12th International Conference on Computer and Knowledge Engineering (ICCKE 2022), November 17-18, 2022, Ferdowsi University of Mashhad, Iran

IranITJobs2021: a Dataset for Analyzing Iranian Online IT Job Advertisements Collected Using a New Crowdsourcing-based Dataset Gathering Process

Fakhroddin Noorbehbahani, Nikta Akbarpour, Mohammad Reza Saeidi
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
noorbehbahani@eng.ui.ac.ir, nikta.akbarpour@mehr.ui.ac.ir, mo.saeidi@mehr.ui.ac.ir

***Abstract*—Gathering and preparing high-quality data is one of the most significant and expensive steps in data analytics. Crowdsourcing is an efficient way to create datasets for machine learning and data science applications. However, it is vital to apply a proper crowdsourcing process for dataset creation to ensure the quality of the collected data. In this paper, a new process to create high-quality datasets based on crowdsourcing is proposed, including the pre-gathering, gathering, and post-gathering phases. Today employers and job seekers benefit from online job postings and social media sites for recruitment more than ever before. Consequently, a huge volume of job posting data is available that enforces the need for data visualization and data analytics for extracting valuable insights to help better decision making. Although there exist several online job advertisement datasets for analyzing job demand and requirements, there is no such dataset about the IT job market in Iran. In this paper, IranITJobs2021, an online IT job posting dataset, is presented, which is produced using the proposed dataset gathering process. IranITJobs2021 includes job advertisements related to information technology from August 2019 to January 2021. The dataset incorporates 1300 instances and 13 features which is publicly available. IranITJobs2021 could be analyzed to find valuable patterns of job requirements and skills in the field of information technology. Furthermore, the proposed dataset gathering process is applicable to create datasets efficiently.**

Keywords—dataset collection; crowdsourcing; job posting; data analytics

I. Introduction

Today, data is one of the most valuable assets. Data could be converted into useful information and practical knowledge for better decision-making in various fields using data mining methods, statistical analysis, machine learning, and other techniques. The information, patterns, and knowledge gained are applicable to multiple areas, such as business intelligence, to increase profits and reduce business costs. The critical element of data analysis is the quality and size of the dataset. Collecting and controlling the data quality requires considerable time, money, and computation.

As the Internet expands, like-minded individuals in society become more linked, as a result, accomplishing tasks through crowdsourcing has expanded rapidly in recent years [1]. Several studies have attempted to define crowdsourcing, however, by summarizing these definitions, the following definition is provided as a comprehensive definition of crowdsourcing:

"Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or a company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge, and/or experience, always entails mutual benefit" [2].

Crowdsourcing may be utilized to collect data -particularly labeled data- more quickly, cheaply, and efficiently. As a result, this approach to data collecting has received a lot of interest from the academic community [3]. Furthermore, the data quality required to implement data mining and analytics techniques is paramount. The existence of outlier or inaccurate data in the dataset misleads data analytics and interpretations.

Although many researchers have utilized crowdsourcing to conduct their research, few studies propose and apply the well-defined process of crowdsourcing. The lack of a proper crowdsourcing process might lessen the success rate of crowdsourced projects [1]. In this paper, a new crowdsourcing data gathering process will be introduced, which could be utilized to acquire high-quality datasets for data analysis. The suggested process covers all stages of data collection, including the pre-gathering, gathering, and post-gathering phases.

One of the most critical deficiencies of educational systems is the mismatch between the requirements of the job market and the graduates' skills, especially in Iran. This gap is more prominent in the field of information technology because many present IT technologies are fast gaining popularity, while others are swiftly fading [4]. This discrepancy is one of the main reasons that a large number of university graduates are unemployed [5], because the job market requires a skilled workforce familiar with and capable of using current technologies. In 2015, Bitkom claimed that 7 out of 10 ICT companies (70%) stated that there is currently a shortage of IT professionals, whereas this rate in 2014 was 60%. In 2015,

978-1-6654-7613-3/22/$31.00 ©2022 IEEE

38.7% of IT jobs in the top 1000 German organizations were filled hardly due to a shortage of qualified and skilled candidates, and 6.1% of vacancies went unfilled due to a lack of appropriate candidates [4].

One of the critical issues in Iran is the high rate of unemployment among university graduates, which might be due to the failure of the country's education system to meet the demands of the job market. Understanding up-to-date skills and requirements needed by the job market helps job seekers develop their qualifications and skills to increase the chance of finding their favorite jobs. On the other hand, universities and educational institutions could benefit from job trends and skill patterns obtained from job data analysis to design and deliver effective education programs. Investigating and analyzing online job posts could reveal valuable patterns and trends related to job market demands.

