Exploring the Utility of Social Content for Understanding Future In-Demand Skills

JALEHSADAT MAHDAVIMOGHADDAM, Ryerson University, Canada
AYUSH BAHUGUNA, Royal Bank of Canada, Canada
EBRAHIM BAGHERI, Ryerson University, Canada

Rapid technological innovations, especially in the information technology space, demand the workforce to be vigilant by acquiring new skills to remain relevant and employable. The workforce needs to be engaged in a continuous lifelong learning process by educating themselves about skills that will be in demand in the future. To do so, it is essential for students, job seekers, and even recruiters to know which skills will be in demand in the future and to invest time and resources in developing these skills. On this basis, the main objective of this paper is to investigate whether social content can offer insight into potential future in-demand skills in the IT job market. Based on the analysis of social content from Reddit and job posting data from Dice and Monster websites, we find that social content related to job skills is a strong indicator for future in-demand skills. We further find that specific social content associated with *recruitment-related* topics are stronger indicators of future skills. Our findings encourage learners and job seekers to pay close attention to online social content to strategically plan new skills and maximize their employability.

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

Additional Key Words and Phrases: lifelong learning, future of work, text classification, user engagement, social media, Reddit

ACM Reference Format:
Jalehsadat Mahdavimoghaddam, Ayush Bahuguna, and Ebrahim Bagheri. 2022. Exploring the Utility of Social Content for Understanding Future In-Demand Skills. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 389 (November 2022), 35 pages. https://doi.org/10.1145/3555114

1 INTRODUCTION

The advancement of technology, such as the recent transformational developments in Artificial Intelligence (AI) and the growth in the amount of accessible data, has significantly impacted the labor market and the workforce [51, 93]. Existing research on the labor market has shown that between 10-50% of the current active jobs are at risk of automation [47]. This means that the skills associated with such jobs will become increasingly outdated and new skills are required for the workforce to be able to stay relevant and employed. The rapid evolution and changes of in-demand skills necessitate constant lifelong learning [29, 100]. Therefore, obtaining new knowledge and acquiring relevant skills have become essential requirements for the workforce to keep up with industrial changes [50, 87].

The literature has shown that the speed of the evolution of job requirements leads to ambiguity about the skills that are immediately in demand by employers, which makes skill acquisition and

Authors' addresses: Jalehsadat Mahdavimoghaddam, Jmahdavi@ryerson.ca, Ryerson University, Toronto, Ontario, Canada; Ayush Bahuguna, Royal Bank of Canada, Toronto, Ontario, Canada, ayush.bahuguna@rbc.com; Ebrahim Bagheri, Ryerson University, Toronto, Ontario, Canada, bagheri@ryerson.ca.

2573-0142/2022/11-ART389 $15.00
https://doi.org/10.1145/3555114

job search challenging for both learners and job seekers. This in turn causes anxiety and exhaustion in the workforce [73, 106, 112]. For this reason, it is imperative for the current and future workforce to have sufficient insight into the changes in the job market and new in-demand skills to help them with continuous skill development and remain relevant and employable

To assist job seekers with obtaining in-demand skills, existing literature has utilized historical recruitment data to identify the popularity of skills in a fast-paced job market [124, 125, 129]. Such studies mainly focus on finding patterns of skill trendiness using machine learning techniques. While these studies can help to analyze the job market, they have limitations in investigating possible relationships between the variables that impact the labor market. For example, these studies overlook the role of social media content in understanding the labor market. There is little work within the Computer-Supported Cooperative Work (CSCW) and Human-Computer Interaction (HCI) communities that has shown the importance of social media in understanding and learning the skills required in emerging jobs [72, 88]. There is still a clear methodological gap in exploring the possible congruence between social content and in-demand skill shift patterns.

To close such a gap, the focus of this paper is to investigate whether *social media content* can be leveraged to understand how future in-demand skills will evolve and whether certain types of social content are more indicative of future skill demand. Utilizing *social content* will help us access large, inexpensive, and near real-time data that contain self-disclosure of personal and industrial experiences and opinions that enable us to study future skill demands. More concretely, and to understand the role of social content in understanding the skill demand shift in the job market, we leverage two data sources, namely Reddit and online public job postings.

This paper focuses on the following three main Research Questions (RQs):

RQ1: Would it be possible to consider online social media content, specifically Reddit content, to develop a realistic understanding of future in-demand skills in the IT job market?

In this first research question, our main objective is to explore how social content can help us understand the trends in future in-demand skills that will appear in future job postings. For this purpose, we study how the frequency of the skills mentioned in Reddit content correlates with skills mentioned in online job postings at different time intervals. We establish that social content observed on Reddit is a practical indicator of how in-demand skills will appear in the future.

RQ2: Are there special types of content on social media that are more indicative of future in-demand skills and would be most appropriate when analyzing skill trends?

To study the findings of RQ1 with further granularity, in the second research question, we adopt a Grounded theory approach [116] to classify social content into suitable topical categories. These categories include topics such as *technical advice*, *career advice*, and *recruitment*, among others. We then classify our social content into these topics using an accurate AI-based deep learning classifier. This allows us to study whether certain online social content types are more indicative of future in-demand skills. We find that the content related to the *recruitment* and *technical advice* topics captures the changes in future in-demand skills better than other topics.

RQ3: Are measures of "user engagement with social content" a strong indicator of content utility when analyzing social content to understand future skills?

In this research question, we explore whether social content that has received different degrees of engagement from the community has different degrees of suitability for our purpose. Our findings suggest that the consideration of the posts that receive a high volume of upvotes leads to a better understanding of the in-demand skills in the IT job market.

This paper is among the first evidence-based works that demonstrate how social content can be a source of phenomenological data to characterize changes and shifts in future in-demand skills. This work provides theoretical implications and practical insights that we discuss in detail in the paper:

- *Theoretically*, instead of studying the future of the job market in isolation by only relying on its history, this study examines this phenomenon through an exogenous data source, i.e., social content, that has often been overlooked in prior labor market research.
- *Methodologically*, using Grounded theory and cutting-edge deep learning classifiers, we model social content posted in a variety of sub-communities on Reddit to further analyze the relationship between social content and future in-demand skills in the job market.
- *Practically*, we illustrate that online social content can help us understand the future of skill demand. Based on the findings of this paper, lifelong learners can 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 by continuously monitoring and engaging with relevant social content and communities.

The rest of this paper is organized as follows. The next section reviews the related literature. In Section 3, we describe our datasets and methodology. We report our findings related to each of our three research questions in Section 4. In Section 5, we elaborate on our findings and discuss the implications of our findings. Finally, we conclude the paper by presenting areas for future work.

2 FUTURE SKILL DEMAND AND PROFESSIONAL DEVELOPMENT

The fourth industrial revolution (Industry 4.0) and the development of new economic models, such as the gig economy [117] and the sharing economy [78, 99] have profoundly impacted the nature of work [46, 77]. As such, a growing body of literature investigates new forms of occupations in the evolving labor market [19, 33]. Many of these works touch on various aspects of digital labor and how it facilitates on-demand work, whenever and wherever, through web-based software applications, such as Uber and Lyft. While such jobs provide workers with time flexibility and an additional source of income, the majority of them do not require a high level of preparation or a diverse set of skills [14, 53]. Therefore, attracting and retaining low-skilled workers who are in turn impacted by the lack of professional growth and job progression in the long run.

Similarly, there has been growing interest among CSCW scholars to explore crowdsourcing and digital work [23, 60]. The majority of these studies adopt qualitative research methodologies, such as survey and interview analysis, to study the labor market [15, 39, 40]. For example, Rivera and Lee investigated career development obstacles among gig workers by surveying 20 Amazon Mechanical Turk workers [102]. Similarly, through qualitative research, Yao et al. and Seetheraman et al. studied online community support and social isolation among gig workers, respectively [109, 128]. There are other CSCW scholars who have adopted quantitative approaches, e.g., to investigate whether on-demand jobs can increase the chance of employment for low-skill workers with a gap in their employment history [80] or explore how users seek career advice on online Q/A forums [119]. While quantitative approaches have been gaining increasing attention over the past few years [96], there has not been much work on how online social data can serve as indicators for future skill demand. This is despite the fact that a strand of literature has already leveraged social data to study and model a variety of social phenomena, such as hate speech [94], nutritional choices [32], mental health [44, 107], physical activities [4], and other social behaviors [85]. For example, in the financial context, Phillips and Gorse [98] have explored future cryptocurrency movements by applying a Hidden Markov model on variables extracted from Reddit, such as the volume of posts and the number of new subscribers. Similarly, Sun et al. [118] have examined the relationship between the stock market and user-generated content by calculating the correlation between stock prices and Twitter activities. Additionally, by using the Granger Causality test, Shen et al. [110] have shown that the number of tweets that contain the word “bitcoin” can help predict future bitcoin volumes. Closer to the theme of the current paper, Wu et al. [123] have studied job burnout, which is an occupational phenomenon, via a social medium called Weibo. Similarly, Saha et al. [108] have explored the relationship between job satisfaction and race, sex, and geographical location using Twitter data.

With regards to employment, there are few existing works that explore the impact of social media on career development, especially for finding the required skills for emerging occupations, such as cloud architectures [68, 88]. These types of studies are timely and important because the transformation of the labor market heavily impacts existing job requirements through the emergence of new skills. This obliges the workforce to acquire new skills to prevent skill gaps that can lead to unemployment [68]. For this reason, learning has turned into an intense lifelong process [65]. Similar to Toffler who proclaimed *"the illiterate of the future will not be the person who cannot read. It will be the person who does not know how to learn"*, Ra et al. [100] emphasize the importance of the ability to learn, including the desire to re-learn and unlearn, over just obtaining new knowledge. In other words, learnability and being able to predict future in-demand skills are required commodities of the new economy [50]. To support students and the workforce to improve their learnability, various researchers have developed tools that provide insights into individuals' learning skills [20, 21, 26] and motivate self-regulated learning [63, 100].

While employers increasingly seek higher skilled workers [27, 66], the pressure to constantly learn and acquire new skills can have an immense psychological burden on the workforce with an increase in the level of anxiety and uncertainty [56, 89]. According to a study by Kluge et al. [70], among different age groups, the younger workforce is at a higher risk of stress and anxiety due to such expectations. Hence, a growing number of research studies are dedicated to investigating occupational changes and new in-demand skills in the knowledge economy to assist job seekers and learners with their professional development [73, 88]. For instance, there has been an effort to develop a variety of tools to recommend skills that are required in the job market to students [30, 92]. To this end, prior work has leveraged different data sources and methodologies. For example, Kaewkiriya [64] gathered variables, such as interpersonal and logical-mathematical intelligence abilities, that are determinants for performing different jobs. Then, they obtained student strengths and weaknesses via manually designed questionnaires and employed recommendation algorithms to identify and suggest customized skills that needed to be acquired by the individual. Similarly, Dave et al. [30] developed a skill recommender technique based on a latent embedding method by utilizing resumes from the CareerBuilder website.

CSCW scholars take a step beyond identifying the skills that individuals need to obtain and propose new tools and ideas that might affect skill development. For example, Dillahunt et al. [41] propose a tool that makes information about employment and volunteering opportunities publicly available in order to help candidates be informed of required skills. The purpose of such tools is to allow the workforce to understand future in-demand skills and help them take steps towards acquiring them. Marlow and Dabbish [88] observed that the interactions that users experience on a community-based social media dedicated to graphic design, called Dribbble, had a favorable influence on their professional growth. Professional development happens for users on Dribbble by viewing other designers' works and getting inspired by them. Additionally, when users see a design technique, they will try to learn and apply it independently, which leads to skill development. Most importantly, Dribbble enables users to find cutting-edge techniques in the industry, which helps the users keep up with the changes in the graphic design market. Moreover, Kou and Gray [72] investigated Reddit, especially the "r/userexperience," subreddit where people contribute expertise and information about UX design. They looked into what UX experts share about their knowledge and how self-disclosure aids professional communication. Finally, they demonstrated that disclosing posts receive more comments from the experts, which also contain disclosed information, leading to professional progress.

