by the change in anxiety level)	Granger causality test + VAR models
	Saha et al., 2021 [80]	CSCW	Twitter	Word counts + ngrams + supervised classifiers	Human Resource (job satisfaction)	Support Vector Machine (SVM) + Logistic Regression (LR) + K Nearest Neighbor (KNN) + Random Forest (RF) + AdaBoost + multilayer perceptron (MLP)
	Wu et al., 2021 [91]	CSCW	Weibo	Word Counts + Posts, comments, and re-post volume + likes volume + posts sentiments + posts words volume	Human Resource (job burnout)	Support Vector Machine (SVM) + Logistic Regression (LR) + Decision Tree (DT) + Random Forest (RF) + XGBoost
	Shen et al., 2019 [83]	Economics Letters	Twitter	Volume of tweets + trading volumes	Finance (Bitcoin trade)	VAR models + linear Granger causality test + nonlinear causality test

this was our inspection on the search results using such skills showed noisy posts. In addition, a number of the skills were not frequent on either Reddit or job postings and hence were not helpful for our purpose. As such, we removed any skills that had an aggregate frequency of less than 100. As a result, we retained **55** skills. These skills are included in Table 8 (in the Appendix).

4.2 Finding IT Skills Job Market and Social Media Data Sources

We further describe how we used the above curated list of skills to acquire job postings and social content. To obtain IT job postings data, we obtained historical job postings published in two career websites, namely "Monster" and "Dice" from October 1, 2015 to July 1, 2016 from DataStock[3]. To maximize the search results and to include different forms of skills, such as "Artificial Intelligence" and "Artificial intelligent", we lowercase, removed the stop words, and stemmed all the words. Using this technique, we retained 1,223,619 job postings that contained at least one of the words in our list of skills in their job requirement, job description, or job title. Finally, we removed duplicate job postings that had the same city, company name, and job title and were posted within 14 days from the initial posting. After these steps, 961,267 job postings remained in our dataset. Table 2 shows the statistics for the job posting dataset. Additionally, we illustrate the distribution of the top-20 skills in this dataset in Figure 1.

Then, we downloaded Reddit posts from Pushshift[4] from October 1, 2015, to July 1, 2016. Similar to job postings, all words in the title or body of the Reddit posts were lowercased, stemmed, and stop words were removed. Then, we removed all the non-English posts, posts whose body or title had the tags "deleted" or "removed" or they contained only digits. Additionally, we deduplicated the posts that had the same title and author. Furthermore, we note there is content on Reddit that

[3]https://datastock.shop/
[4]https://files.pushshift.io/reddit/

Table 2. Statistics of the Job Posting Dataset.

Description	Count
Overall number of job posts	961,267
Number of unique companies	11,230
Number of unique cities	5,672
Number of unique states	50

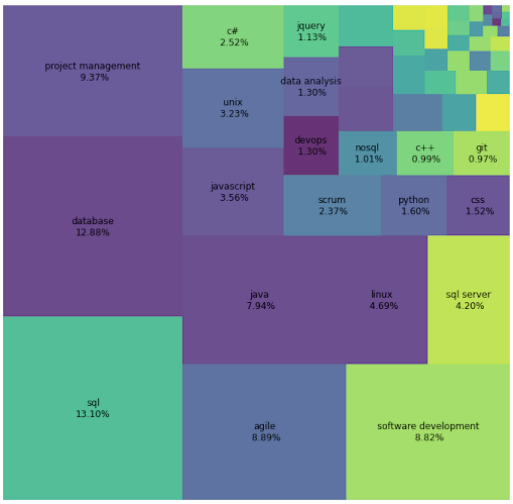


Fig. 1. Distribution of the top 20 skills on job postings.

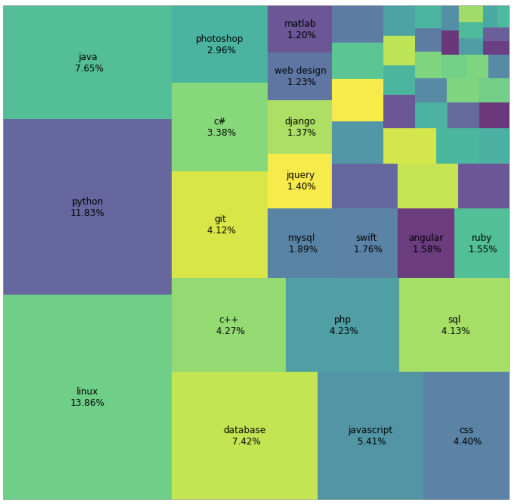


Fig. 2. Distribution of the top 20 skills on Reddit.

is solely for recruitment and job advertising purposes. These are often distinguished by "for hire" and "hiring" tags. Therefore, to separate social content from online job postings, we filtered out the posts that contained "for hire" or "hiring" tags. Finally, we removed the posts that appeared in subreddits with less than 500 relevant posts.

Following the above steps, we were left with 2,426 subreddits, however, in an iterative process through extensive tests, we created and updated a list of subreddit names that mainly contained posts with homonymous skills. For example, posts in the subreddit "r/Taylorswift" had the keyword "Swift" in them, which referred to the singer Taylor Swift and not Swift as the programming language. Another example is C# which also exists in the "r/music" subreddit and refers to a different concept than the programming language. To further eliminate the homonymous skills, we removed 117 subreddits from the dataset that were specifically dedicated to news, academia, or recruitment. Additionally, by removing all the subreddits that referred to non-technical topics, such as "r/relationship", "r/depression", and "r/starwars" we reduced the number of the subreddits to 554. Finally, we filtered out another 71 subreddits with less than 20 posts or based on their title or subreddit description. For example, "r/RecruitCS" was described as a subreddit that recruits CS players for a team and the content was not related to any of the IT skills in our list. Applying this technique, we collected 99,202 posts in 483 subreddits. Statistics of the Reddit dataset and the distribution of the top-20 skills are presented in Table 3 and Figure 2, respectively.

Table 3. Statistics of the Reddit Dataset.

Description	Count
Overall number of Reddit posts	99,202
Number of unique authors	52,785
Number of unique subreddits	483

4.3 Time Series Construction

After curating the two datasets based on Reddit posts and job postings, we built our time series for temporal analysis. In each dataset, we computed the weekly volume of the Reddit content and job postings around our list of skills. Since this frequency is different per week, we normalized our datasets using the cosine normalization method [88]. Before constructing the time series, it is crucial to test the stationarity of our data to detect the potential trending behavior that might lead to fabricated regressions [62]. Appropriate statistical tools should be applied to detect whether series are non-stationary and characterized with a unit root. However, informal methods, such as plotting the data, can aid with stationarity tests. Therefore, we first visualized our time series for each skill to identify any evident trends. In addition to informal methods, there are different stationarity tests that check for unit root. Among these formal methods, similar to the previous literature [3, 28], we applied the two popular tests called Augmented Dickey Fully (ADF) test [24] and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) [50] test for unit root tests.

The intuition behind the ADF test is that if data is characterized by unit root, then the time series is defined by a trend. Therefore, the past values of the time series cannot help to understand the future patterns and will lead to spurious results. In the ADF test, stationarity is checked via the ADF statistic, which is a negative number. The more negative the ADF statistic is, the stronger the rejection of the null hypothesis will be at some level of significance. Table 4 demonstrates the results of the ADF test for some sample variables in our datasets. We can reject the null hypothesis that the unit root exists if the p-value is smaller than 5% significance level and the test statistic is smaller than the significance levels. As an example, we can say "Angular" is stationary with a p-value of 0.0000 and 0.0006 and t-statistics of -6.2739 and -5.1025 in the Reddit and job posting datasets, respectively. In Table 4, the p-value and s-statistics of the variables that reject the null hypothesis are specified with a star. We apply first-order differencing [28] on any of the time series that do not pass any of the stationarity tests, such as MongoDB in the job posting and C++ in Reddit datasets, and then we run the stationarity tests on them again. After the first-order differencing, all of the time series passed the stationarity test except for C++: $t = -1.4$, $p = 0.58$ in the Reddit dataset and MongoDB: $t = -2.22$, $p = 0.19$ in the job posting dataset. Therefore, C++ and MongoDB were excluded from the experiments.

By applying the ADF test, we determined the time series which are stationary without unit root. However, it is possible for time series to not have unit root but be stationary around a deterministic trend [50]. Thus, we applied the KPSS test on the time series that passed the ADF test. In the KPSS test, the null and alternate hypotheses are opposite of the ADF test; the null hypothesis means the data is stationary. So, we interpret the p-value such that if it is smaller than the significant level (5%), then we reject the null hypothesis and we consider the data to be non-stationary. By applying the KPSS test, the calculated p-values for all the time series were greater than the significance value, which shows our data is not trend-stationary and therefore no further steps are required.

Table 4. Augmented Dickey Fuller Test at Level of Some Sample Variables.