In this study, IT job advertisements published on online job search websites in Iran are collected as a dataset utilizing the suggested data collection process. In addition, some frequent challenges with data gathering through crowdsourcing will be discussed. Finally, the specification of the collected dataset (called IranITJobs2021) will be presented.

The paper is structured as follows. In Section II, related studies about crowdsourcing processes are described. Section III, explains the dataset gathering process, including phases and activities. Sections IV and V describe the IranITJobs2021 dataset and conclude the paper, respectively.

II. Related Work

Crowdsourcing benefits from the crowd to accomplish two types of tasks. One type consists of those tasks that require human intelligence to solve complex problems, and the other type includes tasks that could be performed more efficiently using a crowd rather than experts [6]. A general process for crowdsourcing that incorporates four stages, namely designing incentives, collecting and assuring quality, verifying, and aggregating the received information, is presented in [6].

In [7], the authors proposed another general-purpose crowdsourcing process, incorporating initializing, implementation, and finalizing steps. The initialization step comprises task design, task breakdown, and incentive design activities. Finding the crowd and task assigning are two activities that form the implementation step. Quality control and aggregation stages are related to finalizing step.

In [8], a survey of the literature on crowdsourcing has been presented, categorized based on applications, algorithms, performances, and datasets. The crowdsourcing applications are voting systems, information sharing systems, games, and creative systems. Some research has been conducted to analyze crowdsourcing performance in terms of user participation, quality management, and cheating detection. Algorithms could be developed for implementing crowdsourcing systems more effectively. Examples of crowdsourcing algorithms are game-theoretic models for various human computation designs, algorithms for quality management, modeling the completion time as a stochastic process, etc. Furthermore, the authors introduce some crowdsourcing datasets available for further research.

The applications of crowdsourcing in machine learning could be categorized into four main areas: data generation, model evaluation and debugging, hybrid intelligence systems combining human and machine powers to extend the capabilities of AI, and crowdsourced behavioral experiments. The latter could enhance understanding of how humans interact with machine learning systems and technology more widely. Data generation applications of crowdsourcing include labeling, creating transcriptions, translations, and image annotations [9].

Applications of crowdsourcing in data mining have been highlighted in [10]. Classification, clustering, semi-supervised learning, association rule mining, sampling, and validation are data mining tasks that could be accomplished by crowd workers instead of data mining algorithms. For example, distinguishing males and females from social network users is a classification task that could be done by the crowd.

Data collection is a significant bottleneck in machine learning that has become a critical issue recently because of two main reasons. First, as machine learning is becoming more widely used, new applications emerge that do not necessarily have enough labeled data. Second, unlike traditional machine learning, the main advantage of deep learning techniques is generating features automatically, which saves feature engineering costs. But in return, deep learning methods may require larger amounts of labeled data [11].

The role of the crowd in data processing are supplying data to a system, labeling data, and verifying the work of other people or the results of an algorithm. Crowdsourcing is applicable in data management tasks such as data gathering, data integration, data cleaning and validation, operator evaluation, querying, and search [12].

Some challenges of data crowdsourcing have been reported in [13]. For instance, crowd workers are slower and more expensive than computers, and they might generate incorrect answers because they are not qualified or may be biased. Another issue is the existence of spammers who want to make money without any effort. In addition, splitting the problem into tasks and managing a crowdsourcing system could be challenging [13].

In [14], quality, cost, and latency controls have been considered the main problems in crowdsourced data management. The authors categorize tasks as: single choice, multiple-choice, rating, clustering, and labeling in real-world crowdsourcing platforms. The requestors must decide pricing, timing, and quality control as task settings based on their requirements.

One of the most challenging aspects of data analytics projects is data collecting. This task has a direct impact on the quality of extracted patterns, models, and knowledge. With the widespread use of the Internet, crowdsourcing has become a common form of data collecting. Daniel McDuff et al. [15], attempted to crowdsource a collection of human facial picture data to analyze and detect the intensity of their smile. They employed a website to present one of the three most popular promotional videos to the visitors. While visitors were watching the video, their facial pictures were captured by a webcam. The user was requested to answer three questions at the end of the video. Did you like the video? Have you seen this video before? Would you like to watch this video again? They compared the gathered dataset to the laboratory's existing datasets after gathering the photos. The authors concluded that the data gathered by crowdsourcing would have the same quality as those collected in the laboratory [15].

In [16], a news query classification dataset has been generated and validated using Amazon's Mechanical Turk for the labeling dataset. The authors handled two challenges in their research: the workers' lack of information about the news stories, and ensuring the quality of crowdsourced labels. The first challenge was addressed by integrating the news-related content into the labeling interface. The second one was handled by supplying the crowd with Web search rankings or related news article content for the query.