Besides helping lifelong learners with skill development, it is important to systematically explore *the shift in skill demand* in the job market. This will help to understand the upcoming changes in the industry and enable the workforce, educational institutions, policy-makers, and operational managers to proactively respond to the upcoming changes in the labor market. Despite the significance of identifying the evolution of skill demand, only a few studies examine this phenomenon. For instance, a few researchers capture shifts in job market demand by applying unsupervised learning algorithms to historical job postings [125, 129]. Other researchers analyze and recommend future required skills by only relying on historical data obtained from recruitment platforms [113, 114, 124]. One of the areas where the literature can be improved is the consideration of the role of social content in understanding future in-demand skills. Therefore, this paper expands on previous research on professional and skills development by studying how social content can enable us to examine and understand the evolution of future in-demand skills in the IT sector through monitoring and engaging with online social content.

To bridge this gap, in our earlier work [84], we studied possible relationships between social content and future in-demand IT skills. Our first important finding suggested a strong temporal alignment between user-generated data and future requirements of IT occupations. Hence suggesting that social content posted on community-based platforms, such as Reddit, can serve as strong indicators of the future in-demand skills represented in job postings. Further findings indicated that a causality framework, such as Granger-causality, could be beneficial for capturing the linear functional connectivity between Reddit content and job postings. In other words, online social content can Granger-cause future changes in the labor market and can predict future job requirements better than just using historical job demands. Furthermore, our results showed that by using historical data, we can predict the lag at which the maximum correlation between Reddit and job posting time series occurs. This current paper builds on our previous work [84] on professional and skill development by studying ***how*** social content can enable us to examine and understand the evolution of future in-demand skills in the IT sector through monitoring and engaging with online social content.

Our main contribution to the CSCW community is to show that social content has the potential to provide actionable insight into the changes in the skill demand of the job market. In addition to identifying the future in-demand skills, we show that social data can be used to answer questions about the labor market and the future of work. Through the lens of social data, more researchers in the CSCW community will be able to conduct quantitative research to systematically explore different aspects of the future of work. For example, social media can be used to track social isolation among crowdworkers and identify the support that they receive from the community. In terms of skill development, there are important studies reported by CSCW scholars that explore the impact of online social communities on skill development and how individuals leverage their social networks to gain new skills [72, 73]. We consider these works as vital building blocks for gaining insight into the role of social media for professional development in the knowledge economy, and hope that our work contributes to this foundation. Additionally, it is our hope that the insights that we provide in this paper on how individuals in the IT domain consume social media and what type of content they generate can help CSCW and HCI scholars better understand the needs of knowledge workers and develop tools that address those needs.

3 STUDY DESIGN AND METHODOLOGY

3.1 Data Sources

We study the relationship between job skill demand reflected in online job postings and social content posted on Reddit. We chose Reddit as it is one of the largest online communities with more

Table 1. Examples of the CSCW and HCI work that utilize Reddit to answer a variety of questions.

Paper	Venue/Year	Domain	Data
De Choudhury et al. [31]	CHI2016	Public Health (mental health/suicide)	Health subreddits and a suicide support subreddit called "r/SuicideWatch"
Saha and De Choudhury [104]	CSCW2017	Public Health (mental health/gun violence)	Subreddit related to US colleges with gun-related violence
Saha et al. [105]	CSCW2019	Public Health (mental health/LGBTQ)	Posts from "r/lgbt subreddit"
Kou and Gray [72]	CSCW2018	Future of Work (professional development)	A user experience community named "r/userexperience"
Li et al. [81]	CSCW2020	Security and Privacy (software development/personal data)	Personal data discussions on "r/androide"
Garg et al. [49]	CSCW2021	Labor Market (unemployment, job search)	Posts and comments shared in three subreddits called "r/Unemployed", "r/GetEmployed", and "r/Jobs"
Ganesh and Lazar [48]	CSCW2021	Labor Market (workplace)	Posts from "r/migraine" and "r/fibromyalgia"

than 52 million active users per day as reported in January 2021 [https://www.redditinc.com/]. Also, Reddit has more than 100 thousand forums focused on particular topics, called subreddits. On subreddits, both specialists and nonprofessionals can share knowledge about their discipline or seek help [72], gain information by reading posts, voting them up or down (upvote/downvote), or leaving comments on posts. Upvote is a method on Reddit, similar to likes on Facebook, by which users can signal their approval or support for a post [115]. Similarly, downvotes are used to disapprove or downgrade content on a Reddit post. Subreddits have made answering a range of questions from mental health to the labor market possible for CSCW and HCI scholars. Some examples of this work are reported in Table 1, which shows the subreddits that researchers have utilized in their studies. Directly related to this research, prior works show the utility of subreddits as an effective source of understanding the required skills in emerging occupations [73]. Another reason for the popularity of Reddit among researchers is its anonymous environment that allows Redditors to generate content reflecting their genuine opinions [5]. Additionally, unlike Twitter's 140-character restriction, Reddit enables postings of up to 40,000 characters per comment [111], which provides rich content for analysis. Finally, Reddit allows seamless data collection and analysis through an official API [74] and Pushshift Reddit dataset, which made it a popular data source among researchers [9].

To gain content from Reddit and online job postings, we obtained a list of 73 IT skills from the International Standard Classification of Occupations (ISCO), which is a tool for international labor market reporting [12]. The list of these skills is included in Table 10 (in the Appendix). Based on these IT skills, we curated historical job postings and Reddit posts from October 01, 2015, to July 01, 2016. Job postings were obtained from two career websites, namely "Monster" and "Dice" and were downloaded from DataStock[1]. Reddit posts were obtained from Pushshift[2]. Finally, to maximize the search results, we performed text preprocessing by lowercasing, removing stop words, and stemming the gathered textual data.

To find the job postings that contained at least one of the ISCO skills in their job description, job requirement, and job title, we first removed duplicated job postings that were posted within 14 days of the initial posting. Using this technique, we retained 432,177 unique job postings. To create the Reddit dataset, we searched for the ISCO skills in the title or body of the English posts that did not have the tags *"deleted"* or *"removed"*. Then, we created a list containing 2,426 subreddit names

[1] https://datastock.shop/

[2] https://files.pushshift.io/reddit/

Table 2. Statistics of Reddit and job posting datasets.

Dataset	Description	Count
Job Postings	# job posts	432,177
	# unique companies	11,230
	# unique cities	5,672
	# unique states	50
Reddit	# Reddit posts	80,051
	# unique subreddits	483
	# unique authors	52,927
	# upvotes	736,857
	# comments	773,232

from the dataset. Two researchers went through the list and mapped the names of the subreddits to their descriptions to exclude the ones that were not relevant to the scope of our study. For instance, while our skills included terms such as C#, this would return content from Reddit that relates to music (and not IT skills), which is not related to our research questions. As a result, we collected 80,051 unique posts in 483 subreddits. The collected Reddit content was posted by 52,927 unique Redditers and received 736,857 upvotes in total. The statistics of the job posting and Reddit datasets are reported in Table 2.

It is important to note that while the extracted job postings from Dice and Monster are limited to the US and Canada, Reddit does not provide publicly-available geo-location information of its users. Therefore, controlling location via Reddit is not possible. However, according to a survey conducted in 2014 [17] from Reddit users, 65% of the participants were noted to be in the US, 12% in Canada, 6% in the UK, 2% in Australia, and 15% other. We refer to these survey results from 2014 as it aligns well with the dates that our data was collected as mentioned above. Based on this survey, we estimate the location of at least 77% of our social data to match with the job posting locations. While we acknowledge that there might be ≈20% location mismatch between our data sources, inferring the exact location of the users requires surveying 52,927 users in our dataset, which is practically not possible as such, we note this as a potential threat to the validity of our work.

3.2 Overall Analysis

As a part of the first research question (RQ1) and to identify the relationship between user-generated content and future in-demand skills, we construct two time series based on job postings and Reddit data. To model these time series, we first create two vectors based on the number of jobs and Reddit posts that contain at least one of the ISCO skills per week. For example, from October 4th to October 11th, 2015, we captured 2,406 Reddit posts and 23,716 job postings that consisted of IT skills. After creating all time series data points for all weeks in our dataset, we normalize the time series using the cosine normalization method, by taking the weekly data as a vector and dividing it by the square root of the sum of the squared vector elements [118]. This is a common strategy adopted by prior works to represent social content with their volume [98]. For example, Yiming et al. [95] represented user data on Twitter and Linkedin with word counts and investigated the relationship between personality traits and career progression. Similarly, in the mental health context, Dutta et al. [43] employed predictive models on Twitter word counts to identify the effect of social interactions on anxiety levels.

To ensure reliable findings from our time series analysis, it is important to apply stationary tests to detect the potential trending behavior that might lead to fabricated regressions [91]. A time series is described as stationary if its statistical properties, such as mean, variance, and covariance, are

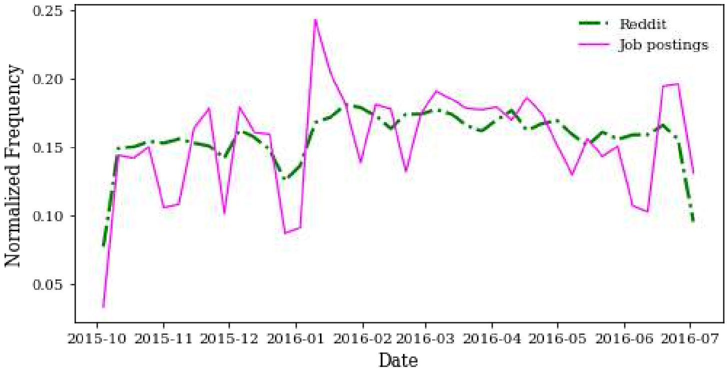


Fig. 1. Temporal behaviors of Reddit and job postings.

constant over time and unaffected by the time at which it is observed [126]. Generally, stationary time series do not exhibit a foreseeable pattern, such as any trend (e.g. upward or downward) or seasonality, in the long term [59]. If these conditions are not fulfilled, we consider the data provided by this stochastic process to come from a separate population of processes [10]. Thus, any analysis of the time series with different statistical characteristics will lead to spurious regression [55]. One of the common techniques for determining non-stationarity time series is by identifying the presence of a unit root in data, which exhibits a stochastic trend [97]. Before applying formal statistical tools, to identify the non-stationary time series characterized by a unit root, we plotted and visually analyzed our time series to detect any visible trends. These time series are shown in Figure 1. This figure illustrates that neither of our time series has a unit root and we can observe a slight upward trend in both of the time series. Therefore, we applied statistical stationarity tests before we explored any relationships between them.

3.2.1 Time Series Stationarity Tests. Similar to the previous literature [6, 44], we applied the two popular methods called Augmented Dickey Fully (ADF) test [38] and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [75] to check whether the time series possess unit root. 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. 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. By applying the ADF test, we noticed that we cannot reject the null hypothesis for Reddit: $t = -0.25$, $p = 0.59$ and job postings: $t = 0.15$, $p = 0.73$, therefore, both time series are non-stationary.