Variables	Dataset	t-statistics	p-value	1% critical value	5% critical value	10% critical value
Angular	Reddit	**−6.2739** *	**0.0000** *	−4.6974	−3.9940	−3.6516
	Job postings	**−5.1025** *	**0.0006** *	−4.9701	−4.1225	−3.7276
MongoDB	Reddit	**−6.3996** *	**0.0000** *	−4.6974	−3.9940	−3.6516
	Job postings	2.5571	1.0000	−4.9701	−4.1225	−3.7276
Ansible	Reddit	−2.8585	0.3666	−4.8664	−4.0742	−3.6992
	Job postings	**−8.0133** *	**0.0000** *	−4.7069	−3.9986	−3.6543
C++	Reddit	−1.7498	0.8960	−4.9701	−4.1224	−3.7276
	Job postings	**−5.0987** *	**0.0006** *	−4.9701	−4.1225	−3.7276
Java	Reddit	−2.7195	0.4431	−4.7392	−4.0140	−3.6636
	Job postings	−1.7837	0.8871	−4.9701	−4.1224	−3.727

5 METHODOLOGY

5.1 Measuring Online Community Activity and Skill Demands Relationship

This section presents our approach to computing the relationship between online community activity and IT skill demands. We define online community activity and skill demand as the frequency of skills appearing on Reddit and job postings, respectively. Methodologically, similar to [54, 77], we use the cross-correlation coefficient (CCF) to discover the relationship between job postings and community activities and determine how well and at what point they best match up to each other.

We compute CCF for every skill in our dataset. It is important to note that the measure of similarity between the Reddit posts and job postings is only compared for two identical skills but not across skills because of the diversity between skills and their behavior as shown in Figure 3. In our work, the number of times that a skill is mentioned by Reddit users at time t is examined against the frequency of that skill showing up in job postings at time $t + \tau$. We measure the CCF value of the Reddit time series for the job postings time series as:

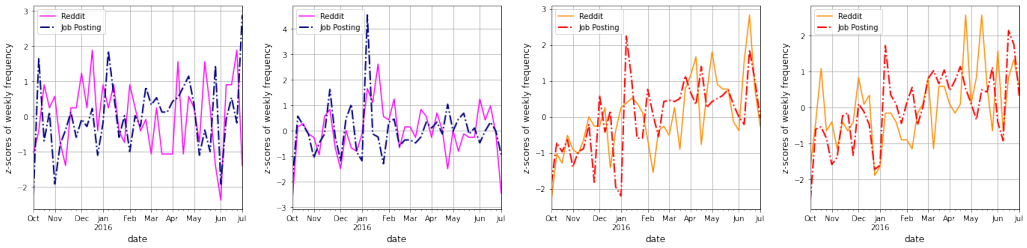
$$R_{Reddit,JobPostings}(t,\tau) = \frac{E[X_{JobPostings}(t+\tau)]E[X_{Reddit}(t)]}{\sigma[X_{JobPostings}(t+\tau)]\sigma[X_{Reddit}(t)]} \quad (1)$$

where τ represents the *shift time* by which the second time series is moved to represent the temporal changes of the first time series, $X_{JobPosting}$ is the moved job posting time series by the shift time τ, and $X_{Reddit}(t)$ is the Reddit time series. When τ is negative, it means that we can use the first time series, i.e. data from Reddit, to understand the second time series, i.e., the frequency of the job postings [77].

5.2 Granger-Causality and Skill Demands Indication

To measure the indicativeness of social content for future in-demand IT skills, we use Granger causality that has been extensively utilized in previous works [27, 83]. Granger causality is a causal inference method that provides us with information about future values of time series Y using values of time series X [35]. This approach can identify the causal dependencies between time series; however, it is important to note that Granger-causation cannot claim that X causes Y, rather *it determines whether X helps to understand the future evolution of Y.*

Although this technique was originally proposed for economic time series data [34], it is now widely used in other domains. For example, Arabzadeh et al. applied Granger causality on Twitter data to discover whether a user's social network can indicate the user's future interests better than their own historical preferences [3]. Dutta et al. [27] investigated Granger causation among temporal anxiety levels of Twitter users and their online interactions. Pavliková and Siničáková [70] used Granger causality in the context of the labor market to investigate the impact of macroeconomic labor market indicators, such as wages and labor cost, on future foreign investments in several



a. “Saas”. b. “Django”. c. “ML”. d. “DevOps”.

Fig. 3. Temporal frequency of skills with different behaviours.

countries. In this study, we utilize the Granger causality technique to understand the relation between online social content and in-demand skills. We can state that the frequency of job postings containing a specific skill is Granger-caused by the volume of social content around the same skill if regressing on both job postings and Reddit provides statistically significant information on the future skills observed in job postings [3]. Mathematically, the two regression models for determining the future frequency of skills in job postings can be defined as follows:

$$
\begin{aligned}
X_{JobPostings}(t) &= \sum_{l=1}^{\tau} \alpha_l X_{JobPostings}(t-1) + \sum_{l=1}^{\tau} \beta_l X_{Reddit}(t-1) + E_1(t) \\
X_{JobPostings}(t) &= \sum_{l=1}^{\tau} \alpha_l X_{JobPostings}(t-1) + E_2(t)
\end{aligned}
\tag{2}
$$

where α and β are the coefficients of the model, τ is the maximum number of shift times, and E_1 and E_2 are the prediction errors. If the results of the Granger Causality test are less than a significant value, i.e., $p = 0.05$, then we can reject the null hypothesis and conclude that the historical values of Reddit Granger-cause the activities of job postings.

6 FINDINGS

6.1 RQ1: Relationship Between the Frequency of Online Community Activity and Frequency of IT Skill Demand in Job Postings

6.1.1 Measuring The Relationship between Community Activity and Skill Demands. Figure 3 shows the temporal frequency of four sample skills in both Reddit and job postings on a weekly-basis. As shown in the figure, skills have different behaviors over time. For example, Figures 3a and 3b illustrate a constant demand and contribution for SASS and Django in 10 months both in job postings and Reddit. Unlike SAAS and Django, the occurrence of Machine Learning and DevOps, shown in Figures 3c and 3d, have an overall increase both in job postings and Reddit. In addition to the similar trend, we can observe that job posting fluctuations are in general similar to Reddit. For example, we observe that both time series in Figure 3b follow the same trends from October 2015 to mid-April 2016, with a few exceptional days in January 2016. Then, the time series have inverse relations from mid-April to June and end with similar patterns in July.

To understand the relationship between skill demand represented in job postings and content on Reddit, we calculate the cross-correlation for the skills that passed the stationarity test. We are looking for any relationship between the Reddit and job posting time series in time t up to 19 weeks after t ($t + 19$). Correlation values based on the weekly frequency of the skills appearing in job postings and Reddit at different shift time can be found in Figure 4. By looking at Figure 4, we can

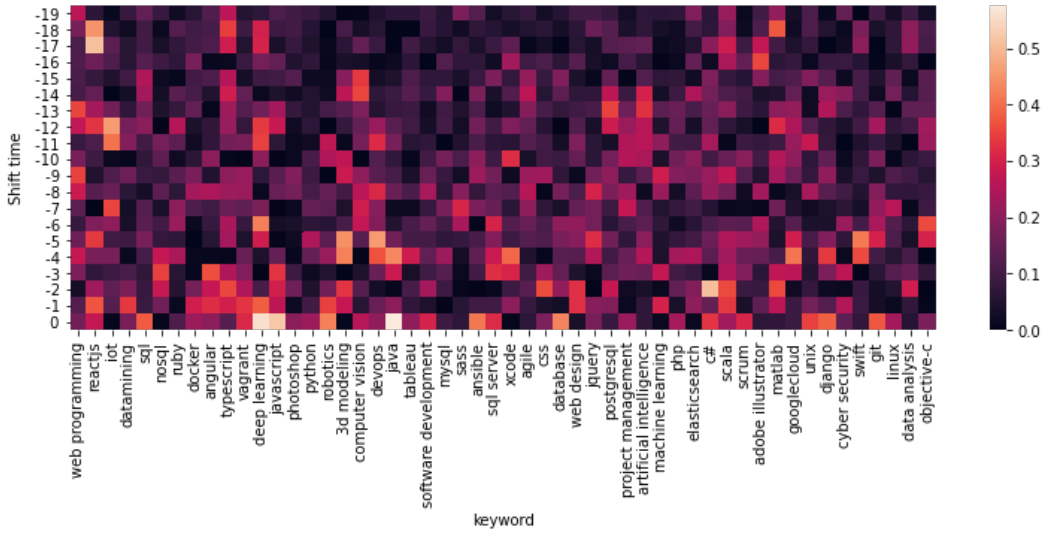
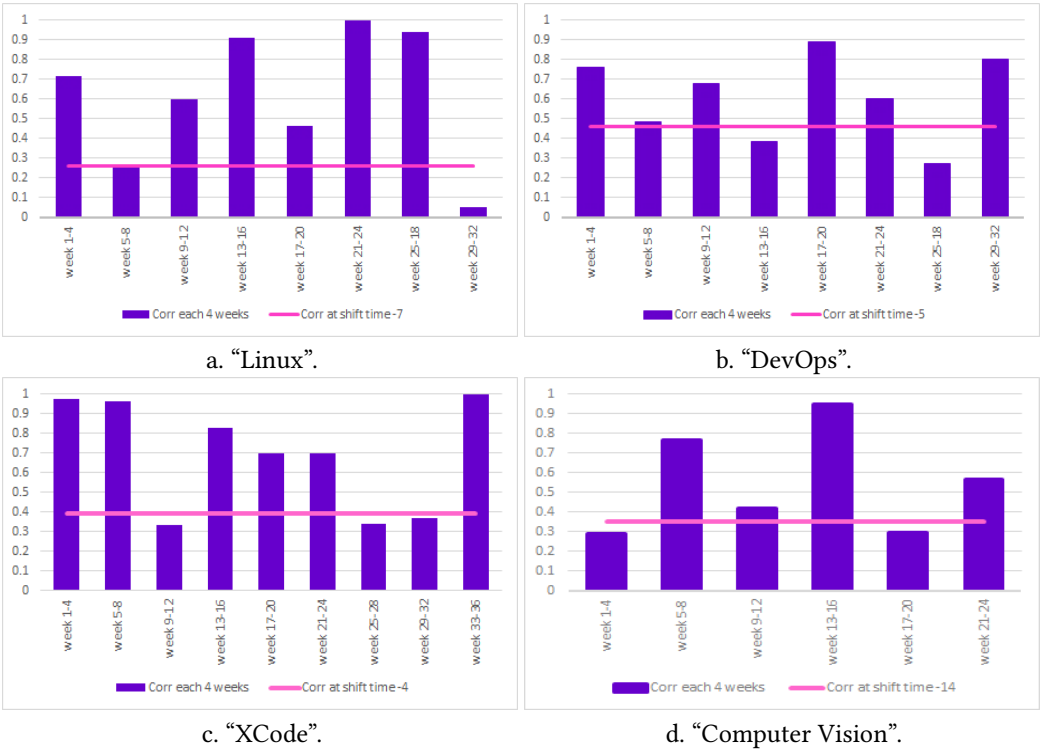