In the study [17], a new crowdsourcing method for generating labeled datasets for machine learning has been proposed called Revolt. The authors have compared their method to traditional crowd labeling methods and have demonstrated that Revolt enables reducing the amount of crowd training and efforts required to learn label guidelines.

A "Hollywood in Homes" approach has been introduced in [18] to collect activity data using crowdsourcing. The video creation process, including script writing, video recording, and annotation, was accomplished by crowd workers to form a dataset called Charades. The dataset consists of videos recorded by people in their homes acting out casual everyday activities. Charades incorporates 9,848 annotated videos with an average length of the 30s displaying activities of 267 people. Each video is annotated by multiple free-text descriptions, action labels, action intervals, and classes of interacted objects. The dataset could be applied for action recognition and automatic description generation.

In research [19], crowdsourcing has been employed to fill missing values in a tabular dataset. They presented a T-Crowd system to integrate each worker's answers on different attributes to learn their trustworthiness and the true values of categorical and continuous attributes effectively. The authors have shown that their proposed method could improve the quality of truth inference while reducing crowdsourcing costs.

In a crowdsourcing project, the number of people involved could be quite enormous. Without pre-defined rules, frameworks, and processes, managing these participants could be challenging. In [1], the authors provide a methodological framework for crowdsourcing. This research offers tips for crowdsourcing projects, however, it focuses mainly on the topic, "How can we persuade people to join in the crowdsourcing?" using The Motive-Incentive-Activation-Behavior model of crowdsourcing or MIAB, which is a model for aligning motives and incentives.

Data cleaning is inevitable in crowdsourcing data collection projects because these projects usually involve the participation of a massive number of people. Missing values, outliers, and noisy data, typographical errors, incomplete data, and the existence of values in multiple languages are issues that need to be addressed when data is collected. In [20], the author examines the errors in massive database data and proposes strategies to correct them. There are six possible strategies, for example, to handle the problem of missing values as follows: ignore the tuple, fill the missing value manually, use a global constant to fill the missing value, use the attribute mean to fill the missing value, use the attribute mean for all samples belonging to the same class as the given tuple, and use the most probable value to fill the missing value. Moreover, the paper provides solutions to the noisy data problem, such as binning, regression, and clustering.

In this paper, we propose a process for dataset gathering based on crowdsourcing. The presented process includes three pre-gathering, gathering, and post-gathering phases. The phases and activities of the proposed process could be followed to generate high-quality datasets suitable for data analytics and machine learning.

III. Dataset Gathering Process

In this section, the proposed dataset gathering process is described, including pre-gathering, during the gathering, and post-gathering phases. Fig. 1, displays phases and related activities that will be described in detail in the following sections.

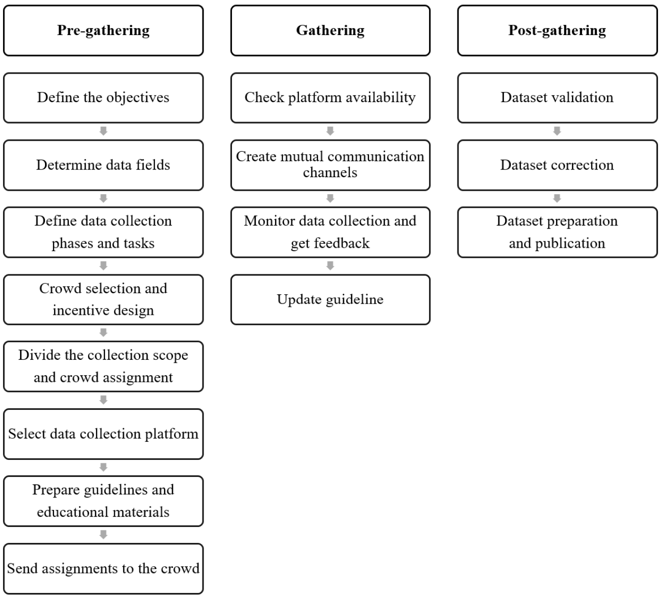
Pre-gathering
Define the objectives
Determine data fields
Define data collection phases and tasks
Crowd selection and incentive design
Divide the collection scope and crowd assignment
Select data collection platform
Prepare guidelines and educational materials
Send assignments to the crowd
Gathering
Check platform availability
Create mutual communication channels
Monitor data collection and get feedback
Update guideline
Post-gathering
Dataset validation
Dataset correction
Dataset preparation and publication

Fig. 1. Dataset gathering process

A. *Pre-gathering*

This stage corresponds to designing and planning crowdsourcing activities for dataset collection. The activities of the pre-gathering phase are explained in the following subsections.