To deal with non-stationary time series, one of the well-known techniques is to use differencing $u_t = y_t - y_{t-n}$, where u_t is called the n-order difference [10]. Differencing is a specific type of filtering by which the mean of the time series will become steady over time. Thus, stationarity is achieved by removing fluctuations in the level of the time series, therefore eliminating (or decreasing) trend and seasonality [59]. Usually, first-order differencing is adequate, however, differencing should be done until the time series becomes stationary [126]. To make the time series stationary, we applied first-order differencing [44] and ran both tests on them again.

Although both time series rejected the null hypothesis after the first-order differencing, it is possible for time series to not have a unit root but be stationary around a deterministic trend [75]. Thus, we also applied the KPSS test to the time series. 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

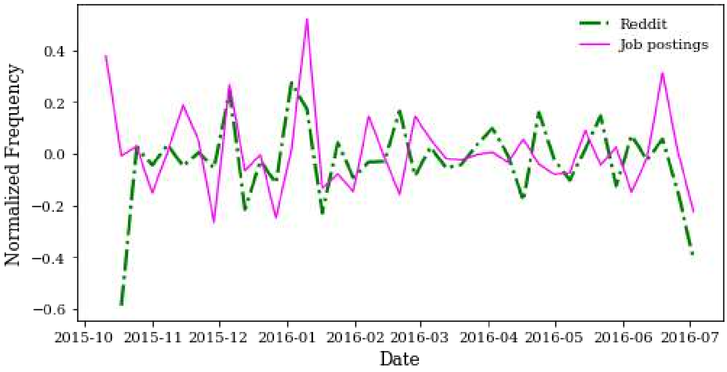


Fig. 2. Weekly temporal frequency of Reddit and job postings at time *t*

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 of job postings were greater than 5%, however, the p-value of Reddit was smaller than the significance value, 0.03. Therefore, we applied the first-order differencing on the Reddit time series and we ran the KPSS test again. After this step, as shown in Figure 2, the two time series become stationary and we cannot see any trends in the data anymore.

3.3 Topic-based Analysis

As a part of the second research question (RQ2), we are interested in understanding whether there are specific types of social content that show a higher association with future in-demand skills. For this purpose, we should identify the main themes, referred to as topics, in our social data. One of the well-known approaches for this purpose is topic modeling methods, such as Latent Dirichlet Allocation (LDA) [16]. The main benefit of topic models is that they detect the topics in an unsupervised manner, which helps save training time. However, understanding the results and verifying the quality of the discovered topics are considered to be difficult tasks [13, 58].

Another popular technique to find the corpus topics is called the Grounded theory [116]. Grounded theory is a general approach by which quantitative researchers collect and analyze data systematically. It is essentially an iterative process that involves applying comparative methods in each iteration to capture similarities and differences between incidents in the data to identify conceptual categories [24]. These tentative conceptual categories are repeatedly revised through addition, deletion, modification, or merging until they adequately represent the thematic view of the observed data [8]. According to Charmaz et al. [24], during this process, memos should be developed to capture the researchers' thoughts, ideas, and questions to avoid losing insight. Researchers showed that the generated results through the Grounded theory approach are interpretable, however, the process is time and energy-consuming, especially for large datasets [7]. Nevertheless, Baumer et al. [8] showed results obtained from a topic modeling method align well with those obtained from a Grounded Theory approach. As such, in favor of interpretability, we follow the Grounded theory process in our work.

To find the content types that represent Reddit posts, we first randomly sampled 8.5% (6,750) of our social data. Three researchers examined the sampled dataset iteratively to identify the conceptual categories. For each conceptual category, the researchers developed and recorded memos iteratively that explained the definition, the number of posts, a sample of representative content,

Table 3. The set of extracted topics, examples of Reddit titles from our dataset in each topic, and the number of posts in each topic.

Topic	Sample Title of Reddit Post	Number of Posts in Each Topic
Technical Advice	Is there any Objective-C cheat sheet for Swift Developers What should I know about SQL+Flask security?	2,196 (32.53%)
Career Advice	How to start with iOS development as a career? How are people getting internships with Google?	1,650 (24.45%)
Learning	Most efficient way to learn NLP? Where to learn to code	1,145 (17%)
Recruitment	Looking for Dev Intern - Possible Hire - London Front End Developer Needed-Cinemoz SAL	955 (14.14%)
Project/Mentor/team Search	[Looking for] C++ Mentor and / or buddies [Hobby] Experienced programmer looking for portfolio work	584 (8.65%)
Other	[Opinion] Xamarin Vs. Native Development... An *Honest* Review of the X1 Carbon 5th Gen	220 (3.25%)

and commonly repeated terms. In these memos, the similarities between categories were captured that could potentially cause overlaps between them. After the initial categorization, the conceptual categories were compared to each other and reassessed to represent the data sufficiently. For example, in the initial iterations, the researchers found categories that were in detailed descriptions of observed data. However, as the process progressed, categories were grouped to represent a broader description. When researchers could not merge more categories and new categories were not found, they stopped the process. This final set of conceptual categories formed the *topics* that are of interest to our study.

In the initial iteration of the Grounded theory process, the researchers came up with 15 categories that represent the majority of Reddit posts. The name of these topics and detailed information on each topic can be found in Table 11 (in the Appendix). After 5 iterations, the three annotators labeled the data and finalized 6 topics. A list of these topics and a few examples in each topic can be found in Table 3.

3.3.1 *Building a Machine Learning Classifier.*

To train a learning-based text classifier that can assist in labeling all posts in our dataset, we use the 8.5% (6,750 posts) of the Reddit dataset that is labeled through the Grounded theory process as the training set for machine learning-based text classifiers. We train different text classification algorithms, such as Logistic Regression [1, 54], Random forest [120], Linear Support Vector Classifier (SVC) [61] models on term frequency–inverse document frequency (TF-IDF) vectors of the documents, Bidirectional Encoder Representations from Transformers (BERT) [2, 90], DistillBERT, and XLNet [127]. In all models, we used a k-fold (k=5) cross-validation technique to evaluate our models. Using these text classifiers, we classify all posts into one of the identified topics. We provide details on the training specifications of each of these models in Appendix C.

To compare the performance of the models, we report the precision, recall, f1-score of each topic, and the mean weighted-average F1 score of the trained classifiers on all topics in Table 12. As shown, the performance metrics of BERT and DistillBERT are the highest and quite competitive with each other. As such, we select DistillBERT to classify our unlabeled dataset and we provide its performance metrics for each topic in Table 4. Figure 3 illustrates the percentage of Reddit posts on each topic after classifying the unlabeled data using the DistillBERT model. As shown in this figure, the top three topics with the most data are *"technical advice"*, *"career advice"*, and *"learning"* with 83%, 11%, and 3% posts, respectively. It is important to note that because the *"Other"* class does not represent any particular topic, we do not consider it for further analysis.

Table 4. Performance metrics of our selected topic classifier (DistillBERT) for each of the 6 topics. DistillBERT has an overall accuracy of 97.80% on our test dataset.

Classifier	Topics	Title + Text		
		Precision	**Recall**	**F1-Score**
DistilBERT (base uncased)	Other	0.92	0.95	0.94
	Technical Advice	0.99	0.98	0.99
	Career Advice	0.96	0.96	0.96
	Recruitment	0.96	0.94	0.95
	Learning	0.98	0.99	0.99
	Project/Mentor/team Search	0.92	0.95	0.94
Average Accuracy		**0.97**		

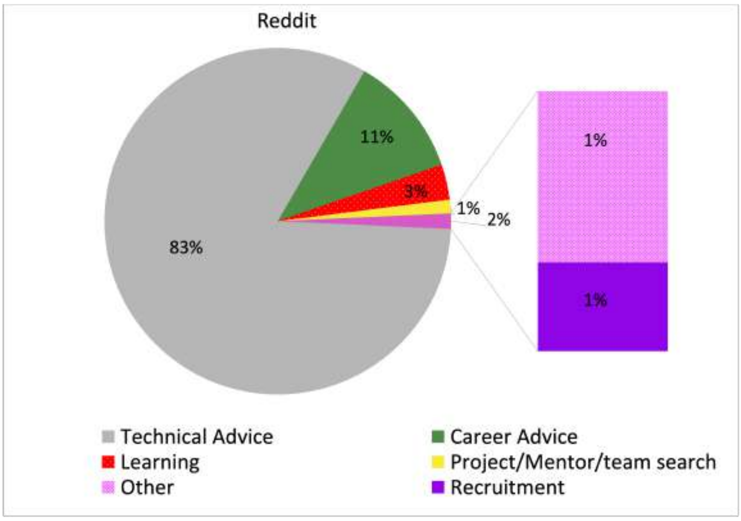
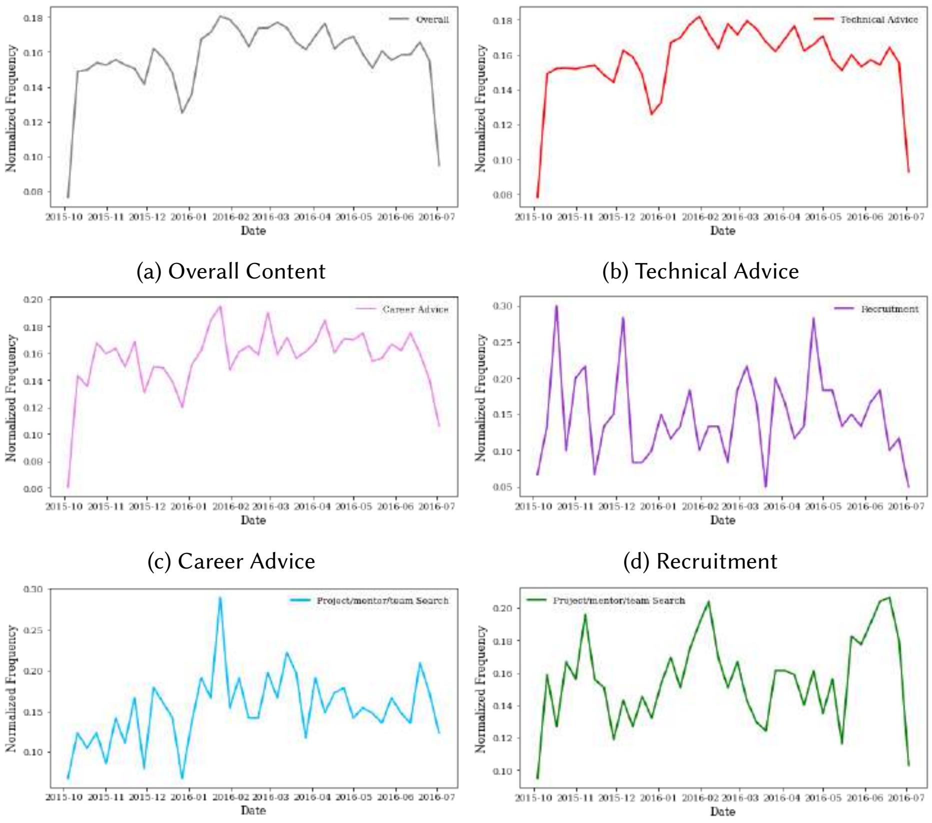


Fig. 3. Distribution of posts in each Reddit topic.

3.3.2 Topic-based Time Series Construction. To understand the impact of social content related to a particular topic on future in-demand skills, we create 5 Reddit time series based on the data on each topic. Similar to Section 3.2, we consider the weekly normalized frequency of the content containing one of the skills to build our time series. Temporal frequencies of topic-related time series are illustrated in Figure 4. We observe that each of these time series demonstrates a unique characteristic. For example, we can find that the *overall*, *technical advice*, and *career advice* time series have similar drops and peaks over time with a slight upward trend. In contrast, the *Recruitment* time series does not have any obvious trends, while the time series of discussions around *project, mentors, and teams* has a gradual increase over time. Interestingly, the observed pattern in the *Learning* topic time series is periodic with three repeated patterns, which makes this time series different from the other topics.