Fig. 4. cross-correlation of all the skills in different shift times.

see that the shift time that different skills correlate with each other are different. For example, the maximum correlation for Angular, Java, CSS, and Swift occur at shift time -3, -4, -2, and -5 with 0.36, 0.43, 0.36, and 0.41 coefficients, respectively. Therefore, we cannot select only one shift time for all the skills that will correlate Reddit content and job postings with the maximum coefficient. This finding is consistent with findings in information diffusion patterns on online social networks, such as [55, 74] that showed topics become viral with different velocities in different time windows depending on the topic.

In addition to finding a meaningful relationship between social content and job postings, we are interested to know how long this correlation may hold. To answer this question, for each skill we move the job posting time series by the best-discovered shift time. Then, we calculate the correction between Reddit content and job postings every 4 weeks and compare it with the overall CCF, as demonstrated in Figure 5. In Figure 5, this comparison is shown for 4 time series named Linux, Devops, Xcode, and Computer Vision, where the constant line represents the overall correlation at the best shift time and the histograms illustrate the correlation between Reddit and job posting every 4 weeks. According to our observation, for the majority of skills, the correlation every 4-weeks is higher than the overall correlation. This observation shows that the relationship between social content and job postings is reliable. Therefore, job seekers and hiring managers can rely on social content to continuously become aware of the skills that will become in-demand in the future. The monthly correlation of all skills can be found in Table 10 (in the Appendix).

The first finding of this study is that there is a meaningful relationship between the social content related to IT skills and in-demand skills represented in job postings. By analyzing the cross-correlation results, we observe a strong temporal alignment between the number of jobs that require IT skills and the number of times that Reddit users talk about those skills on a weekly basis. For example, there is an absolute CCF value of 0.35 between the number of times that "Web Programming" appeared in job postings at time t and "Web Programming" mentioned on Reddit 13 weeks earlier $t - 13$.



a. “Linux”. b. “DevOps”.

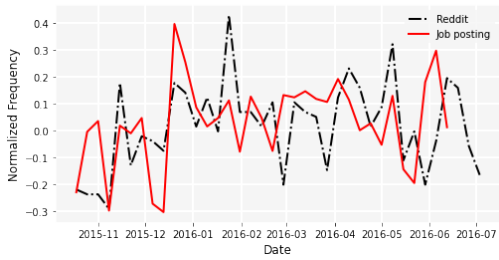
c. “XCode”. d. “Computer Vision”.

Fig. 5. Correlations between Reddit and job postings for each skill at the best shift time and every 4-week.

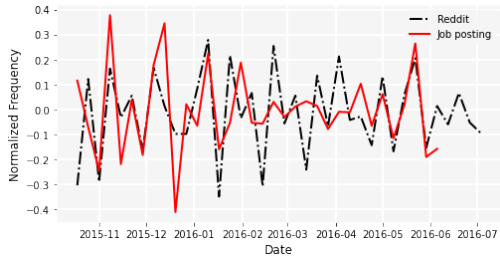
6.1.2 Minimizing data sampling bias. Social content has already enabled researchers to study a variety of social phenomena. However, storing and analyzing large quantities of data has led to computational challenges. Therefore, researchers are forced to apply different sampling techniques to select and examine a subset of data that sufficiently represents the population under study [20]. If the selected content differs from the whole data in important ways, the findings might be biased and the conclusions may not be generalizable [6] due to sample selection bias.

In the context of RQ1, one of the main threats to this quantitative study is sample selection bias [69]. To evaluate whether our extracted data properly represents the whole social content, when answering RQ1, we created two datasets by extracting Reddit posts in two ways: Dataset 1, which included posts that had at least one of our skills in their titles or body (99,202 posts), and Dataset 2 that consisted of all posts that had our skills in the subreddit names (69,202 posts). For example, “r/cpp” for the skill C++ and “r/csharp” for the skill C#. When there were more than one subreddits for the same skill, we selected the subreddit with more members, posts, and older creation dates. Furthermore, we selected the skills with the frequency of more than 100, as mentioned in Section 4.1, from both datasets and we ran the ADF and KPSS tests on time series of the skills-based on Datasets 1 and 2. We observed that two skills (C++ and MongoDB) in Dataset 1 and four skills (MongoDB, Business Intelligence, Project Management, SQL) in Dataset 2 did not pass the test after first-order differencing. Therefore, they were removed from the datasets.

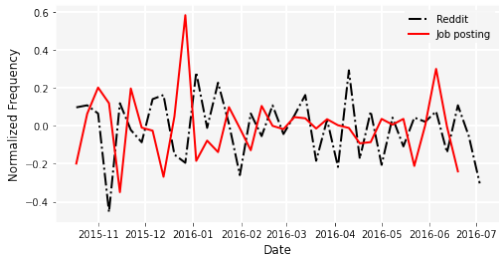
To compare the relationship between each Reddit dataset and job postings, we calculated the cross-correlation of each skill in Datasets 1 and 2. The correlation results for each dataset per skill can be found in Table 9 (in the Appendix). By comparing the correlation coefficients of each skill



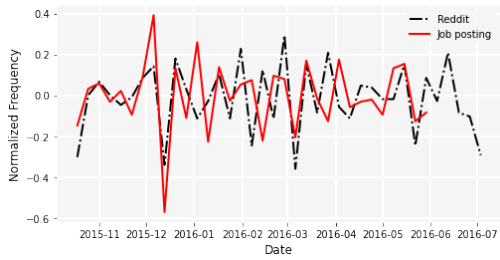
a. Angular Reddit and job posting frequency at shift time -3.



b. Java Reddit and job posting frequency at shift time -4.



c. CSS Reddit and job posting frequency at shift time -2.



d. Swift Reddit and job posting frequency at shift time -5.

Fig. 6. Time series with the shift time at the maximum absolute correlation.

in Datasets 1 and 2 we noticed that out of all of the skills, in 18 cases the correlation between the skills in Reddit and job posting was equal or slightly more in Dataset 1 compared to Dataset 2. The correlation coefficients of the rest of the skills in Dataset 1 were only slightly less than Dataset 2. Based on this comparison across two datasets that were extracted differently, we conclude that the adopted data extraction approach can be considered to be valid.

6.2 RQ2: Relation between Online Social Content and the Future of In-demand IT Skills

In this research question, we examine the potential of using social content to understand the distribution patterns of skills in future IT-related job postings. In this section, we present our findings and insights that were obtained through a Granger-causality framework, already described in the Methodology Section. To identify whether Reddit reflects changes of future job posting representations, we apply the Granger causality test to each skill in our dataset. If the obtained p-value from the test is smaller than the significance level, i.e., 0.05, we can reject the null hypothesis and conclude that the past values of Reddit Granger-cause future values of job postings. The results of the Granger causality test for some sample skills are shown in Table 5. The columns and the rows in this table represent the Reddit and job posting time series, respectively. For example, by looking at this table we can infer that the frequency of the "Web Design" skill mentioned on Reddit can indicate the future frequency of this skill in job postings. We report that the p-value of all skills was less than 0.05 except for "PostgreSQL" with a p-value of 0.14. Therefore, drawing from this analysis, **the second finding of this study** is that it is possible to use online content from a community-based social platform, such as Reddit, to better understand the temporal changes in job requirements. In other words, there is a significant linear functional connectivity between Reddit content and job postings, i.e., Reddit time series Granger-causes in-demand skills in job postings.