1) Define the objectives

Before gathering any dataset, the purpose of collecting the dataset should first be determined. The answer to the question "What results do we want to find after gathering data and analyzing it?" could help define the objectives. At this point, a brainstorming session is started, and a list of all potential outcomes from dataset analysis is made. Some objectives may seem complex and challenging at first glance, but by breaking them down, they could be made more manageable. "Face recognition in live video streaming," for example, may seem sophisticated at first glance. Considering that a live broadcast comprises a large number of images, the first objective could turn into "face recognition in a photo." Even if face recognition in a photo looks difficult, each photo could be considered a set of color pixels for finding a solution to process each pixel.

The principal purpose of collecting IranITJobs2021 is to "discover hidden patterns in IT job postings published on online job websites". The main objectives of creating the IranITJobs2021dataset are to analyze:

- What skills are essential to be employed in governmental and non-governmental IT companies?
- What are the current trends in the IT job market?
- How much do companies allow remote work in the case of IT jobs?
- What are the most popular IT job opportunities in each location?
- Is there any gender difference in IT job requests?
- What are the skills dependencies for each IT job category?

2) Determine data fields

Once the objectives were identified, the data fields needed to meet the data gathering objective should be determined. To collect IranITJobs2021, the specified data fields are displayed in Table 1.

3) Define data collection phases and tasks

Data collection by crowd workers could be performed in multiple phases regarding defined objectives. Each phase may contain several tasks that should be completed by participants. Initial data collection, evaluation, modification, and labeling are examples of such tasks. Hence, in this step, phases and related tasks should be defined.

Regarding task complexity, there are four kinds of tasks, namely micro-task, complex task, macro task, and creative task. Task decomposition methods include sequential, parallel, recursive, iterative, and hybrid task implementation (For more details, please see [21]). Task complexity and decomposition method are required to be decided in this step.

The collection of IranITJobs2021 has been divided into three phases. In the first phase, each participant was requested to complete one task: to create the initial dataset, they must gather 30 job advertisements and ensure that none of them are duplicated. The second phase is to extract data from the collected job postings and complete dataset fields. The third phase corresponds to data validation and correction.

TABLE I. IRANITJOBS2021 DATA FIELDS

Column name	Data type	Description
Company name	String	-
Company type	Nominal	Type of company: governmental, non-governmental, or non-company
Ad date	Date	The date that the job advertisement is posted
Ad title	String	Example: hiring a senior Android programmer
Remote work	Boolean	Yes, if the company allows remote work, otherwise, No.
Location	String	The city/distinct where the job seeker is supposed to work
Knowledge enterprise	Boolean	Yes, if the company is knowledge-based, otherwise, No.
Part-time	Boolean	Yes, if the company supports part-time recruitment, otherwise, No
Gender	Nominal	Requested gender: Female, Male, or Both
Military service	Boolean	Yes, if the job seeker has an option to work in a company in exchange for doing his military service, otherwise, No.
Project-based	Boolean	Yes, if the company supports project-based recruitment, otherwise, No.
Ad text	String	The whole text of the advertisement
Keywords	String	List of technologies and skills required for a job

4) Crowd selection and incentive design

The crowd selection is performed to select appropriate crowd workers for a specific task. At this step, the required features and number of crowd workers should be determined. Then crowd workers are selected regarding features needed. Demographic factors, level of knowledge, ethnicity, worker experience, level of required training, familiarity with the collection platform, etc., are examples of worker features.

Because crowdsourcing involves human collaboration, people should be given some incentives to be engaged in completing tasks. Hence, incentives are decided and planned to stimulate crowd workers. Extrinsic incentives could be categorized as entertainment, social recognition, and financial compensation [6].

IranITJobs2021 dataset is related to the IT industry, so the crowd workers should be familiar with IT terms, such as programming languages and associated technologies. 50 BSc computer and IT engineering students of the University of Isfahan are chosen to collect IranITJobs2021. Data collection is a part of the final project related to the “Fundamental of Information Technology” course. Therefore, the project point is considered an incentive for crowds to accurately complete the tasks.

5) Divide the collection scope and crowd assignment

To prevent duplicate work and manage the crowdsourcing process correctly, it is necessary to divide the collection space into several scopes and assign scopes to specific participants. Regarding IranITJobs2021, the job advertisements could be divided according to the location of the workplace, and 30 instances of a location were assigned to each crowd worker.

6) Select data collection platform

A proper data collection platform should be chosen for crowdsourcing-based data gathering. In the past, data were collected through paper questionnaires and in person, but today, with the spread of the Internet and the creation of various platforms such as Amazon Mechanical Turk, the required data can be easily collected worldwide. To choose a suitable platform, the following tips should be considered:

- First, the data format needed to be collected should be determined. For example, if the data is only textual and numerical, Google Sheets is a good option. But if it is required to collect photos, files, and similar data, Google Forms could be suitable. If the format of requested data is such that there is no appropriate platform for collecting it, a website or a mobile application should be developed for data collection.