Figure 4 shows that the topic-related time series are not stationary and, similar to Section 3.2, we need to run the ADF and KPSS tests on them. We noticed that in all topics, after the first-order differencing, all topic-related Reddit time series passed the ADF stationarity test. After the second first-order differencing, all time series passed the KPSS test as well.



(a) Overall Content

(b) Technical Advice

(c) Career Advice

(d) Recruitment

(e) Project/Mentor/Team Search

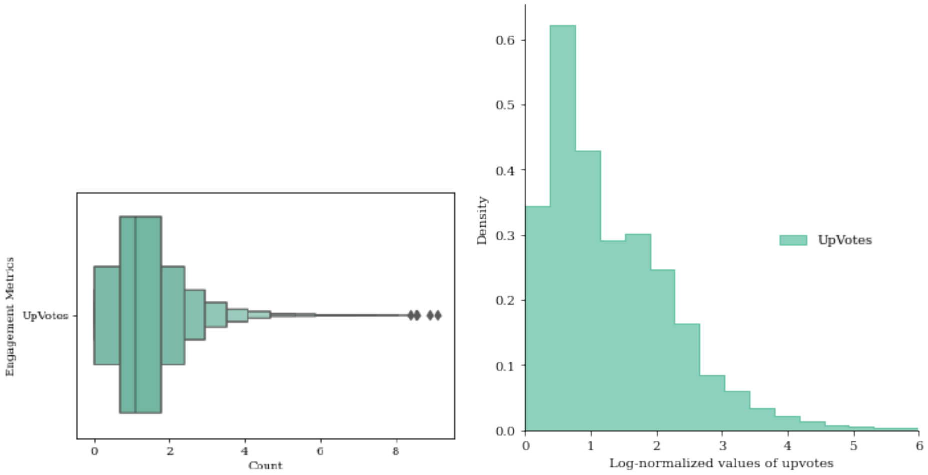
(f) Learning

Fig. 4. Temporal frequencies of topic-based time series.

3.4 Engagement-based Analysis

As part of the third research question (RQ3), we are interested in studying the effect of user engagement with social content on capturing the relationship between social Reddit content and online job postings. We define user engagement as the intensity of users' involvement with content on social media platforms [37]. Risch and Krestel [101] claim that the number of likes and upvotes indicates the popularity of content. In contrast, they argue that comments can be posted in reply to a given social content for a variety of reasons and are hence not as indicative of popularity or informativeness. For example, comments might answer someone's questions, provide feedback, correct a mistake, or express dissatisfaction or disagreement. Therefore, a higher number of comments does not indicate the popularity of a post. Therefore, similar to [101], we classify engaging posts based on the number of upvotes that they have received. To identify the posts that received different levels of engagement from the community, we follow a similar approach proposed by Aldous et al. [3] where we divide the content into two main categories, namely high and low engaging content. The high and low-engaging content are the top and bottom 33% of content sorted based on the number of upvotes they have received.

Before we examine how engaging posts relate to the skills on the job market, we plot our Reddit dataset to better visualize the degrees of engagement with content in Figure 5. As shown



(a) Box plot of upvotes for the Reddit posts. (b) Histograms of upvotes for the Reddit posts.

Fig. 5. Frequency distribution of upvotes across the Reddit dataset. The x-axis is the log-normaized value of the engagement metric.

in Figure 7b, the number of upvotes is highly skewed, thus we use the log-normalized values for our data analysis and we add one to all the upvotes to control for 0. Based on Figure 7a, we can see that 75% of the posts received less than 6 (log-norm=2) upvotes. Using 33% percentile without considering the topics of the posts, 28,806 posts are categorized as low engaging posts and they have received either 0 or 1 upvotes. In the high-engaging category, we have 25,056 posts that have received more than 7 upvotes. We will explain time series development for high and low-engaging content in the next section.

3.4.1 Engagement-based Time Series Construction. For our experiments in the third research question, we build time series for high and low engaging content overall (explained in Section 3.2) and for each of the five topics (introduced in Section 3.3). Similar to the previous time series, to ensure that all of the developed time series passed stationarity tests, we applied both ADF and KPSS tests. We realized that both high and low engaging time series based on Overall and topic-based content failed the ADF test the first time. After the first-order differencing, we could reject the null hypothesis for all time series. By applying the KPSS test, we noticed that all low engaging-based time series passed the stationarity test immediately. However, only the high engaging content related to *recruitment*, *learning*, and *project/mentor/team search* passed the KPSS test the first time. To make the Overall, *technical advice*, and *career advice* high engaging content stationary, we applied first order differencing, which made all time series stationary.

4 RESULTS AND FINDINGS

In this section, we report the results of our analysis and structure our findings around the three main research questions of this study.

4.1 RQ1: The Utility of Online Social Content for Understanding Future In-demand Skills

Our first research question is focused on studying the possible relationship between social content and future in-demand skills as represented in online job postings. The purpose of RQ1 is to explore whether social content can help the workforce to develop an understanding of the evolution of required skills in the future. To this end, we investigate any possible relationships between social content and in-demand skills and determine how well and at what point they may be correlated with one another. Methodologically speaking, we use the cross-correlation coefficient (CCF) to measure the association between the time series for social content and online job postings. Our approach is similar to previous works [79, 103] that utilized CCF to identify how two data sources are related to each other.

To measure CCF, we employ the time series developed in Section 3.2 based on the frequency of the skills mentioned on Reddit at time t against the number of times that skills appear in job postings at time $t + \tau$. Here, τ represents the shift time by which the job posting time series is moved to represent the delayed temporal alignment with the Reddit time series. To understand at which shift time the correlation between Reddit and job postings is the maximum, we look at the relationship between the two time series at time t up to 19 weeks after t ($t + 19$). The absolute CCF values of different shift times are reported in Table 5. As shown in this table, the correlation between the two time series is the highest at time $t + 0$, when the two time series are temporally aligned. However, because our purpose is to explore future trends, we are interested in the best correlation that happens in time $t + \tau$. In other words, using the maximum CCF that happens in time 0 does not help us to understand how well Reddit posts correlate with job postings in the future. Based on Figure 6, it is possible to see that at shift time $t + 5$ job postings correlate with Reddit content the best. We note that due to the job posting time series being shifted by 5 weeks, as it is shown in Figure 6, this time series ends 5 weeks sooner than the Reddit time series.

To test whether the correlation holds for other time granularities, we also measured the CCF for daily, semi-monthly (1st and 15th), and monthly time series. We created these time series by following the same steps mentioned in Section 3.2. More details on the conducted stationarity tests and the results of these experiments are included in Appendix D. By looking at the results reported in Tables 13 and 14, we can see that a relationship exists between Reddit and job postings, regardless of whether we consider the frequency of IT skills appearing on social media and recruitment data on a daily, weekly, or semi-monthly basis, and the findings are consistent across different time granularities. However, based on the size of our dataset, we chose a weekly frequency for our further analysis. Note that although for each time granularity we report the time at which the correlation is the maximum, the goal of this study is not to identify the best time that Reddit correlates with future IT skills on job postings. Instead, our purpose is to shed light on the important relationship between social content and job postings and introduce a new source of data that helps to understand the changes in the job market.

First finding: there is a meaningful relationship between the social content around in-demand skills on Reddit and the future skill demands represented in online job postings. In other words, the correlation values suggest that there is a temporal alignment between the number of jobs that require skills and the number of times that Reddit users have talked about those skills in the past.

Next, we are interested in exploring the frequency of different ISCO skills on Reddit and job postings at shift time $t + 5$. To be able to visually compare the highly posted contents in both data sources, we select the week at which the normalized frequency of Reddit and job postings is close

Table 5. Absolute CCF values of Reddit and job postings in different shift times.

Shift time (weeks)																			
19	18	17	16	15	14	13	12	11	10	9	8	7	6	**5**	4	3	2	1	0
0.09	0.11	0.00	0.09	0.08	0.04	0.04	0.22	0.00	0.08	0.00	0.12	0.17	0.01	**0.24**	0.19	0.17	0.09	0.13	0.30

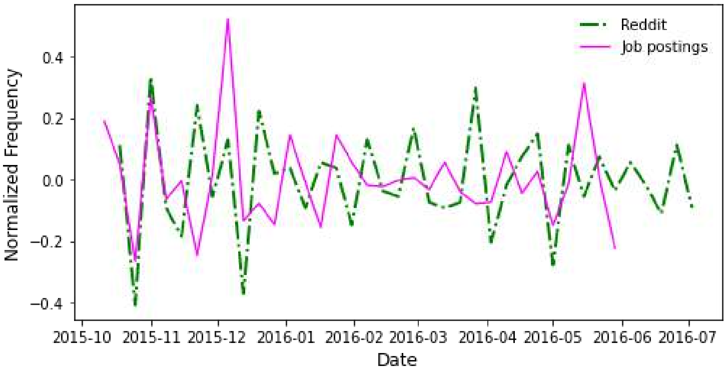
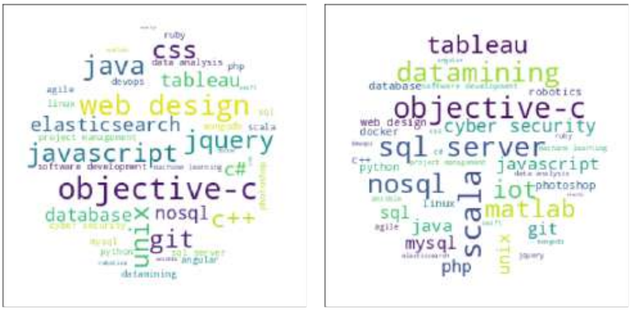


Fig. 6. Weekly temporal frequency of Reddit and job postings at time $t + 5$ with CCF value of 0.24.



(a) Job postings

(b) Reddit

Fig. 7. Word clouds of the most frequently ISCO skills posted in job postings and Reddit from Nov 3, 2015 to Nov 11, 2015 at shift time $t + 5$.

to each other. Therefore, based on Figure 6, we selected the data posted between Nov 3, 2015, and Nov 11, 2015, for this experiment. Figure 7 shows the most frequently posted skills on Reddit and in job postings. As illustrated in Figure 7, in both data sources, skills such as Objective-C, Javascript, Tableau, Java, NoSql, and Git were highly mentioned both in recruitment and social data.

Our results indicate that by using social content, it would be possible to recognize future skill trends that may not be immediately recognizable otherwise. This finding is aligned with prior work [72, 73, 88] that showed how online communities are reflective of the newest skill sets and knowledge that are required in those jobs that are rapidly evolving.

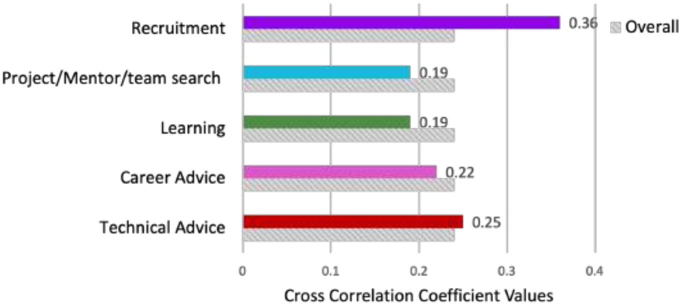


Fig. 8. Comparison between the CCF of topic-related time series and all Reddit content time series.

4.2 RQ2. The Utility of Social Content Topic Types for Understanding Future In-demand Skills

The second research question is concerned with understanding whether different topics discussed on Reddit sub-communities have differing degrees of association with future in-demand skills. To this end, and as outlined in Section 3.3, we identified five topics using a Grounded theory approach and then developed effective deep learning classifiers on expert-annotated data to classify the Reddit posts into these topics. Based on the topics, our objective is to discover whether the frequency of the skills mentioned within a particular topic on Reddit is a better indicator for future in-demand skills in online job postings compared to when all social contents is used without considering topics. To this end, we employ the topic-based time series that were described in Section 3.3.2.