Table 5. Granger Causality test of selected skills; the columns are the skills from Reddit and are the determiner series (X) and the rows are the skills from Job posting response (Y).

Job posting	Reddit					
	PostgreSQL_R	WebDesgin_R	Angular_R	Tableau_R	Ansible_R	Robotics_R
PostgreSQL_J	0.1462					
WebDesgin_J		0.0045				
Angular_J			0.0000			
Tableau_J				0.0086		
Ansible_J					0.0000	
Robotics_J						0.0056

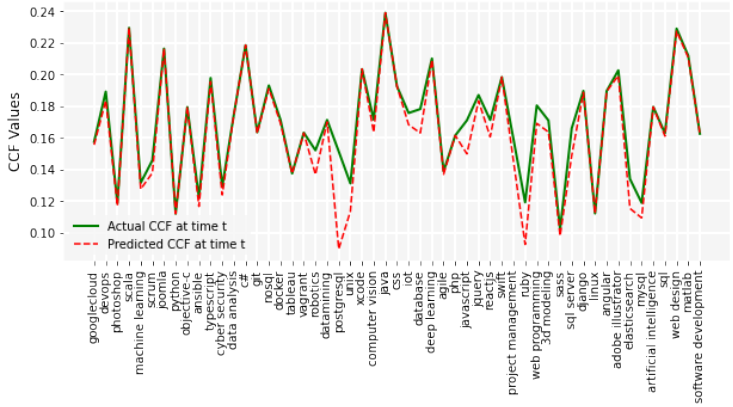


Fig. 7. Observed Average CCF at Time t vs Predicted Average CCF at time t.

In addition to our interesting finding of Reddit content Granger-causing job postings, we further investigate whether we can identify the optimal shift time to understand the future of IT job postings. The reason that we are interested in the optimal shift time is that the speed and delay for the spread of information on social networks are not the same for different topics [55, 74]. Therefore, it would not be possible to find a single shift time that would be universally applicable for all skills in which a meaningful relationship exists. So, to identify the optimal shift time, we use all observations that occurred up to $t-1$ to find the shift time at which the maximum correlation between job postings and Reddit happens. Then, we use the optimal shift time up to $t-1$ to find the CCF between the two-time series at time t. Figure 6 shows the time series for four sample skills, namely Angular, Java, CSS, and Swift when the optimal shift time is used. In some cases, the estimated optimal shift time is different from the actual best shift time. For instance, the CCF of the "Web Design" skill is 0.36 at shift time -1, which is estimated as the optimal shift time. However, the actual best shift time is -2, which leads to a CFF of 0.42. Figure 7 shows the CCF of each skill at time t computed using the estimated optimal shift time versus the calculated CCF using their actual optimal shift time. By inspecting Figure 7, we observe that the historical relationship between job postings and Reddit has aligned with their relationship at time t and the difference between their estimated and actual CCF is not significant.

These results point to **the third finding of our study** that historical information of Reddit and job postings can be used to estimate the optimal future shift time at which the maximum correlation occurs between the Reddit and job posting time series.

7 DISCUSSIONS

We began this study by asking whether social media can be considered as an indicator of changes in the IT job market. To answer this question, we first investigated whether there is any relationship between social content produced by online communities and in-demand IT skills. By applying a quantitative method in Section 5.1, we suggested that the frequency of IT-related social content posted on community-based platforms, specifically Reddit, has a noticeable correlation with future in-demand IT skills within job postings. Also, we showed that the shift time at which the maximum correlation between Reddit and job posting time series occurs can be estimated via historical data. Then, we asked whether social content can provide insight on future IT job requirements. By applying a Granger-causality framework introduced in Section 5.2, we showed linear functional connectivity between Reddit content and job postings that suggests one can get insight into future job requirements by tracking the frequency of skills mentioned in online communities.

Note that although our results shed light on the importance of social content for understanding the IT job market, our findings are limited by the scope of our dataset. Therefore, our observations are only valid for IT professionals and further studies are required to generalize the proposed methodology for other skills and occupations. Moreover, we note that the goal of this study is not to propose a skill prediction model. Rather, we focus on understanding whether, in the absence of any other indicators, social content can be used to get insight into future job requirements. Therefore, we do not examine the existence of any other relationships between our datasets. For example, even though job postings might help us understand the social activities around a particular skill on Reddit, we did not study this reverse relationship because it is beyond the scope of this research.

Previous research in CSCW touches on the future of work and labor market evolution [29], and how social media can support skill development [48]. Despite the popularity of online social networks, as a social analysis tool within the CSCW community, to the best of our knowledge, there has been no work that investigates the potential of social media as a data source for indicating the changes in the IT job market. Current state-of-the-art methods mainly rely on historical recruitment data to find patterns of skill trendiness [92–94]. However, with rapidly evolving in-demand skills and emergent disciplines, it is crucial for CSCW researchers to leverage suitable sources of information to track the job requirements more systematically. This study provides the first empirical study that highlights the possibility of investigating the future changes in the job market via a highly available, large-scale, instant source of data. In the following sections, we discuss the implications of our work for job seekers, hiring managers, and policymakers and conclude with limitations and future works.

7.1 Implications for Job Seekers

In the knowledge economy, workers are required to be independent fast-learners who can identify new opportunities, such as learning opportunities, and upgrade or gain emerging skills [32, 60]. Self-regulated and continuous learning already encourages the workforce to be proactive in learning the skills that are required to complete their tasks, even if those skills are not a part of their job mandates. This requires the workforce to be vigilant and proactive in gaining new skill sets. In this context, McGregor et al. [60] finds that despite the importance of continuous learning, obtaining new skills is not necessarily happening within the context of on-the-job training and by the employers, but rather is happening through personal networks and outside-of-work interaction. Similarly, CSCW scholars showed that frequent occupational changes make online communities a more effective source of information for skill development because of the fast reflection of the most needed skills [58].

Table 6. Real examples of Reddit titles and posts - how job applicants use Reddit.

Title	Summary of the post	Subreddit
Help needed how to continue	I am a novice web designer and wanted to expand my resume more and be able to do more than just HTML, CSS,basic jquery,bootstrap.I am delving into Ajax,Json but am confused and don't want to waste time on outdated technology. I know basic ASP classic. **I am so confused about what to learn next.** Ruby/ruby on rails. Django .net not so much as my job is a .net shop and I am too far behind and not in the team that does .NET as I am on 2.0 and they are on 4+.	r/IWantToLearn
Info Sys student taking gap semester, have questions.	I am a junior in university taking a semester off for personal reasons. **I have decided to gather up some certifications to boost my resume** (my GPA is quite low from switching majors). What are the most useful/best certifications to snag up in my time? I am looking at security+ at the moment. **Also what languages should I learn?** I am skilled in Java and C# and of basic skill in R (I think python might be a good move). What would you guys do with a semester off in my shoes? Any advice is appreciated.	r/Sysadmin
What API/Platforms should I learn for web programming?	**I need to update my skills.** Currently I am a backend programmer who does front end once in a while. I'd like to use my vacation to learn more frontend and put currently trending APIs/platforms on my resume. I don't want any language specific suggestions. I'm not convinced SPA is a good idea but I'm considering looking at angular 2. **React seems good do any of you recommend learning it? Is there anything else I may want to use with JQuery or bootstrap? Is there a list of very popular JQuery and Wordpress plugins?**	r/Webdev
For those that made the career change into a programming position with virtually no experience to begin with, what did you use as resume builders?	when it comes to that i would think either a degree in CS, or certificates would be best rather than "completed X tutorial on X website etc ..." I have been thinking about this awhile now and was dead set on paying for an online programming that would net me a certificate that I could prove that I learned, but **wondering those that have taken the leap into coding and have landed a job, what did you add to your resume?**	r/Learnprogramming

Further, our results shed light on the importance of online communities as a potential source of information for monitoring the skills that can become highly in-demand and necessary requirements in future jobs. Our finding is in accordance with prior works that showed the role of OSNs in understanding the required skills for emerging jobs [32, 48]. We observed that some of the job seekers already rely on online experts' opinions to improve their interviewing skills and to obtain insights on the most in-demand skills that need to be reflected in their resumes to maximize their employability. Table 6 illustrates the titles of 4 posts, a shortened version of their post content made by job seekers, and the name of the subreddits. We observe in these examples that users usually provide a summary of their background and knowledge, the reason that they want to upgrade their skills or resumes, such as entering the job market or career transition and then seek advice from the community. In the "summary of the post" column of Table 6, the main sentence that clarifies the type of required help from the community is highlighted. The findings in the paper hint that the workforce may rely on their own social circles and online communities to smoothen their learning curve, get insight into the skills that will assist them to perform their job more effectively, and understand important skill trends within the labor market to stay in demand.