- Choose a platform that is easy to work with. Of course, in this case, the level of knowledge of the people participating in the crowdsourcing should be considered, and the appropriate platform is selected accordingly. For example, attendees may not be able to work with Google Sheets, so a more straightforward platform such as notepad should be chosen.

- In crowdsourcing, crowd workers might have different writing habits when collecting data. On the other hand, the defined data format should be followed when entering data. Suppose one of the required fields is the date. Different people may enter the date in different ways. To minimize user errors as much as possible, the platform could limit data entry to conform to the predefined format.

- Since the dataset is collected by the crowd, it is likely that duplicate information is entered into the dataset because of poor communication or management of the crowd. To avoid this problem, it is recommended to select a platform that can recognize duplicate data and prevent users from entering it. If the chosen platform does not support this feature, a specific column (such as customer id, company name, etc.) could be added to the dataset to detect and remove duplicate data.

Google Sheets platform was applied to collect IranITJobs2021 because of its ease of use, popularity, publicity, and appropriate output format for data collection.

7) Prepare guidelines and educational materials

To collect a dataset by crowd workers effectively and efficiently, it is vital to prepare clear guidelines to inform participants about crowdsourcing phases and tasks. In this instruction, the type of data required for each column and how to enter the information, crowdsourcing platform, validation process, and so on should be clarified. Moreover, the crowd workers should be educated about their tasks, so educational material such as video tutorials, workshop content, slide decks, social network posts, or online learning management systems content should be prepared.

Here are some suggestions for developing a guideline to collect the dataset.

- Specify a single language for the data. For example, the instruction states that all columns should be filled out in English.

- A column's values may consist of just a few constant values. For example, the gender column can only have two values for Male and Female. In such cases, it is necessary to control the user's input using the platform selected for data collection and prevent the user from making any mistakes. But if the chosen platform did not support this option, it should be explicitly stated in the instruction to enter only the values of "Male" and "Female". Try to keep the selected words as simple and short as possible to reduce the possibility of error. For example, the letters 'M' and 'F' could represent males and females respectively.

- Design a special template for short text columns. For example, the following template was applied for the "Ad title" column in the IranITJobs2021 dataset:

  (Field of Work) (Skill Level) (Position)

  Example: Android Senior Developer

- If there is a column with multiple values, first try to select a field using the platform features so the user can enter the values one by one with a convenient and simple user interface. If the chosen platform does not support this feature, the instructions should state the value separation character. For example, the "keywords" column contains several short words that can be split using the '|' character. Notably, the selected character should not be presented in the options that the user enters. For example, the character ',' may not be appropriate in some cases.

- For multi-part words, it is recommended to specify a character or standard to divide the parts. For example, the react js library consists of two parts, so the word must be entered in the dataset as react.js or react-js.

8) Send assignments to the crowd

Once the data collection guideline is prepared, the document should be available to the participants. Furthermore, educational materials are delivered to crowd workers, and training workshops, meetings, or learning management system sessions are held. A convenient and straightforward communication way to solve any problems is provided to participants. It is also necessary to inform each crowd worker about their data collection scope.

B. Gathering

1) Check platform availability

Some problems may occur when data collection begins. One problem is the unavailability of the data collection platform. The platform chosen for data collection may be temporarily or even permanently unavailable for various reasons, including internet disconnection, expiration of the platform (for non-free platforms), technical issues, and so on. Therefore, the availability of the platform should be periodically checked, and the collectors should be notified as soon as any problem occurs and after the issue is addressed. It is also necessary to periodically back up data collected by the crowd.

2) Create mutual communication channels

In some cases, participants may need to keep abreast of news and information related to the collection process. For example, it is likely to make changes to the fields of the dataset, in this case, the participants should be informed about the changes. Questions or ambiguities may also arise for participants, hence, it is necessary to consider a way of communication with collectors. In general, depending on the people chosen for the crowdsourcing, it is a good idea to set up a two-way communication channel with them. This communication channel may be an information channel on social network services, social messaging platforms, or even an email.

3) Monitor data collection and get feedback

Participants may make mistakes, and the quality of the final dataset may be reduced due to errors and ambiguities in the collection process and guidelines, or issues that were not previously considered in the collection process. So, after some data has been collected, it is necessary to check it to ensure it is of good quality. Collectors' feedback during data collection is precious and helpful in solving problems and improve the collection process.

4) Update guideline

It is possible to see that the participants committed mistakes with a recurring pattern after each review of the collected dataset and also according to the participants' comments. Such problems may occur because error, vagueness, and shortage exist in the guideline. Consequently, the guideline might be edited and updated, and the participants should be notified about the changes.