Figure 8 illustrates the correlation observed between each topic-based time series and the online job posting time series. The figure also provides a contrast with the correlation that is observed based on the analysis in RQ1. As shown in this Figure, different topics correlate with the job postings in time $t + \tau$ differently. For example, contents that are related to the *recruitment* topic exhibit a correlation of 0.36 at shift time 4. In other words, by only using Reddit posts that are related to the *recruitment* topic, one can better understand future in-demand skills with an increase of 44% over when the topic distinction is not considered. In addition to the *recruitment* topic, Figure 8 shows that the *technical advice* topic can improve the correlation by 4.4% and yield a better CCF value compared to the case when we use all the posts regardless of their topics. However, we also observe that unlike *recruitment* and *technical advice*, the content in other topics is not correlated with future in-demand skills as effectively as when topics are not considered.

Second finding: the utility of IT-related social content in understanding the future of in-demand skills is highly dependent on the topic they cover. We found that posts discussing *recruitment* and *technical topics* are the best indicators of future in-demand skills.

As shown in Figure 3, we note that while only one percent of the social content in our dataset is related to the *recruitment* topic, our results show that this one percent of the data by itself can show a notably higher correlation with future in-demand job skills. In contrast, while the majority of our Reddit content, i.e., 83% of the content, is related to the *technical advice* topic, we observe a lower degree of correlation when compared to the *recruitment* topic. Our findings also show that by only considering the posts that are related to *career advice*, help on learning technical skills or finding a project/team, we cannot observe a correlation as strong as that observed in the other two topics and also when topics were not considered.

Third finding: More data does not necessarily lead to a more meaningful relationship between social content and future in-demand skills. In other words, the temporal characteristics of specific social content types, even with low volumes, are strong indicators for future in-demand skills.

To further analyze the notable increase in correlation within the *recruitment* topic, we show the characteristics of the data associated with this topic. When reviewing the notes from the Grounded theory process related to the *recruitment* topic, we note that this topic consists of posts related to a variety of purposes, such as formal job advertisements that are shared by hiring managers, as well as individuals looking to hire a mentor, tutor, or contributor to partner up with. Many of the posts belong to hiring managers who rely on online communities to understand what skills they should look for in applicants, how to interview them, and more importantly, where to look for qualified candidates. The fact that both individual employees as well as hiring managers rely on crowd sourced content from Reddit to decide about job descriptions and recruiting strategies is an indication of why future in-demand skills correlate highly with content on this topic. We will further elaborate on this issue in Section 5.

4.3 RQ3. The Utility of User Engagement Indicators for Understanding Future In-demand Skills

In the past two research questions (RQs 1 and 2), we explored the relationship between the frequency of Reddit content about in-demand skills and the future demand for those skills in online job postings. In RQ3, our main objective is to study if and to what extent content with certain user engagement characteristics is more appropriate for determining future in-demand skills. To this end, we employ the time series associated with high and low engaging content, described in Section 3.4, in order to measure their association with future in-demand job skills. It is important to note that when we divide our social data into high and low-engaging content, the frequency of skills mentioned in each category becomes different compared to when we consider all content in a topic. Therefore, the time series developed for RQ3 will not be identical to the other time series developed for the previous RQs. Thus, the shift time at which the maximum correlation between job postings and low/high engaging content happens may be different from previous instances. For example, when we consider all content related to career advice, the maximum correlation of 0.22 occurs at shift time $t + 5$. However, low-engaging posts related to the career advice topic exhibit a correlation of 0.28 at shift time $t + 8$ while high-engaging posts have a maximum correlation of 0.39 at shift time $t + 5$. We compare the obtained CCF in each engagement category with the results obtained from RQs 1 and 2. We report these results in Table 6. Table 6 shows that in the majority of our content types, only considering high-engaging posts when building the social content time series can lead to notable improvement of the CCF values. As one example, when considering all social content regardless of its type, high-engaging posts increase the CCF value by 79.1%. Therefore, we conclude our fourth finding as follows:

Fourth finding: consideration of high engaging posts in online social content, i.e., those that have received the highest number of upvotes, is a strong indicator for future in-demand skills.

To further validate this finding, we conduct a robustness check by modifying our low and high-engagement definitions. To do so, instead of dividing the social content into 3 quantiles, we divide it into 5 quantiles. Therefore, the high and low engaging contents are the top and bottom 20% of content sorted based on the number of upvotes they have received. Finally, we follow the steps in Section 3.4 to construct new time series and calculate the CCF between the social content and the

Table 6. Comparison of the CCF between low and high-engaging posts. In this table, the Overall row refers to the case when we consider all the social content regardless of its topic. "CCF diff" refers to the percentage of the difference between the CCF of all the posts in the time series and the different engagement levels. The upward and downward arrows identify improvement and deterioration of CCF in each row, while the grey rectangle indicates no significant change.

Low User Engagement				**High User Engagement**			
Time Series	**#Posts**	**CCF**	**CCF diff**	**Time Series**	**#Posts**	**CCF**	**CCF diff**
Overall	28,806	0.30	↑ 25%	Overall	25,056	0.43	↑ 79.1%
Technical Advice	24,861	0.30	↑ 20%	Technical Advice	20,308	0.43	↑ 72%
Career Advice	3,457	0.28	↑ 27.27%	Career Advice	2,064	0.39	↑ 77.27%
Recruitment	129	0.39	↑ 8.33%	Recruitment	100	0.26	↓ 27.78%
Learning	836	0.19	▬ 4.12%	Learning	728	0.34	↑ 78.95%
Project/Mentor/team Search	489	0.33	↑ 73.68%	Project/Mentor/team Search	198	0.31	↑ 63.16%

Table 7. Comparison of the CCF between low and high engaging posts based on 5 quantiles. In this table, the Overall row refers to the case when we consider all the social content regardless of its topic. CCF Diff refers to the percentage difference between the CCF of all the posts in the time series and the different engagement levels. The upward and downward arrows identify improvement and deterioration of CCF in each row, while the gray rectangle indicates no significant change.

Low User Engagement				**High User Engagement**			
Time Series	**#Posts**	**CCF**	**CCF diff**	**Time Series**	**#Posts**	**CCF**	**CCF diff**
Overall	28,806	0.30	↑ 25%	Overall	15,087	0.42	↑ 75%
Technical Advice	24,861	0.30	↑ 20%	Technical Advice	11,948	0.33	↑ 32%
Career Advice	2,478	0.26	↑ 18%	Career Advice	1,440	0.38	↑ 73%
Recruitment	123	0.39	↑ 8%	Recruitment	65	0.42	↑ 17%
Learning	792	0.19	▬ 0%	Learning	453	0.20	↑ 5%
Project/Mentor/team Search	327	0.32	↑ 68%	Project/Mentor/team Search	188	0.31	↑ 63%

job postings. Results from this experiment are reported in Table 7. As shown in Table 7, the obtained results based on 5 quantiles are consistent with those reported based on 3 quantiles reported in Table 6. Therefore, our earlier findings that high-engaging posts increase our understanding of the future in-demand skills are robust from the point of view of the definition of low and high engagement content.

To understand the obtained results, we analyzed our data in further detail. By considering the data in each topic, we find that the majority of the posts that are part of the high-engaging content are mainly neutral posts about sharing news, experience, or opinions and are not looking for help or any particular advice from the community. Additionally, the purpose of some of these posts is to start a conversation between knowledgeable experts and to require broader contributions from the user base.

In addition to our finding that the consideration of high-engaging posts can be a strong indicator for future in-demand skills, we make further observations based on content in the *recruitment* topic. We found that *recruitment* content correlates with future skills with a CCF value of 0.36, while high-engaging posts in this topic reduce the obtained CCF by 0.1. This finding motivated us to further explore the data in the *recruitment* topic to identify the types of posts that are classified in this group. We realized that the *recruitment* topic consists of four main classes of content, each with its own unique characteristics. These classes and their definitions, along with three examples for each class, are reported in Table 8. As shown in the table, the first class of content represents job postings posted by hiring managers and how social media assists recruiters with talent acquisition and hiring [67, 71]. The second class represents hiring managers looking for advice from the

Table 8. Different types of classes in the recruitment content type.

Class	Type Definition	Post Title
1	international job posting advertisements posted by companies hiring managers or recruiters	Job opening in Seattle for senior level Devs
		[PAID] Head of Analytics at Bossa Studios, London, UK
		[Job Ad] Haskell Developer in Ankara, Turkey
2	content posted by hiring managers or recruiters to obtain advice on where, what, and how to look for talents	Looking for part time help. Preferably Intern. What should I be looking for?
		Where to hire django programmers/installers
		What should I ask to hire a good python developer?
3	Labour demands by individuals for a variety of purposes	Looking to hire a coder to help with JAVA 1 homework
		[MENTOR] Software Developer Taking On Newbies!
		Looking to hire iOS tutor to help me implement my startup idea
4	content posted by job seekers who are looking for the paid and unpaid positions	[Unpaid] [Programmer] Student programmer looking for experience
		[Paid] Programmer looking for short term Job
		[OFFER] Building Portfolio - I'll design anything

Table 9. Number of posts belonging to each type in high and low engaging categories.

Engagement Level	Class 1	Class 2	Class 3	Class 4	Class - other
High	61%	6%	9%	14%	9%
Low	34%	5%	28%	25%	7%

community in the recruiting process, e.g., where to hire talent or what to consider when hiring for a position. The second class demonstrates that not only job seekers and lifelong learners, but also hiring managers need the help of online communities to make hiring decisions. The third and fourth classes consist of content posted by individuals who are looking for or providing labor, respectively.

As demonstrated in Table 9, we find that among the four mentioned classes of content, 61% of the high-engaging posts are in the first class and are mainly paid job advertisements that are more likely posted by companies hiring managers or recruiters. However, in the low-engaging group, we find that in addition to the first class, the third and fourth classes, i.e., those content posted by individuals who look for or provide labor, are also prevalent. In other words, most of the posts that did not receive many upvotes are those looking for hiring tips or are posted by job hunters, or users who are looking to hire tutors or people who can help them with their projects or homework. Therefore, we can say that the characteristics of the high and low-engaging posts are different in a way that the high-engaging group mainly represents *labor demand* while the low-engaging group constitutes *labor supplies*. By showing different types of posts in low and high-engaging categories and the CCF values reported in Table 6, we can draw the conclusion that the behavior of the time series based on labor supply correlates better with the future of in-demand skills than just considering the content from labor demand. Thus, our observations indicate that only using historical job postings without considering the social media content limits our obtained insight to understanding shifts in future in-demand skills. Thus, our fifth finding is as follows:

Fifth finding: social content by job applicants who are looking for opportunities or recruiters who are looking for advice from the community correlates quite well with future in-demand skills. This correlation is much more notable compared to official job advertisements posted on Reddit.

5 DISCUSSIONS

In this paper, we have examined whether user-generated social content from Reddit can provide insight into the future skill demand in the job market. Summarily, we have found that:

- Social content posted on community-based platforms, such as Reddit, can serve as strong indicators of the future in-demand skills represented in job postings;

- Relevant social content have differing degrees of utility depending on the type of the content for the sake of understanding the future of in-demand skills. We have found that social content discussing recruitment and technical topics have the most utility for determining future in-demand job skills;
- High engaging social content are better indicators of the future in-demand skills compared to low engaging posts;
- Social content originating from job seekers and learners are better indicators of future in-demand skills compared to those posted by hiring managers and recruiters.