7.2 Implications for Hiring Managers

Besides the significant role of social media in knowledge production, previous studies highlight the importance of social networks in recruitment and hiring processes [44]. It is important to note that in previous studies the main focus has been on understanding how social media, such as Linkedin, facilitates talent acquisition [47]. However, in this study we noticed that similar to lifelong learners and job seekers, hiring managers rely on online communities to become aware of the latest in-demand skills for different occupations and reflect them in their job requirements. In Table 7, we show the title, "summary of the post", and subreddit of 4 posts shared by hiring managers or recruiters. In the "summary of the post" column of Table 7, the main sentence that clarifies the type of required help from the community is highlighted. As shown in Table 7, hiring managers may be aware of the general requirements and the job title that they are looking for. However, they

Table 7. Real examples of Reddit titles and posts - how hiring managers use Reddit.

Title	Summary of the post	Subreddit
How do I hire an AI specialist?	I'm a Director of Research for a financial services company, and we're looking to hire an AI specialist for the first time (right now we've focused on having PhD in Maths that generate models). While I'm eager to hire someone with this skill-set, **I'm finding it difficult to know what to look for.** Does anyone have some advice on what it takes to find the good AI specialist?	r/Artificial
Looking for advice about hiring the first engineer for our SaaS startup	I'm the technical co-founder in a startup that creates software solutions for 3D graphics designers. The software is quite complex, combining multiple programming languages (full stack web + windows native app), and it is also poorly documented. To meet with our users' expectations we are planning on hiring our first engineer to help me out with development, but I've never done this before. **Should I try getting someone part time and then scale up to a fulltime dev later?** Should I look for someone who could work on the **whole product alongside me, or should I divide it up**, so for example I would only work with the frontend and they could work on the backend? Is it possible to have a remote team right away in these early stages, or is it necessary to start out in an office working together?	r/Startups
We're hiring a CTO with node.js background, but is that actually what we need?	We've recently started to hire for a CTO position to lead the tech side of our startup. We're at a pretty important point of transition where the team is starting to grow quite quickly, so we need someone with great team management skills, but for the next couple of months it'll also be important for them to dive into the code and really help on delivery as we ship products. **Would love input on the job spec** we've put together - I've already made quite a few updates from other input received so far. It's a really tricky hire, so I'm keen to make the spec / role as compelling as possible...	r/Node
Advice for hiring a biz development person	I run a web design and development agency and we're looking to hire somebody to help us bring new business. We're calling the position "Head of New Business" and I'd like to get some thoughts on **what we should expect from experience, expectations, and salary/compensation)**. A few details to help paint a picture of the role and company: [The author provides details on the company, the available budget, the timeline, location, department] We're looking for somebody who can prospect, find new leads, and close new projects. We don't have experience hiring a biz dev person and would love some advice. **What should we expect experience wise? How many years? What are the top attributes to look for in this position, e.g. prospecting ability, previous goals, experience?**	r/Sales

still seek advice from online experts to understand what skills they should look for, whether they need full-time or part-time employees, the suitable working environment, compensation, and years of experience, to name a few.

As mentioned in Section 4.2, in the pre-processing step, we removed the subreddits that are particularly dedicated to recruitment posts, such as "r/recruitment" and "r/recruiting". However, by looking at Table 7, we can see that hiring managers and recruiters also use other non-recruitment subreddits, such as "r/startups" and "r/sale" and technical subreddits, such as "r/node" and "r/artificial" to get hiring advice. This observation might indicate that recruiters and hiring managers in the IT industry are also active consumers of Reddit and they rely on social data for in-demand skill awareness as much as lifelong learners and job seekers do. Our findings and observations suggest that technical online communities may have an impact on how IT job requirements are formed and reflect in-demand skills. Therefore social content posted on community-based platforms, such as Reddit, can serve as an adequate indicator of the future in-demand skills represented in job postings. This study introduces a beneficial line of research that is in accordance with CSCW's growing interest in implementing frameworks that help industries to get insights into the recent changes in the job market.

7.3 Implications for Policymakers

As shown in Table 6, one of the common aspects of the Reddit posts is that applicants need to know about the new skills that help them improve their resumes. However, our further analysis illustrates that finding the answer to the "what to learn next" question can be a frustrating process

for many job seekers. Here are 4 examples that show the confusion that users expressed around skill development either on the title or body of their posts:

Sample Title 1: **Overwhelmed** with what to learn next

Sample Title 2: Does anybody else **feel overwhelmed** looking at how much there is to learn?

Sample Body 1: I am looking to become a sysadmin or DevOps engineer and am struggling to find out what I need to know or where to start. **It is very overwhelming**.

Sample Body 2: I've been working as a Data Analyst for the past 3 years and also trained on Oracle Pl/SQL. Now **working on Oracle too**, **I just wanted to change but was confused about "What should I learn"**.

These examples further explain that although individuals use social media as a source of skill development, there is still a need for tools that help new graduates, job seekers, and lifelong learners to systematically get informed of the changes in the job market by getting some insight into these changes. Researchers have already proposed techniques that rely on historical recruitment data to identify the popularity of skills in a fast-paced job market [92–94]. These studies mainly focus on finding patterns of skill trendiness using Machine Learning techniques. While they can help analyze the job market, they have limitations in investigating possible relationships between the variables that impact the labor market. Additionally, this line of literature only relies on historical recruitment data and disregards the influence of social content in the labor market.

Our findings suggest that by leveraging social content, we may be able to capture the recent changes in the IT job market, which might not always be instantly mirrored in job postings. However, it is important to note that although we utilize Granger-causality, we do not claim that our framework will be an application of job market prediction. One of the main reasons is that to make a predictive framework, multiple variables should be considered within and outside the social media context. Also, we acknowledge that the labor market cannot be studied in isolation from the other factors, such as the stock market, major medical events, such as COVID 19, and technological news, to name a few. This study can aid academic institutions and policymakers to identify the in-demand skills empirically with access to large-scale and appropriate data that can be gathered with no time gaps from online communities. Therefore, universities can design forward-looking curricula based on the future job market demand. Similarly, the human resource sector can conduct talent acquisition, career advice, and future investments based on the observed demand.

7.4 Limitations and Future Works

While our findings illustrate the effectiveness of social content in the context of our study, we recognize some limitations in this research that we hope to address in our future work. One of the limitations is that we only examine the relationship between the frequency of online content and IT related job requirements, however, the semantics of this content remains unexplored. For instance, in our current study, we treat the following two Reddit posts in the same vein: 1) *"'I have been a Spark user for a few years, However, I can't take it anymore! Every day, I have to spend so much time trying to reverse engineer all bugs and limitations of the dataframes framework. It is impossible to go beyond a simple word count project in Spark!".* 2) *"Spark is great for dealing with massive data, you can query your data very quickly using the distributed infrastructure."* However, the first post is discussing the disadvantages of "Spark" while the second post mentions the advantages of this skill. Therefore, It would be important to examine the impact of Reddit content semantics on job demands in addition to considering their frequency.

The next limitation relates to the distribution of the skills in our datasets. Currently, the job titles in our job posting dataset are biased toward software development positions, such as "Java developer", "software engineer", and ".NET developer", to name a few. This bias causes an unbalanced

distribution toward the skills, such as, "Python", "Java", "Linux", and "Javascript", as illustrated in Figures 1 and 2. Therefore, our study might be considered to be biased towards more popular job skills and prevent us from exploring the impact of less popular skills, such as "SparQL" and "GraphQL". Moreover, although in the current work, we consider the speed and delay of information spread, we do not examine the effect of information longevity on in-demand skills. Skill longevity can be determined by considering the time window in which a skill remains popular within the community. For example, a particular skill might attract the community's attention for a few weeks through continuous posts and comments as opposed to another skill that might only be discussed occasionally. Addressing these challenges will provide significant insights into in-demand skills and their evolution.

Additionally, in this study, our focus has been on Granger-causation between technology-related content and in-demand skills. In other words, we extract content from Reddit sub-communities that are related to technology, such as "r/AskProgramming", "r/ProgrammingLanguages", and "r/technology". However, other factors are also known to influence employment, such as the stock market [12]. Therefore, it remains an open avenue for the future to address the relationship between the content of sub-communities, such as those related to the stock market, and in-demand skills.

Other future work includes extending our research to understand how topics of online content impact the future representation of skills. For instance, in the Java subreddit, there could be posts around job search, learning, and advice on projects. By applying appropriate topic modeling techniques, we can cluster the posts that are semantically related to each other. Then, the relationship between each of these topics and the future representation of skills can be determined. The effect of topics on skill demand would be an interesting direction for the future.