C. Post-gathering

1) Dataset validation

Despite crowd education, preparing guidelines, and controlling data collection, there may exist some inconsistencies, errors, and missing values in the collected dataset. As a result, it is vital to review the data once it has been collected. The validation could be performed by experts, the crowd, or both. If the crowd is supposed to validate the dataset, it is suggested that each participant validates some of the data that they have not collected. For this purpose, it is necessary to define the validation phase and related tasks and assign data to the crowd for evaluation. Expert validation is more expensive than crowd evaluation, but it may be more accurate and effective.

2) Dataset correction

After reviewing the dataset by experts or the crowd, a correction step is needed. Correction operations are generally unavoidable in most data-related projects. The following are some of the inconsistencies in the data collected, as well as ways to correct them.

a) Existence of values with different languages and typographical mistakes

In the guideline document, the language of the dataset should be determined and followed by crowd workers, however, values with different languages may be included in the dataset, and spelling errors may exist. For example, although the IranITJobs2021 guideline document stated that the advertisement's title should be in English, some collected ads had Persian titles. To solve this problem, sentences containing Persian words were first identified using regular expressions. These sentences were then standardized using the Normalizer class of the Parsivar library. The benefits of this class include converting Persian numbers to English and eliminating extra spaces [22]. After standardization, the SpellCheck class of this library was employed to correct possible spelling mistakes. Finally, these sentences were translated into English using Google Translate library.

b) The presence of empty cells in some columns

Empty cells are prevalent in datasets collected through crowdsourcing. This is because people may become confused in some cases and leave some cells empty instead of connecting with the support team and clearing up the ambiguity. One way to handle missing values is to re-enter data by experts or another crowd worker. Another way is to fill the missing value automatically using mean, median, most frequent, or constant value. In the IranITJobs2021 dataset, the missing values were completed using the most frequent approach.

c) Failure to follow the instructions recommended in the guideline

Crowd workers may not follow the dataset collection guideline completely. In such cases, it may be required to delete or correct some records. For example, in the IranITJobs2021 dataset for a keyword column, the crowd workers were asked to add at least five keywords for each ad, but, some records had less than five keywords. To solve this problem, a column for saving the whole text of the ad was considered to extract the keywords after the data collection. Furthermore, some crowd workers did not follow the multi-part words standard declared in the guideline. Therefore, based on a dictionary of multi-part words of IT terms, these words are corrected as a defined format.

d) Low-density values in samples

In some cases, a data field may have a wide range of values, so there may be very few instances of each value in the dataset. In these cases, a common feature between low-density values could be considered, and the values could be generalized. For example, after a careful review of the IranITJobs2021 dataset, it was found that while the diversity of cities is relatively high, the quantity of ads in small cities is very low. To address this problem, it was decided to generalize cities to the corresponding distinct.

3) *Dataset preparation and publication*

After collecting, validating, and correcting the dataset, the dataset is published. The dataset should be uploaded to an appropriate source, and a dataset description should be prepared, including a description of each column, the number of samples, the date of collection, etc. This document could be made available to the public as a paper, or dataset description document. It is also possible to make several versions of the dataset. These versions may vary in terms of language, number of instances, etc.

IV. IranITJobs2021 Dataset description

During the data collection stage, 1300 ads related to information technology from August 2019 to January 2021 were collected. These ads are collected from Iranian online job search websites. These websites are as follows: quera.ir, jobinja.ir, karboom.io, cheragh.com, jobvision.ir, e-estekhdam.com, iranestekhdam.ir, karinsoo.com, computerjobs.ir, javacup.ir, and daneshkar.net. Each instance in the IranITJobs2021 dataset incorporates 13 fields. The statistics of the dataset fields are listed below. The collected dataset for public use can be downloaded from the following link:

https://github.com/mrezasaeidi/IranITJobs2021/blob/main/IranITJobs2021.xlsx?raw=true

Fig. 2, displays advertisements distribution per distinct. Tehran, Esfahan, Khorasan, and Fars are top locations regarding the number of ads. Fig. 3 to 8, display advertisements distribution per cooperation type, gender, knowledge enterprise, military service support, remote work allowance, and project-based recruitment, respectively. Fig. 9 to 14, show advertisements distribution based on different fields per four top locations.