We note that the findings in this paper corroborate observations made in human capital research, which discuss the importance of human capital development in the evolving knowledge economy [65]. These findings may enable academic institutions and policymakers to empirically identify in-demand skills with access to inexpensive, large-scale, and naturalist social data. Thus, academic institutions can create curricula according to the future in-demand skills and industries can execute talent recruitment and job guidance more effectively. Additionally, this work incorporates a valuable line of research for future scholars that corresponds to the literature's increasing focus on building tools that assist industries in gaining insights into the recent developments in the job market. In the remainder of this section, we discuss the implications of the proposed methodology for the general public and governments.

5.1 Implications for the Public

In the knowledge economy, where on-site training has become outdated, employers expect the workforce to be independent self-learners with the ability to learn on the job [100]. Therefore, individuals are dependent on their own social circles to acquire new skills and gain input into the qualities that can help them do their job more successfully and remain in demand [88]. Building on the existing literature that highlights the importance of online communities in professional skill development [73], we shed light on the significance of online communities as a possible source of information for tracking talents that may become highly in-demand and required prerequisites in future occupations. Beyond the importance of personal networks, our findings indicate that online communities discussing highly engaging technical and *career advice* content and posts representing labor supply have the most impact on how job specifications (skills) are shaped.

Currently, many of the existing skill recommendation tools are developed by detecting trendy skills appearing in historical recruitment data, e.g. job postings, and overlooking the significance of other job market identifiers, such as social data [30, 113, 114, 124]. Similarly, it is common practice among job seekers and new graduates to regularly explore job postings to identify the skills that they require to enter the job market. One of the main shortfalls of this approach is looking at static data and missing the opportunity to engage with the professional community to find direction and get answers to recruitment and development inquiries. This is while prior work shows how professional workers can take advantage of the social interactions and self-disclosure that occur in online communities. [72]. Similarly, we observe in our data that, according to Figure 3, although the majority of the individuals use Reddit as a source of skill development, with 83% of the content dedicated to technical inquiries, there are already some learners, job seekers, and recruiters who reach out to the community experts to get career advice or get insights on the "what to learn next" question, which can be frustrating to find an answer to. Here are two examples that illustrate how users take advantage of their interactions with the online community:

Example 1:

- Number of comments: 19
- Title: Depressed, intimidated, and feeling lost. Help?

- Body: I'm having a hard time career-wise and thought it might be worth reaching out here. I graduated a design program and have been working since 2014, but I feel I have had little growth since then due to having trouble finding work I care about. I had a great job coming out for school that I loved, but was laid off in a restructure and struggled through mediocre bill-paying freelance for a couple years before getting my current role. In that time my sensibilities/skills kind of dulled and my enthusiasm for the work dwindled sharply . . . If anyone has gone through anything remotely similar I would appreciate hearing how you navigated it all. Thank you for reading.
  TLDR: aimless career path poisoned by self doubt and severe depression, struggling with what to do with myself and reconnecting with my interest in design.
  EDIT: Thank you for the really thoughtful replies, I was feeling super vulnerable putting this out there but it's been really encouraging to not only hear people going through similar experiences but sharing their journeys through it. I hope I can find the drive to be as strong as you guys.

Example 2:

- Number of comments:46
- Title: overwhelmed
- Body: I'm looking to land an entry-level web developer position with the skills of a full-stack web developer. ... Question: The methods of some data structures alone seem incredibly difficult to memorize without peeking into one's notes. Maybe I feel this way because I've not finish my data structures and algorithms book yet to begin testing myself on the topic. During the interview process, what are developers expected to memorize?
  Edit: This kind of blew up. I expected like five replies. Thanks for the super valuable and informative information you guys and gals shared with me. It was very eye opening.

These examples indicate that although learners, employers, and job designers can passively consume the information shared on Reddit technical communities to detect in-demand skills, more benefits can be gained if they actively engage with online experts. Additionally, these examples show that although some users reach out to the community to get advice on available job posts, e.g., Example 2, many others rely on social interactions to seek mental support, get career advice, or find learning materials. Therefore, we show that social data is as effective as the baseline methods that only rely on job postings and is also better in terms of providing a social engagement advantage. Thus, we hope that these results motivate the public to pay attention to social content beyond skill development purposes and to direct more individuals to uncover the changes in the job market and ease their lifelong learning journey by taking advantage of online social engagement. Moreover, our results may encourage the workforce, job applicants, and educators to pay close attention to specific types of content shared in online communities in order to gain insight into essential skill trends within the labor market in order to stay in demand.

5.2 Implications for Governments

Previous research shows clear evidence of a skilled workforce shortage in Science, Technology, Engineering, and Mathematical (STEM) fields across the world, especially in the USA, UK, and Australia [45]. A skill shortage happens when there is a mismatch between labor skills and vacant jobs in high-tech industries [11]. The skill shortage refers to both the quality (the relevance of the skills to the job requirements) and the quantity (the number of people who can fill the available jobs) of the workers. However, Deming and Noray [34] show that in STEM fields, shortages mostly involve a lack of relevant job skills and not the worker's quantity. Skill mismatch is one of the widely-acknowledged reasons that can limit economic growth [22]. As a result, many national governments have made measuring the skill gap and closing it a high-priority policy [25, 28, 121].

There are several strategies to address the talent shortage, such as internal skill development and international talent migration, to name a few [45]. Skill development is known as a significant component of the United Nations Sustainable Development 2030 Agenda [28]. This Agenda consists of 17 Sustainable Development Goals (SDGs) with the purpose of ending poverty, tackling inequalities, promoting decent work, and encouraging lifelong learning opportunities for everyone [35]. Transforming the workforce into lifelong learners and upskilling/reskilling them will provide the expertise required in the digital and postindustrial economies, which can lead to economic growth [11, 122].

Despite the importance of continuous skill development and its impact on living standards, the training content and the vocational curricula are often outdated, and they fail to reflect the skills required in the fast-evolving economy. As shown in an MGI report, in 9 countries, 60 percent of employers indicated that new graduates do not have the adequate technical and soft skills that they are looking for [86]. To close this gap, the CSCW community has already proposed a variety of approaches by which skill development can happen outside the educational context [41, 73, 88]. Another direction to address this issue is to form social media teams in the government and educational institutes with the aim of regularly tracking the career and learning-related content that appears on social media. Currently, the primary goal of social media engagement teams, especially in governments, is to track public political opinions and engage communities [42]. In educational institutions, online social networks are mainly utilized for marketing purposes [62]. However, as shown in this study, engaging with online communities will help policy-makers design new curricula that reflect the future requirements of the job market, which is crucial for training the next generation of lifelong learners.

Another solution to address the skill shortage is talent migration, temporary or permanent. To attract foreign talent, governments have altered their immigration policies and laws to approach human capital and focus on the labor market needs [83]. Increasingly, governments introduce new employer-driven immigration programs that help recruit the international skilled workers faster than other immigration programs [76]. Therefore, the importance of defining skills and identifying the changes in the labor market for designing the skill migration programs, recruiting talent workers, and building the required capabilities becomes obvious [18]. Yet, there are few practical techniques that provide insight into the job market transformations and identify the in-demand skills in high-tech industries that are crucial to being reflected in skill immigration policies and skill development curricula. This paper introduces a new avenue that may assist governments and policymakers in designing immigration and reskilling programs by recognizing the skills that need to be prioritized in recruiting and learning activities.

6 LIMITATIONS AND FUTURE DIRECTIONS

This study demonstrates the efficacy of social content in understanding the future in-demand skills of the job market specifically for the Information Technology (IT) sector and based solely on Reddit content. Therefore, we cannot make broad claims about the effectiveness of the proposed methodology for other professions beyond the IT sector and other social platforms other than Reddit. Therefore, our findings are limited to the scope of our datasets. Additionally, we acknowledge certain shortcomings in this analysis that we intend to overcome in future work. The first limitation refers to our feature selection approach, which overlooked the lexico-syntactic features of the Reddit content. As such, our work does not model the semantic relationship between the posts and their impact on the representation of job postings. A possible solution to this limitation is to utilize representation learning approaches [52, 104]. Combining lexico-syntactic features with the word counts of skills to capture future patterns of in-demand skills remains an open avenue for the future. Another limitation of this work is in regards to the size of the collected dataset. In our future work, we intend to re-train our machine learning models on larger datasets in order to ensure higher degrees of robustness and generalizability.

Another interesting future direction is to investigate the role of content longevity in determining future in-demand skills. According to prior studies, one of the main factors that can affect the longevity of social content is repeated content sharing over extended periods of time [69]. Therefore, in the context of our work, in addition to only considering the number of upvotes to detect social engagement, one can calculate the time period over which a post containing a skill remains popular within the community and attracts attention.

REFERENCES

[1] Opeyemi Aborisade and Mohd Anwar. 2018. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 269–276.

[2] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398* (2019).

[3] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. 2019. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 47–57.

[4] 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.

[5] Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. 2019. New data sources in social science research: things to know before working with reddit data. *Social science computer review* (2019), 0894439319893305.

[6] 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.

[7] Lorenzo Barberis Canonico, Nathan J McNeese, and Chris Duncan. 2018. Machine learning as grounded theory: Human-centered interfaces for social network research through Artificial Intelligence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 1252–1256.

[8] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[9] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[10] Eduard Baumohl and Stefan Lyocsa. 2009. Stationarity of time series and the problem of spurious regression. *Available at SSRN 1480682* (2009).

[11] World Economic Forum Boston Consulting Group (BCG). 2018. Towards a reskilling revolution: A future of jobs for all. World Economic Forum, Geneva, Switzerland.

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

[13] Mark Belford and Derek Greene. 2020. Ensemble topic modeling using weighted term co-associations. *Expert Systems with Applications* 161 (2020), 113709.

[14] 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).

[15] Allie Blaising, Yasmine Kotturi, Chinmay Kulkarni, and Laura Dabbish. 2021. Making it Work, or Not: A Longitudinal Study of Career Trajectories Among Online Freelancers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–29.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[17] Toine Bogers and Rasmus Nordenhoff Wernersen. 2014. How'Social'are Social News Sites? Exploring the Motivations for Using Reddit. com. In *iConference 2014: Breaking Down Walls: Culture-Context-Computing*. iSchools, 329–344.

[18] Anna Katherine Boucher. 2020. How 'skill'definition affects the diversity of skilled immigration policies. *Journal of Ethnic and Migration Studies* 46, 12 (2020), 2533–2550.

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

[20] 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.

[21] Tom Broos, Maarten Pinxten, Margaux Delporte, Katrien Verbert, and Tinne De Laet. 2020. Learning dashboards at scale: early warning and overall first year experience. *Assessment & Evaluation in Higher Education* 45, 6 (2020), 855–874.

[22] Belinda Brucker Juricic, Mario Galic, and Sasa Marenjak. 2021. Review of the Construction Labour Demand and Shortages in the EU. *Buildings* 11, 1 (2021), 17.

[23] 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.

[24] Kathy Charmaz and Linda Liska Belgrave. 2007. Grounded theory. *The Blackwell encyclopedia of sociology* (2007).

[25] David Chinn, Solveigh Hieronimus, J Kircherr, and Julia Klier. 2020. The future is now: closing the skills gap in Europe's public sector. *McKinsey & Co. https://www. mckinsey. com/industries/public-and-social-sector/our-insights/the-future-is-now-closing-the-skills-gap-in-europes-public-sector* (2020).

[26] Ruth Cobos and Juan Carlos Ruiz-Garcia. 2020. Improving learner engagement in MOOCs using a learning intervention system: A research study in engineering education. *Computer Applications in Engineering Education* (2020).