8 CONCLUSION

In this paper, we have presented one of the first attempts to quantitatively study the relationship between social content and IT skill demand. We captured 961,267 job postings and 99,202 posts around information and communication technology skills in 484 different subreddits, such as "r/learning", "r/programming", and "r/gamedev". We examined how social content reflects upon the future of skill representations on IT-related job postings by measuring the congruence between the frequency of skill in social content and IT job postings. We showed that there is a meaningful relationship between online content and skill demand. Also, we applied the Granger-causality analysis to understand whether online community activity Granger-causes skill demands. We showed that Reddit content can assist with understanding the in-demand skills. We believe that our work in this paper is aligned with how workers in the knowledge economy are acquiring skills, namely through their social connections and interactions, a reflection of which can be seen in online social networks. Hence, this work provides an organic and efficient way of understanding important skill trends within the labor market based on online social network content.

REFERENCES

[1] Fabian Abel. 2015. We know where you should work next summer: job recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems*. 230–230.

[2] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proceedings of the tenth ACM international conference on web search and data mining*. 537–546.

[3] Negar Arabzadeh, Hossein Fani, Fattane Zarrinkalam, Ahmed Navivala, and Ebrahim Bagheri. 2018. Causal dependencies for future interest prediction on Twitter. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1511–1514.

[4] Miroslav Beblavỳ, Mehtap Akgüc, Brian Fabo, and Karolien Lenaerts. 2016. What are the new occupations and the new skills? And how are they measured.

[5] Janine Berg, Marianne Furrer, Ellie Harmon, Uma Rani, and M Six Silberman. 2018. Digital labour platforms and the future of work. *Towards Decent Work in the Online World. Rapport de l'OIT* (2018).

[6] Richard A Berk. 1983. An introduction to sample selection bias in sociological data. *American sociological review* (1983), 386–398.

[7] Daren C Brabham. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* 14, 1 (2008), 75–90.

[8] Tom Broos, Laurie Peeters, Katrien Verbert, Carolien Van Soom, Greet Langie, and Tinne De Laet. 2017. Dashboard for actionable feedback on learning skills: scalability and usefulness. In *International Conference on Learning and Collaboration Technologies*. Springer, 229–241.

[9] Yujia Cao, Jiri Vasek, and Matej Dusik. 2018. Design Towards AI-Powered Workplace of the Future. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 3–20.

[10] Melissa Cefkin, Obinna Anya, Steve Dill, Robert Moore, Susan Stucky, and Osariemo Omokaro. 2014. Back to the future of organizational work: crowdsourcing and digital work marketplaces. In *Proceedings of the companion publication of the 17th ACM conference on computer supported cooperative work & social computing*. 313–316.

[11] Stevie Chancellor and Scott Counts. 2018. Measuring employment demand using internet search data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[12] Gabriel Chodorow-Reich, Plamen T Nenov, and Alp Simsek. 2019. *Stock market wealth and the real economy: A local labor market approach.* Technical Report. National Bureau of Economic Research.

[13] Michael F Clarke, Joseph Gonzales, Richard Harper, David Randall, Thomas Ludwig, and Nozomi Ikeya. 2019. Better supporting workers in ML workplaces. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 443–448.

[14] Emilio Colombo, Fabio Mercorio, and Mario Mezzanzanica. 2019. AI meets labor market: Exploring the link between automation and skills. *Information Economics and Policy* 47 (2019), 27–37.

[15] McKinsey & Company and James Manyika. 2017. *Technology, jobs, and the future of work.* McKinsey Insights.

[16] David Constant, Lee Sproull, and Sara Kiesler. 1996. The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization science* 7, 2 (1996), 119–135.

[17] Olivia Crosby. 2002. New and emerging occupations. *Occupational Outlook Quarterly* 46, 3 (2002), 16–25.

[18] Vachik S Dave, Baichuan Zhang, Mohammad Al Hasan, Khalifeh AlJadda, and Mohammed Korayem. 2018. A combined representation learning approach for better job and skill recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1997–2005.

[19] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2098–2110.

[20] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, Kasim Selcuk Candan, Lexing Xie, and Aisling Kelliher. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 4.

[21] Munmun De Choudhury, Sanket Sharma, and Emre Kiciman. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*. 1157–1170.

[22] Yuri Oliveira de Lima and Jano Moreira de Souza. 2017. The future of work: Insights for CSCW. In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 42–47.

[23] Valerio De Stefano. 2015. The rise of the just-in-time workforce: On-demand work, crowdwork, and labor protection in the gig-economy. *Comp. Lab. L. & Pol'y J.* 37 (2015), 471.

[24] David A Dickey and Wayne A Fuller. 1981. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: journal of the Econometric Society* (1981), 1057–1072.

[25] Tawanna R Dillahunt and Alex Lu. 2019. DreamGigs: Designing a Tool to Empower Low-resource Job Seekers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[26] Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[27] Sarmistha Dutta, Jennifer Ma, and Munmun De Choudhury. 2018. Measuring the Impact of Anxiety on Online Social Interactions.. In *ICWSM*. 584–587.

[28] Sindhu Kiranmai Ernala, Tristan Labetoulle, Fred Bane, Michael L Birnbaum, Asra F Rizvi, John M Kane, and Munmun De Choudhury. 2018. Characterizing Audience Engagement and Assessing Its Impact on Social Media Disclosures of Mental Illnesses.. In *ICWSM*. 62–71.

[29] Anne-Laure Fayard. 2019. Notes on the meaning of work: Labor, work, and action in the 21st century. *Journal of Management Inquiry* (2019), 1056492619841705.

[30] Carl Benedikt Frey and Michael A Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114 (2017), 254–280.
[31] Snorre S Frid-Nielsen. 2019. Find my next job: labor market recommendations using administrative big data. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 408–412.
[32] Myk Garn. 2015. AfterNext: Decoding the Future of Higher Education in 2030. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 54–65.
[33] Camilla Gjellebæk, Ann Svensson, and Catharina Bjørkquist. 2020. The Dark Sides of Technology-Barriers to Work-Integrated Learning. In *International Conference on Human-Computer Interaction*. Springer, 69–85.
[34] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.
[35] Clive WJ Granger. 1988. Causality, cointegration, and control. *Journal of Economic Dynamics and Control* 12, 2-3 (1988), 551–559.
[36] Alan Greenspan. 2000. The Evolving Demand for Skills. (2000).
[37] Keman Huang, Jinhui Yao, and Ming Yin. 2019. Understanding the skill provision in gig economy from a network perspective: A case study of fiverr. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[38] Julie Hui, Justin Cranshaw, Yasmine Kotturi, and Chinmay Kulkarni. 2019. The Future of Work (places) Creating a Sense of Place for On-demand Work. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 487–491.
[39] Kazushi Ikeda and Keiichiro Hoashi. 2017. Crowdsourcing GO: Effect of worker situation on mobile crowdsourcing performance. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1142–1153.
[40] Shagun Jhaver, Justin Cranshaw, and Scott Counts. 2019. Measuring professional skill development in US cities using internet search queries. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 267–277.
[41] Thongchai Kaewkiriya. 2015. Design of Framework for Students Recommendation System in Information Technology Skills. In *International Conference on Human Interface and the Management of Information*. Springer, 109–117.
[42] Kyriaki Kalimeri and Ingvar Tjostheim. 2020. Artificial Intelligence and Concerns About the Future: A Case Study in Norway. In *International Conference on Human-Computer Interaction*. Springer, 273–284.
[43] Lawrence F Katz and Robert A Margo. 2014. Technical change and the relative demand for skilled labor: The united states in historical perspective. In *Human capital in history: The American record*. University of Chicago Press, 15–57.
[44] Adnan Q Khan and Steven F Lehrer. 2013. The impact of social networks on labour market outcomes: New evidence from cape breton. *Canadian Public Policy* 39, Supplement 1 (2013), S1–S24.
[45] Zachary Kilhoffer. 2020. REPORT ON HOW TO IDENTIFY AND COMPARE NEWLY EMERGING OCCUPATIONS AND THEIR SKILL REQUIREMENTS. (2020).
[46] Johanna Kluge, Julian Hildebrandt, and Martina Ziefle. 2019. The Golden Age of Silver Workers?. In *International Conference on Human-Computer Interaction*. Springer, 520–532.
[47] Tanja Koch, Charlene Gerber, and Jeremias J De Klerk. 2018. The impact of social media on recruitment: Are you LinkedIn? (2018).
[48] Yubo Kou and Colin M Gray. 2018. " What do you recommend a complete beginner like me to practice?" Professional Self-Disclosure in an Online Community. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–24.
[49] Yubo Kou and Colin M Gray. 2018. Towards professionalization in an online community of emerging occupation: Discourses among UX practitioners. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*. 322–334.
[50] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, Yongcheol Shin, et al. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of econometrics* 54, 1-3 (1992), 159–178.
[51] Airi Lampinen, Victoria Bellotti, Coye Cheshire, and Mary Gray. 2016. CSCW and theSharing Economy: The Future of Platforms as Sites of Work Collaboration and Trust. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 491–497.
[52] Airi Lampinen, Victoria Bellotti, Andrés Monroy-Hernández, Coye Cheshire, and Alexandra Samuel. 2015. Studying the" Sharing Economy" Perspectives to Peer-to-Peer Exchange. In *Proceedings of the 18th ACM conference companion on computer supported cooperative work & social computing*. 117–121.
[53] Airi Lampinen, Christoph Lutz, Gemma Newlands, Ann Light, and Nicole Immorlica. 2018. Power struggles in the digital economy: platforms, workers, and markets. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–423.
[54] Eugene Leypunskiy, Emre Kıcıman, Mili Shah, Olivia J Walch, Andrey Rzhetsky, Aaron R Dinner, and Michael J Rust. 2018. Geographically resolved rhythms in twitter use reveal social pressures on daily activity patterns. *Current Biology* 28, 23 (2018), 3763–3775.
[55] Lingfei Li, Yezheng Liu, Qing Zhou, Wei Yang, and Jiahang Yuan. 2020. Targeted influence maximization under a multifactor-based information propagation model. *Information Sciences* 519 (2020), 124–140.