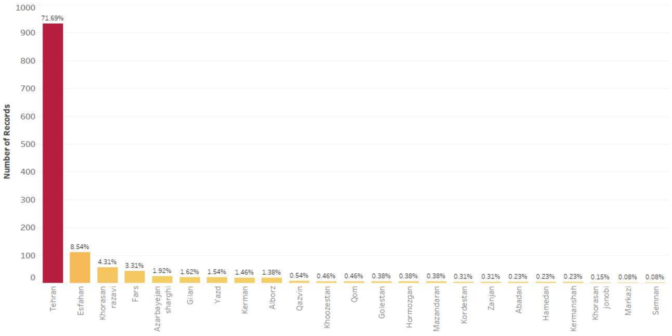

Fig. 2. Ads distribution per location

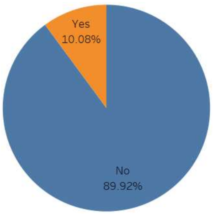

Fig. 3. Ads distribution per part-time work allowance value

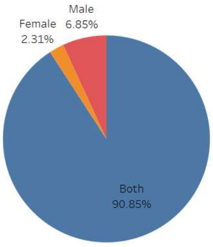

Fig. 4. Ads distribution per gender value

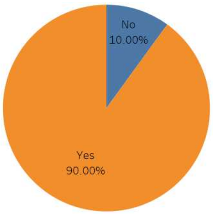

Fig .5. Ads distribution per knowledge enterprise value

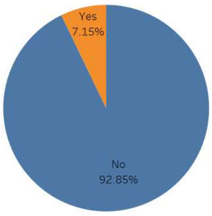

Fig. 6. Ads distribution per military service support value

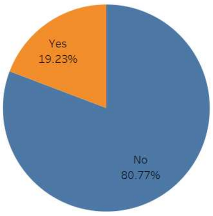

Fig. 7. Ads distribution chart per remote work allowance value

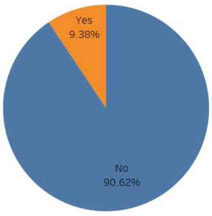

Fig. 8. Ads distribution per project-based recruitment value

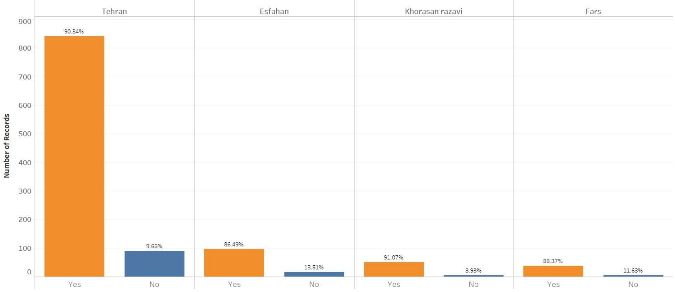

Fig. 9. Ads distribution based on part-time work allowance per top location

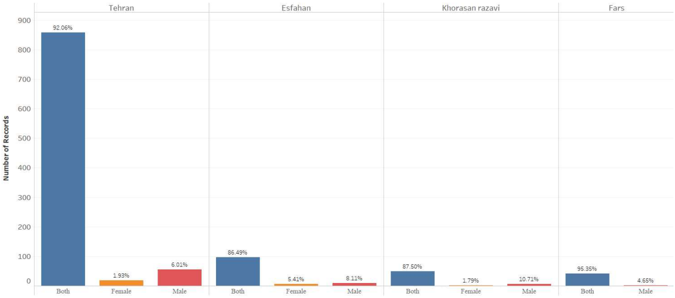

Fig. 10. Ads distribution based on gender per top location

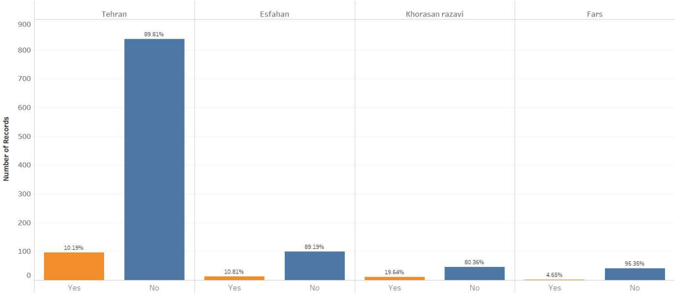

Fig. 11. Ads distribution based on knowledge enterprise per top location

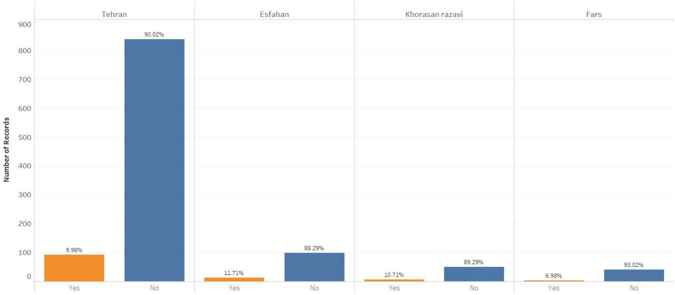

Fig. 12. Ads distribution based on support military service per top location

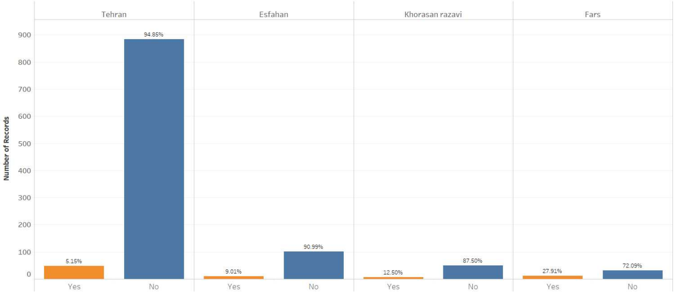

Fig. 13. Ads distribution based on project-based recruitment per top location

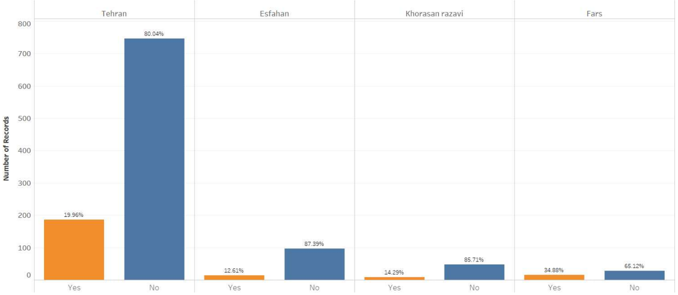

Fig. 14. Ads distribution based on remote work allowance per top location

V. Conclusions and Future Work

Correcting datasets after crowdsourcing-based data collection could be very complex and costly. The volume of correction operations can be significantly reduced if a proper and pre-planned process is employed for data collection. This process should include all stages of the project, including before, during, and after gathering because all these steps will have a direct impact on the quality of the final dataset. In this paper, a process for collecting and preparing datasets using crowdsourcing was presented. Furthermore, a set of IT job advertisements published on Iranian online job websites was collected to form a dataset through the proposed process. Based on the proposed crowdsourcing process, a platform specially designed for collecting datasets could be developed.