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

[28] Paul John Comyn. 2018. Skills, employability and lifelong learning in the Sustainable Development Goals and the 2030 labour market. *International Journal of Training Research* 16, 3 (2018), 200–217.

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] David J Deming and Kadeem L Noray. 2018. *STEM careers and the changing skill requirements of work*. Technical Report. National Bureau of Economic Research.

[35] UN Desa et al. 2016. Transforming our world: The 2030 agenda for sustainable development. (2016).

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[37] Paul M Di Gangi and Molly M Wasko. 2016. Social media engagement theory: Exploring the influence of user engagement on social media usage. *Journal of Organizational and End User Computing (JOEUC)* 28, 2 (2016), 53–73.

[38] 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.

[39] Tawanna R Dillahunt, Matthew Garvin, Marcy Held, and Julie Hui. 2021. Implications for Supporting Marginalized Job Seekers: Lessons from Employment Centers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[40] Tawanna R Dillahunt and Joey Chiao-Yin Hsiao. 2020. Positive Feedback and Self-Reflection: Features to Support Self-Efficacy among Underrepresented Job Seekers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[41] 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.

[42] Elizabeth Dubois, Anatoliy Gruzd, Philip Mai, and Jenna Jacobson. 2018. Social Media and Political Engagement in Canada. (2018).

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

[44] 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.

[45] Elaine Farndale, Mohan Thite, Pawan Budhwar, and Bora Kwon. 2021. Deglobalization and talent sourcing: Cross-national evidence from high-tech firms. *Human Resource Management* 60, 2 (2021), 259–272.

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

[47] 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.

[48] Kausalya Ganesh and Amanda Lazar. 2021. The Work of Workplace Disclosure: Invisible Chronic Conditions and Opportunities for Design. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[49] Radhika Garg, Yash Kapadia, and Subhasree Sengupta. 2021. Using the Lenses of Emotion and Support to Understand Unemployment Discourse on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24.

[50] 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.

[51] 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.

[52] Kristina Gligorić, Ashton Anderson, and Robert West. 2019. Causal effects of brevity on style and success in social media. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[53] Mareike Glöss, Moira McGregor, and Barry Brown. 2016. Designing for labour: uber and the on-demand mobile workforce. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1632–1643.

[54] Steven L Gortmaker. 1994. Theory and methods–Applied Logistic Regression by David W. Hosmer Jr and Stanley Lemeshow. *Contemporary sociology* 23, 1 (1994), 159.

[55] Clive WJ Granger and Paul Newbold. 1974. Spurious regressions in econometrics. *Journal of econometrics* 2, 2 (1974), 111–120.

[56] Alan Greenspan. 2000. The Evolving Demand for Skills. (2000).

[57] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. *arXiv preprint arXiv:1908.05620* (2019).

[58] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine learning* 95, 3 (2014), 423–469.

[59] Rob J Hyndman and George Athanasopoulos. 2018. *Forecasting: principles and practice*. OTexts.

[60] 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.

[61] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. Support vector machines. In *An Introduction to Statistical Learning*. Springer, 337–372.

[62] Waldemar Jędrzejczyk. 2021. Barriers in the Use of Social Media in Managing the Image of Educational Institutions. *Procedia Computer Science* 192 (2021), 1904–1913.

[63] 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.

[64] 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.

[65] Carol Kasworm. 2011. The influence of the knowledge society: Trends in adult higher education. *The Journal of Continuing Higher Education* 59, 2 (2011), 104–107.

[66] 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.

[67] 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.

[68] Zachary Kilhoffer. 2020. Report on How to Identify and Compare Newly Emerging Occupations and Their Skill Requirements. (2020).

[69] Hyun Suk Kim. 2020. How Message Features and Social Endorsements Affect the Longevity of News Sharing. *Digital Journalism* (2020), 1–22.

[70] Johanna Kluge, Julian Hildebrandt, and Martina Ziefle. 2019. The Golden Age of Silver Workers?. In *International Conference on Human-Computer Interaction*. Springer, 520–532.

[71] Tanja Koch, Charlene Gerber, and Jeremias J De Klerk. 2018. The impact of social media on recruitment: Are you LinkedIn? (2018).

[72] 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.

[73] 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.

[74] Yubo Kou, Colin M Gray, Austin L Toombs, and Robin S Adams. 2018. Understanding social roles in an online community of volatile practice: A study of user experience practitioners on reddit. *ACM Transactions on Social Computing* 1, 4 (2018), 1–22.

[75] 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.

[76] Shauna Labman and Sarah Zell. 2021. The shift towards increased citizen-driven migration in Canada. In *Research Handbook on the Law and Politics of Migration*. Edward Elgar Publishing.

[77] Airi Lampinen, Victoria Bellotti, Coye Cheshire, and Mary Gray. 2016. CSCW and the Sharing 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.

[78] 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.

[79] 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.

[80] Linfeng Li, Tawanna R Dillahunt, and Tanya Rosenblat. 2019. Does Driving as a Form of. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–16.

[81] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[82] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. *arXiv preprint arXiv:1910.00883* (2019).

[83] Lucia Lo, Wei Li, and Wan Yu. 2019. Highly-skilled Migration from China and India to Canada and the United States. *International migration* 57, 3 (2019), 317–333.

[84] Jalehsadat Mahdavimoghaddam, Niranjan Krishnaswamy, and Ebrahim Bagheri. 2021. On the Congruence Between Online Social Content and Future IT Skill Demand. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[85] Feng Mai, Zihan Chen, and Aron Lindberg. 2019. Does Sleep Deprivation Cause Online Incivility? Evidence from a Natural Experiment. (2019).

[86] James Manyika. 2017. Technology, jobs and the future of work. (2017).

[87] 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).

[88] 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.

[89] Judy McGregor, David Tweed, and Richard Pech. 2004. Human capital in the new economy: devil's bargain? *Journal of Intellectual Capital* (2004).

[90] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.

[91] Rizwan Mushtaq. 2011. Augmented dickey fuller test. (2011).

[92] 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.

[93] James Ness et al. 2020. Technology in education. In *International Conference on Human-Computer Interaction*. Springer, 574–585.

[94] 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.

[95] 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.

[96] 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.

[97] Peter CB Phillips and Pierre Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75, 2 (1988), 335–346.

[98] 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.

[99] Thomas Puschmann and Rainer Alt. 2016. Sharing economy. *Business & Information Systems Engineering* 58, 1 (2016), 93–99.

[100] Sungsup Ra, Unika Shrestha, Sameer Khatiwada, Seung Won Yoon, and Kibum Kwon. 2019. The rise of technology and impact on skills. *International Journal of Training Research* 17, sup1 (2019), 26–40.

[101] Julian Risch and Ralf Krestel. 2020. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 579–589.

[102] Veronica A Rivera and David T Lee. 2021. I Want to, but First I Need to: Understanding Crowdworkers' Career Goals, Challenges, and Tensions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–22.

[103] 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.

[104] Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27.

[105] Koustuv Saha, Sang Chan Kim, Manikanta D Reddy, Albert J Carter, Eva Sharma, Oliver L Haimson, and Munmun De Choudhury. 2019. The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–22.

[106] 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.

[107] 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 Web and Social Media*, Vol. 12.

[108] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[109] Bhavani Seetharaman, Joyojeet Pal, and Julie Hui. 2021. Delivery Work and the Experience of Social Isolation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–17.

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

[111] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*. 58–65.

[112] 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.

[113] 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.

[114] 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.

[115] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of reddit: from the front page of the internet to a self-referential community?. In *Proceedings of the 23rd international conference on world wide web*. 517–522.

[116] Susan Leigh Star. 1998. Grounded classification: Grounded theory and faceted classification. (1998).

[117] 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.

[118] 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.

[119] Maria Tomprou, Laura Dabbish, Robert E Kraut, and Fannie Liu. 2019. Career mentoring in online communities: Seeking and receiving advice from an online community. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[120] Pucktada Treeratpituk, Pradeep Teregowda, Jian Huang, and C Lee Giles. 2010. SEERLAB: A system for extracting keyphrases from scholarly documents. In *Proceedings of the 5th international workshop on semantic evaluation*. 182–185.

[121] Anneleen Vandeplas, Anna Thum-Thysen, et al. 2019. *Skills Mismatch & Productivity in the EU*. Publications Office of the European Union.

[122] J Woetzel, J Seong, N Leung, J Ngai, LK Chen, V Tang, S Agarwal, and W Bo. 2021. Reskilling China: Transforming the world's largest workforce into lifelong learners. *McKinsey Global Institute (MGI)* (2021).

[123] 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.

[124] 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.

[125] 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).

[126] Kiyoung Yang and Cyrus Shahabi. 2005. On the stationarity of multivariate time series for correlation-based data analysis. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 4–pp.

[127] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).

[128] Zheng Yao, Silas Weden, Lea Emerlyn, Haiyi Zhu, and Robert E Kraut. 2021. Together But Alone: Atomization and Peer Support among Gig Workers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[129] 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 ISCO SKILLS

Here, we present a list of all the skills that were used in this study in Reddit and job posting datasets.

Table 10. List of ISCO IT skills.

Skill			
Cloud technologies	Neural networks	haskell	chatbots
xquery	opencv	graphql	cobol
Apache kafka	Ios programming	tensorflow	vbscript
software testing	coffeescript	apache spark	data engineering
business intelligence	cyber security	cloud computing	3d modeling
adobe illustrator	agile	angular	ansible
artificial intelligence	c#	c++	computer vision
css	data analysis	database	datamining
deep learning	devops	django	docker
elasticsearch	iot	joomla	machine learning
git	java	jquery	matlab
googlecloud	javascript	linux	mongodb
mysql	objective-c	php	project management
nosql	photoshop	postgresql	python
reactjs	ruby	scala	software development
robotics	sass	scrum	sql
sql server	swift	tableau	typescript
unix	vagrant	web design	web programming
xcode			

B LIST OF MEMOS

In the initial iteration of the Grounded theory process, the researchers came up with 15 categories that represent the majority of Reddit posts. Name of these topics and detailed information on each topic is provided in this section.

Table 11. Sample of the initial memo for 15 categories on 8.5% of the data.