[56] Feng Mai, Zihan Chen, and Aron Lindberg. 2019. Does Sleep Deprivation Cause Online Incivility? Evidence from a Natural Experiment. (2019).
[57] James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko, and Saurabh Sanghvi. 2017. Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Institute* 150 (2017).
[58] Jennifer Marlow and Laura Dabbish. 2014. From rookie to all-star: professional development in a graphic design social networking site. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 922–933.
[59] Andrew McAfee and Erik Brynjolfsson. 2016. Human work in the robotic future: Policy for the age of automation. *Foreign Affairs* 95, 4 (2016), 139–150.
[60] Judy McGregor, David Tweed, and Richard Pech. 2004. Human capital in the new economy: devil's bargain? *Journal of Intellectual Capital* (2004).
[61] Phoebe V Moore. 2019. OSH and the future of work: benefits and risks of artificial intelligence tools in workplaces. In *International Conference on Human-Computer Interaction*. Springer, 292–315.
[62] Rizwan Mushtaq. 2011. Augmented dickey fuller test. (2011).
[63] Naomi Nagata and Tomofumi Uetake. 2018. An e-Learning System Using Gamification to Support Preliminary Learning for Job Hunting. In *International Conference on Learning and Collaboration Technologies*. Springer, 173–184.
[64] James Ness et al. 2020. Technology in education. In *International Conference on Human-Computer Interaction*. Springer, 574–585.
[65] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[66] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 370–386.
[67] Yiming Pan, Xuefeng Peng, Tianran Hu, and Jiebo Luo. 2017. Understanding what affects career progression using linkedin and twitter data. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2047–2055.
[68] Sirous Panahi, Jason Watson, and Helen Partridge. 2012. Social media and tacit knowledge sharing: Developing a conceptual model. *World academy of science, engineering and technology* 64 (2012), 1095–1102.
[69] Maria Papoutsoglou, Apostolos Ampatzoglou, Nikolaos Mittas, and Lefteris Angelis. 2019. Extracting Knowledge from on-line Sources for Software Engineering Labor Market: A Mapping Study. *IEEE Access* 7 (2019), 157595–157613.
[70] L'udmila Pavliková and Marianna Siničáková. 2012. Labor market indicators and their causalities: the case of the new European Union member states. *Procedia Economics and Finance* 3 (2012), 1229–1237.
[71] Ross C Phillips and Denise Gorse. 2017. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 1–7.
[72] Thomas Puschmann and Rainer Alt. 2016. Sharing economy. *Business & Information Systems Engineering* 58, 1 (2016), 93–99.
[73] Alberto Rivas, Pablo Chamoso, Alfonso González-Briones, Roberto Casado-Vara, and Juan Manuel Corchado. 2019. Hybrid job offer recommender system in a social network. *Expert Systems* 36, 4 (2019), e12416.
[74] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. 2014. Uncovering the structure and temporal dynamics of information propagation. *Network Science* 2, 1 (2014), 26–65.
[75] Veronica Rossano, Rosa Lanzilotti, and Teresa Roselli. 2020. A Simulation Game to Acquire Skills on Industry 4.0. In *International Conference on Human-Computer Interaction*. Springer, 730–738.
[76] David Rotman. 2013. How technology is destroying jobs. *Technology Review* 16, 4 (2013), 28–35.
[77] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 513–522.
[78] Koustuv Saha, Manikanta D Reddy, Stephen Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[79] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, Vol. 2018. NIH Public Access, 320.
[80] KOUSTUV SAHA, ASRA YOUSUF, LOUIS HICKMAN, PRANSHU GUPTA, LOUIS TAY, and MUNMUN DE CHOUDHURY. 2021. A Social Media Study on Demographic Differences in Perceived Job Satisfaction. (2021).
[81] Rukman Senanayake, Grit Denker, and Patrick Lincoln. 2018. bRIGHT–workstations of the future and leveraging contextual models. In *International Conference on Human Interface and the Management of Information*. Springer,

346–357.

[82] Sanket S Sharma and Munmun De Choudhury. 2015. Measuring and characterizing nutritional information of food and ingestion content in instagram. In *Proceedings of the 24th International Conference on World Wide Web*. 115–116.

[83] Dehua Shen, Andrew Urquhart, and Pengfei Wang. 2019. Does twitter predict Bitcoin? *Economics Letters* 174 (2019), 118–122.

[84] Sheng-Pao Shih, James J Jiang, Gary Klein, and Eric Wang. 2013. Job burnout of the information technology worker: Work exhaustion, depersonalization, and personal accomplishment. *Information & Management* 50, 7 (2013), 582–589.

[85] Elisa Margareth Sibarani and Simon Scerri. 2020. Generating an evolving skills network from job adverts for high-demand skillset discovery. In *International Conference on Web Information Systems Engineering*. Springer, 441–457.

[86] Elisa Margareth Sibarani, Simon Scerri, Camilo Morales, Sören Auer, and Diego Collarana. 2017. Ontology-guided job market demand analysis: a cross-sectional study for the data science field. In *Proceedings of the 13th International Conference on Semantic Systems*. 25–32.

[87] Andrew Stewart and Jim Stanford. 2017. Regulating work in the gig economy: What are the options? *The Economic and Labour Relations Review* 28, 3 (2017), 420–437.

[88] Andrew Sun, Michael Lachanski, and Frank J Fabozzi. 2016. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis* 48 (2016), 272–281.

[89] Fenglian Wang, Mingqing Hu, and Min Zhu. 2020. Threat or Opportunity–Analysis of the Impact of Artificial Intelligence on Future Employment. In *International Conference on Human-Computer Interaction*. Springer, 296–308.

[90] Barry Wellman, Anabel Quan Haase, James Witte, and Keith Hampton. 2001. Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment. *American behavioral scientist* 45, 3 (2001), 436–455.

[91] Jue Wu, Junyi Ma, Yasha Wang, and Jiangtao Wang. 2021. Understanding and Predicting the Burst of Burnout via Social Media. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.

[92] Xunxian Wu, Tong Xu, Hengshu Zhu, Le Zhang, Enhong Chen, and Hui Xiong. 2019. Trend-Aware Tensor Factorization for Job Skill Demand Analysis.. In *IJCAI*. 3891–3897.

[93] Tong Xu, Hengshu Zhu, Chen Zhu, Pan Li, and Hui Xiong. 2017. Measuring the popularity of job skills in recruitment market: A multi-criteria approach. *arXiv preprint arXiv:1712.03087* (2017).

[94] Chen Zhu, Hengshu Zhu, Hui Xiong, Pengliang Ding, and Fang Xie. 2016. Recruitment market trend analysis with sequential latent variable models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 383–392.

APPENDIX

A LIST OF ESCO SKILLS

Here, we present a list of all the skills that were used in this study in Reddit and job posting datasets. Additionally, we report the distribution of each keyword in percentage, which is the number of occurrence of each skill divided by the total number of all skills in each dataset.

Table 8. List of ESCO skills and their distribution.

Skill	Reddit	Job Postings
3d modeling	0.75	0.04
adobe illustrator	0.12	0.02
agile	0.37	8.89
angular	1.58	0.96
ansible	0.32	0.14
artificial intelligence	0.23	0.02
c#	3.38	2.52
c++	4.27	0.99
computer vision	0.17	0.03
css	4.40	1.52
data analysis	0.37	1.30
database	7.42	12.88
datamining	0.16	0.32
deep learning	0.31	0.02
devops	0.43	1.30
django	1.37	0.08
docker	0.88	0.19
elasticsearch	0.14	0.10
git	4.12	0.97
googlecloud	0.16	0.06
iot	0.30	0.18
java	7.65	7.94
javascript	5.41	3.56
joomla	0.20	0.02
jquery	1.40	1.13
linux	13.86	4.69
machine learning	1.17	0.30
matlab	1.20	0.15
mongodb	0.42	0.28
mysql	1.89	0.88
nosql	0.15	1.01
objective-c	0.31	0.19
photoshop	2.96	0.28
php	4.23	0.91
postgresql	0.30	0.17
project management	0.73	9.37
python	11.83	1.60
reactjs	0.17	0.05
robotics	1.08	0.13
ruby	1.55	0.75
sass	0.29	0.13
scala	0.37	0.21
scrum	0.15	2.37
software development	0.87	8.82
sql	4.13	13.10
sql server	0.92	4.20
swift	1.76	0.20
tableau	0.34	0.50
typescript	0.24	0.02
unix	0.61	3.23
vagrant	0.25	0.03
web design	1.23	0.49
web programming	0.10	0.12
xcode	0.75	0.09

B TEST FOR DATA EXTRACTION BIAS

The following table shows the common skills in Dataset 1 and Dataset 2. Additionally, we illustrate the maximum CCF between Reddit skills and job postings with their corresponding shift time in both datasets. The CCF values in Dataset 1 that are greater than or equal to the CCF in Dataset 2 are marked as bold.