Initial topic names	Categories definition	Number of posts
Challenge	Posts relating to technical tests/challenges/brain-teasers. E.g. post asking people to decode a cipher the user has created.	1
General advice	Posts regarding general advice. E.g. post seeking advice on how to fix/manage a broken/problematic team.	1
Certification	Post regarding certifications such as AWS, GCP, etc and people giving their feedback or asking for which one to go for.	2
Seeking friends/teams	posts relating to users looking for a partner/team of collaborators to work with. E.g. user wants to establish a team to work on a project, lists skills/requirements needed.	6
Job posts	Recruiters posting the jobs with duties and responsibilities.	18
Recruitment	Posts regarding recruiter asking for advice on hiring.	4
Promotion	self-promoting posts or advertisements/affiliate links/promotional codes. E.g. ``We are the founders of _____. Sign-up for/check out our _____, here's a promo code."	206
Share Knowledge/experience	People sharing their experience related to a task they accomplished successfully. For example, How to set up mongodb on linux.	128
Program/language/tool/tech search	People asking for the programming language that helps them with a particular task or tool.	187
Resume Advice	People posting their resume and asking for advice/feedback.	205
Project Advice	People asking a specific question regarding their project, what technology they should use	224
Seeking Feedback/Code review	posts seeking design/architectural/visual/technical feedback on their code/products/websites/ applications. E.g. posting their code/application/website link and asking for criticism/feedback on it.	240
Advice on learning	People asking for what programming languages they should learn, courses they should take in their university/college, online courses they should opt for.	296
Job search	People searching for a job, post their skills and experience and want someone to hire them based on that.	258
Troubleshoot	People post technical questions such as stack trace from java error, or any other language and ask for help. This class also includes other technical queries outside of computer science field such as plumbing, electrical, etc.	534

C TRAINED TEXT CLASSIFIERS

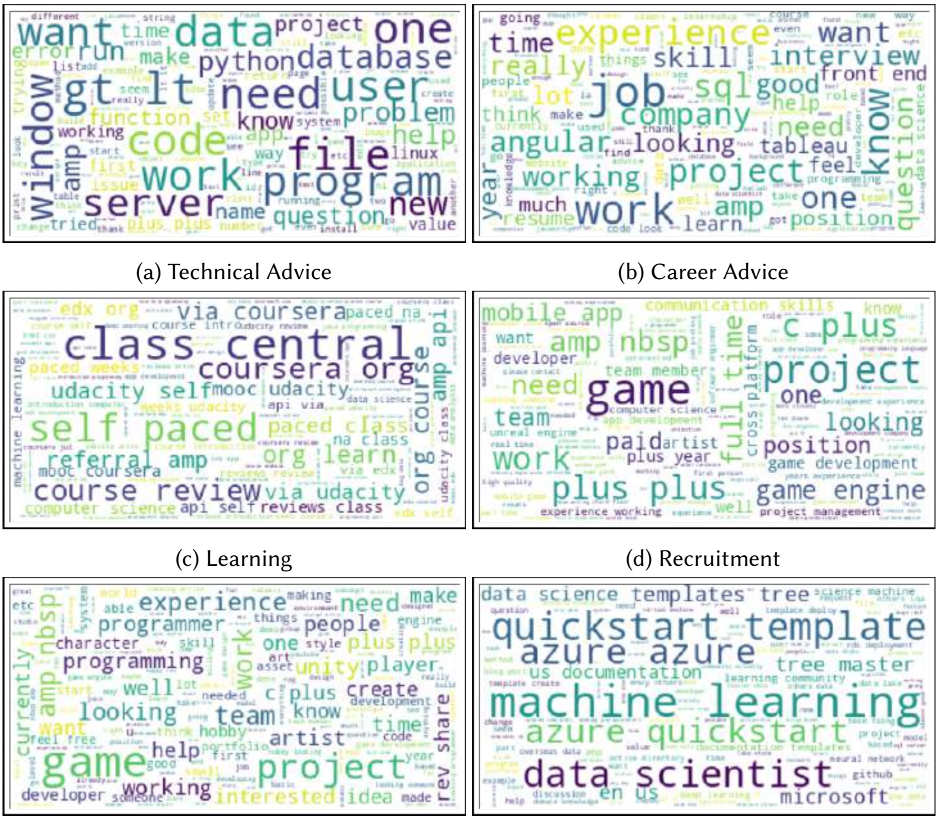
To build our text classifiers, we applied preprocessing steps to clean up our input data. First, we made a contraction dictionary to convert short words, such as “I’ve” and “I’ll”, to their complete format, such as “I have” and “I will”. Then, we removed square brackets, round parentheses, links, and words containing numbers. Additionally, we also performed lowercasing, stop word removal, stemming, and punctuation removal. During the initial phase of training the text classifiers, the models were trained and tested against the Reddit dataset. Hyper-parameter tuning was done using the GridSearchCV from the sklearn library.

To implement our deep neural networks, we used pre-trained bidirectional transformers, namely BERTbase (12-layer, 768-hidden, 12-heads), DistillBERTbase (6-layer, 768-hidden, 12-heads), and XLNetbase (12-layer, 768-hidden, 12-heads). Existing literature shows that fine-tuning pre-trained bidirectional transformers, even with small labeled training examples, increases the performance of the models and helps them generalize better [36]. Additionally, case studies show that fine-tuning pre-trained models, such as BERT, DistillBERT, and XLNET, is robust to overfitting [57, 82].

To fine-tune BERTbase, DistillBERTbase, and XLNetbase, based on our experiments, we set the batch size to 16, the learning rate to $2E - 5$, and use the ADAM optimizer. Due to the large length of the Reddit posts, we set the maximum sequence length to 512, and any longer text was truncated to the maximum length, while the shorter text was padded by zero. As the input of BERT and DistillBERT classifiers, we tokenized each Reddit post with BERT and DistilBert tokenizer to lowercase the input, perform basic tokenization, remove invalid characters, and split by punctuation. For the XLNet classifier, we used the XLNet tokenizer to lowercase and remove excess spaces before and after the string when tokenizing the input.

To evaluate the performance of our text classifiers, we used a k-fold (k=5) cross-validation technique on our labeled data (6,750 posts) and calculated precision, recall, and F1-score for each class in each fold. We report the mean of the achieved metrics in 5 folds for each text classifier in Table 12. To select the best model, we considered the weighted-average F1-score because, as reported in Table 3, our dataset is imbalanced with various class distributions. Therefore, the weighted average F1-score is the most robust metric that accounts for class imbalance. By looking at the F1 scores reported in Table 12, we see that the performance of all models in the majority of our classes, except for the ‘other’ class, is above 80% on our test set from 5-fold cross-validation. To better understand the reason that all classifiers performed well in most classes, we plotted the top 1,000 words in each class, shown in Figure 9. As illustrated in Figure 9, the most frequent words in each class are discriminatory and representative of their topic. For example, the top words in the “recruitment” class are terms, such as work, position, full time, and communication skills, while the top words in the “learning” class are related to learning platforms and courses. Therefore, classifiers are able to easily distinguish between the classes.

Table 12 indicates that BERT and DistillBERT models are able to effectively classify Reddit posts in all classes. Therefore, we select our final model between BERT and DistillBERT, both with a weighted-average F1 score of 0.97% in all classes. Although both BERT and DistillBERT have equal weighted-average F1 score, we noticed a slightly better performance in “Project/Mentor/team search” and “Recruitment” topics classified by DistillBERT model. Thus, we select DistillBERT as our final model for the classification of unlabelled Reddit posts.



(a) Technical Advice

(b) Career Advice

(c) Learning

(d) Recruitment

(e) Project/Mentor/Team Search

(f) Other

Fig. 9. Distribution of top 1000 words in each class.

Table 12. Comparison between performance of the 6 trained machine learning models for topic classification task. The overall accuracy for all the topics is reported at the end of the table. The outperformed models based on F1-score are BERT and DistillBERT.

Topics	Classifier	Title + Text		
		Precision	Recall	F1-Score
Other	BERT (base uncased)	0.95	0.94	0.95
	DistilBERT (base uncased)	0.92	0.95	0.94
	XLNet	0.75	0.77	0.76
	Logistic Regression	0.85	0.55	0.67
	LinearSVC	0.97	0.52	0.68
	Random Forest	0.98	0.36	0.52
Technical Advice	BERT (base uncased)	0.99	0.98	0.99
	DistilBERT (base uncased)	0.99	0.98	0.99
	XLNet	0.97	0.98	0.98
	Logistic Regression	0.94	0.97	0.95
	LinearSVC	0.92	0.99	0.95
	Random Forest	0.85	0.99	0.92
Career Advice	BERT (base uncased)	0.96	0.96	0.96
	DistilBERT (base uncased)	0.96	0.96	0.96
	XLNet	0.92	0.96	0.94
	Logistic Regression	0.93	0.94	0.93
	LinearSVC	0.93	0.94	0.94
	Random Forest	0.93	0.89	0.91
Recruitment	BERT (base uncased)	0.95	0.94	0.94
	DistilBERT (base uncased)	0.96	0.94	0.95
	XLNet	0.89	0.90	0.90
	Logistic Regression	0.93	0.88	0.90
	LinearSVC	0.96	0.87	0.91
	Random Forest	0.96	0.81	0.88
Learning	BERT (base uncased)	0.98	0.99	0.99
	DistilBERT (base uncased)	0.98	0.99	0.99
	XLNet	0.97	0.98	0.98
	Logistic Regression	0.97	0.98	0.98
	LinearSVC	0.98	0.98	0.98
	Random Forest	0.98	0.98	0.98
Project/Mentor/team search	BERT (base uncased)	0.91	0.94	0.92
	DistilBERT (base uncased)	0.92	0.95	0.94
	XLNet	0.72	0.77	0.76
	Logistic Regression	0.86	0.92	0.89
	LinearSVC	0.87	0.93	0.90
	Random Forest	0.90	0.83	0.86
Average Accuracy	BERT (base uncased)	0.97		
	DistilBERT (base uncased)	**0.97**		
	XLNet	**0.93**		
	Logistic Regression	0.93		
	LinearSVC	0.94		
	Random Forest	0.91		

D CORRELATION BETWEEN REDDIT AND JOB POSTINGS IN DIFFERENT TIME GRANULARITIES

To measure CCF values of Reddit posts and IT skills in job postings in different time granularity, we developed daily, semi-monthly, and monthly time series for Reddit content and job postings. Then, we applied stationarity tests on them, as explained in Section 3.2.1. In daily time series, we noticed that both Reddit and job postings were non-stationary as they both failed the ADF test. After the first-order differencing, Reddit and job postings became stationary and they passed both ADF and KPSS tests. We report the daily relationship between Reddit and job postings in time t up to 60 days after t, $(t + 60)$, in Table 13. As seen in Table 13, although the correlation between Reddit and job postings is the maximum at time $t + 1$, we see meaningful relationships at other times as well.

Similar to the time series created on a daily frequency, Reddit and job postings semi-monthly time series did not pass the ADF test. To make the semi-monthly time series stationary, we applied first-order differencing on both Reddit and job postings time series. However, first-order differencing only made Reddit time series stationary and job postings time series passed the ADF test after third-order differencing. After first-order and third-order differencing on Reddit and job postings, respectively, both time series passed the KPSS test. We report the CCF values of the semi-monthly time series in Table 14. As demonstrated in Table 14, although absolute CCF values of Reddit and job postings are maximized at $t + 2$, the CCF values are also significant at all other times.

Finally, before we measure the correlation between Reddit and job postings with monthly frequency, we applied stationarity tests on them. In the monthly time series, the job postings time series was originally stationary and passed both ADF and KPSS. Unlike job postings, Reddit time series required second-order differencing to pass the ADF test. However, even after the second-order differencing, the Reddit time series did not pass the KPSS test. Due to the limited monthly data (only 10 months) in our dataset, we stopped the stationarity test and concluded our monthly time series are non-stationary and our monthly data is not sufficient for this experiment.

Received July 2021; revised January 2022; revised April 2022; accepted July 2022

Table 13. Absolute CCF values of Reddit and job postings in different daily shift times.

Shift Time	CCF
0	0.402
1	0.628
2	0.043
3	0.284
4	0.297
5	0.154
6	0.119
7	0.379
8	0.463
9	0.062
10	0.35
11	0.108
12	0.149
13	0.054
14	0.263
15	0.359
16	0.052
17	0.247
18	0.177
19	0.063
20	0.027
21	0.28
22	0.226
23	0.38
24	0.258
25	0.102
26	0.075
27	0.01
28	0.239
29	0.27
30	0.057
31	0.251
32	0.124
33	0.066
34	0.015
35	0.254
36	0.278
37	0.09
38	0.255
39	0.13
40	0.04
41	0.067
42	0.17
43	0.221
44	0.047
45	0.239
46	0.129
47	0.066
48	0.1
49	0.208
50	0.154
51	0.069
52	0.241
53	0.078
54	0
55	0.031
56	0.22
57	0.126
58	0.085
59	0.19
60	0.187

Table 14. Absolute CCF values of Reddit and job postings in different semi-monthly shift times.

Shift Time	coefficient
0	0.19
1	0.358
2	0.362
3	0.293
4	0.34
5	0.493
6	0.433
7	0.328