Table 9. Cross-correlation coefficients and sift times of Dataset 1 and Dataset 2.

Skill	DS1-coefficient	DS1-maxLag	DS2-coefficient	DS2-maxLag
docker	**0.29**	-1	-1	0.29
web design	0.32	-2	-2	0.35
django	**0.33**	-4	-13	0.15
swift	**0.41**	-5	-1	0.40
ruby	0.25	-12	-9	0.36
css	**0.36**	-2	-3	0.34
tableau	0.28	-4	-1	0.35
joomla	0.37	-1	-3	0.43
machine learning	**0.30**	-8	-1	0.30
scala	**0.35**	-2	-5	0.32
python	**0.24**	-5	-10	0.22
database	**0.21**	-2	-13	0.21
postgresql	**0.34**	-13	-12	0.34
photoshop	0.19	-9	-3	0.34
javascript	**0.33**	-3	-2	0.30
mysql	0.22	-4	-6	0.39
reactjs	0.50	-17	-14	0.60
devops	**0.45**	-5	-12	0.22
3d modeling	**0.44**	-5	-4	0.36
ansible	0.22	-9	-2	0.38
agile	0.26	-14	-1	0.41
jquery	0.32	-5	-1	0.35
sql server	**0.32**	-3	-3	0.20
git	**0.30**	-5	-7	0.25
robotics	0.34	-1	-2	0.47
php	0.23	-4	-8	0.31
unix	0.28	-11	-1	0.39
matlab	**0.37**	-18	-11	0.36
c#	**0.50**	-2	-9	0.31
java	**0.43**	-4	-3	0.27
linux	0.26	-7	-11	0.35
computer vision	**0.35**	-14	-14	0.30

C TEST FOR CORRELATION DURATION

Here, we demonstrate the 4-week correlation between social content and job postings in addition to the correlation at the best shift time. As it is shown in this Table, if the best shift time is more than 4 weeks, the data will not be available for 4-week correlation for some weeks. For example, for the skill "DevOps" because the best shift time is -5, therefore we cannot calculate the 4-week correlation in week 33-36. To distinguish the weeks without data for correlation, we mark these cells with gray color.

Received October 2020; revised April 2021; accepted July 2021

Table 10. Correlations between Reddit and job postings for each skill at the best shift time and over 4-week periods.

Skill	Best shift time	Correlation at the best shift time	Correlation over 4-week time periods								
			week 1-4	week 5-8	week 9-12	week 13-16	week 17-20	week 21-24	week 25-28	week 29-32	week 33-36
joomla	-1	0.37	0.73	0.95	0.93	0.57	0.51	0.72	0.67	0.27	0.10
mysql	-4	0.22	0.17	0.80	0.86	0.08	0.51	0.61	0.79	0.24	1.00
tableau	-4	0.29	0.60	0.35	0.68	0.87	0.06	0.38	0.08	0.40	1.00
devops	-5	0.46	0.76	0.48	0.68	0.38	0.89	0.60	0.27	0.80	
sql server	-3	0.32	0.81	0.93	0.37	0.28	0.08	0.24	0.63	0.80	0.58
postgresql	-13	0.35	0.84	0.52	0.78	0.37	0.62	0.75			
sass	-7	0.31	0.88	0.80	0.97	0.95	0.13	0.19	0.99	0.97	
java	-4	0.43	0.51	0.89	0.42	0.59	0.78	0.03	0.16	0.92	1.00
Web design	-1	0.33	0.14	0.92	0.60	0.31	0.50	0.65	0.89	0.81	0.16
web programming	-13	0.35	0.14	0.86	0.02	0.19	0.88	0.95			
angular	-3	0.36	0.49	0.24	0.97	0.66	0.61	0.67	0.28	0.93	0.67
vagrant	-1	0.33	0.93	0.45	0.86	0.37	0.72	0.47	0.59	0.35	0.80
javascript	-3	0.33	0.28	0.89	0.53	0.95	0.47	0.56	0.49	0.61	0.01
iot	-12	0.46	0.86	0.03	0.92	0.04	0.03	0.55	1.00		
xcode	-4	0.39	0.97	0.97	0.33	0.83	0.70	0.70	0.34	0.37	1.00
python	-5	0.25	0.16	0.25	0.59	0.78	0.72	0.11	0.05	0.81	
robotics	-1	0.35	0.40	0.12	0.74	0.95	0.75	0.91	0.57	0.83	0.72
php	-4	0.23	0.32	0.89	0.68	0.29	0.51	0.73	0.25	0.20	1.00
scrum	-5	0.23	0.53	0.83	0.64	0.38	0.37	0.28	0.58	0.41	
agile	-14	0.27	0.77	0.86	0.14	0.25	0.90	0.51			
datamining	-1	0.34	0.73	0.98	0.05	0.78	0.07	1.00	0.22	007	0.64
css	-2	0.36	0.32	0.56	0.82	0.45	0.54	0.91	0.13	0.63	0.05
matlab	-18	0.37	0.47	0.79	0.98	0.65	0.92				
jquery	-5	0.32	0.68	0.16	0.64	0.03	0.90	0.77	0.88	0.03	
c#	-2	0.51	0.70	0.96	0.89	0.17	0.02	0.28	0.82	0.84	0.82
sql	-15	0.25	0.49	0.03	0.37	0.17	0.75	0.77			
nosql	-3	0.34	0.27	0.36	0.82	0.78	0.40	0.66	0.78	0.73	0.37
data analysis	-2	0.29	0.48	0.91	0.76	0.78	0.81	0.93	0.07	0.30	0.02
ruby	-12	0.25	0.60	0.88	0.46	0.06	0.90	0.46	1.00		
django	-4	0.34	0.34	0.98	0.93	0.02	0.36	0.62	0.66	0.97	1.00
deep learning	-6	0.42	0.94	0.93	0.70	0.10	0.21	0.30	0.38	0.72	
machine learning	-1	0.30	1.00	0.68	0.60	0.98	0.99	0.89	0.74	0.00	0.57
artificial intelligence	-13	0.33	0.38	0.66	0.92	0.67	1.00	0.79			
project management	-7	0.28	0.61	0.45	0.50	0.89	0.76	0.06	0.36	0.36	
docker	-1	0.30	0.57	0.94	0.50	0.89	0.59	0.10	0.91	0.13	0.68
reactjs	-17	0.51	0.00	0.30	0.33	0.96	0.90				
linux	-7	0.26	0.71	0.26	0.60	0.91	0.46	1.00	0.94	0.05	
objective-c	-6	0.36	0.53	0.86	0.16	0.26	0.02	0.95	0.19	0.44	
typescript	-2	0.36	0.65	0.21	0.96	0.12	0.12	0.52	0.64	0.98	0.02
photoshop	-9	0.19	0.04	0.24	0.70	0.49	0.04	0.09	0.53		
swift	-5	0.42	0.95	0.85	0.69	0.55	0.30	0.56	0.17	0.77	
unix	-11	0.28	0.86	0.33	0.69	0.33	0.47	0.06	0.85		
scala	-2	0.36	0.90	0.95	0.32	0.02	0.73	0.89	0.68	0.78	0.31
3d modeling	-5	0.45	0.89	0.10	0.47	0.18	0.18	0.92	0.58	0.97	
computer vision	-14	0.35	0.29	0.76	0.42	0.95	0.30	0.57			
database	-2	0.21	0.83	0.37	0.89	0.07	0.04	0.27	0.24	0.30	0.37
software development	-8	0.24	0.71	0.16	0.85	0.79	0.69	0.07	0.57	1.00	
cyber security	-3	0.27	0.32	0.71	0.41	0.90	0.39	0.52	0.24	0.51	0.95
ansible	-9	0.22	0.85	0.52	0.13	0.28	0.09	0.92	0.36		
git	-5	0.30	0.38	0.36	1.00	0.69	0.17	0.09	0.71	0.89	
elasticsearch	-14	0.25	0.65	0.06	0.34	0.94	0.94	0.61			
googlecloud	-4	0.41	0.31	0.62	0.93	0.56	0.55	0.67	0.60	0.92	1.00
adobe illustrator	-16	0.35	0.56	0.02	0.85	0.98	0.78	1.00			