In the future, we are going to collect more job postings about IT and other industries to create the next versions of IranITJobs2021. Moreover, we will analyze the keywords and the text of the ads of IranITJobs2021, and separate the ads based on different areas of work and the skills required for each field. In addition, using data analytics methods, a roadmap for learning the skills needed for each job could be extracted and suggested to job seekers.

References

[1] M. Keating and R. D. Furberg, "A methodological framework for crowdsourcing in research," in Proceedings of the Federal Commite on Statistical Methodology Research Conference, 2013.

[2] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, "Towards an integrated crowdsourcing definition," Journal of Information Science, vol. 38, no. 2, pp. 189–200, Apr. 2012, doi: 10.1177/0165551512437638.

[3] M. Lease, "On Quality Control and Machine Learning in Crowdsourcing.," in Workshops at the Twenty-Fifth AAAI Conference on Artificial Inteligence, Jun. 2011.

[4] J. Grüger and G. J. Schneider, "Automated analysis of job requirements for computer scientists in online job advertisements," in WEBIST 2019 - Proceedings of the 15th International Conference on Web Information Systems and Technologies, 2019, pp. 226–233. doi: 10.5220/0008068202260233.

[5] S. A. M. Nasir, W. F. Wan Yaacob, and W. A. H. Wan Aziz, "Analysing Online Vacancy and Skills Demand using Text Mining," in Journal of Physics: Conference Series, Jun. 2020, vol. 1496, no. 1. doi: 10.1088/1742-6596/1496/1/012011.

[6] L. Nassar and F. Karray, "Overview of the crowdsourcing process," Knowledge and Information Systems, vol. 60, no. 1, pp. 1–24, Jul. 2019, doi: 10.1007/s10115-018-1235-5.

[7] A. I. Chittilappilly, L. Chen, and S. Amer-Yahia, "A Survey of General-Purpose Crowdsourcing Techniques," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9, pp. 2246–2266, Sep. 2016, doi: 10.1109/TKDE.2016.2555805.

[8] M.-C. Yuen, I. King, and K.-S. Leung, "A Survey of Crowdsourcing Systems," Jan. 2012, pp. 766–773. doi: 10.1109/passat/socialcom.2011.203.

[9] J. Wortman Vaughan, "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research," 2018. [Online]. Available: http://jmlr.org/papers/v18/17-234.html.

[10] G. Xintong, W. Hongzhi, Y. Song, and G. Hong, "Brief survey of crowdsourcing for data mining," Expert Systems with Applications, vol. 41, no. 17. Elsevier Ltd, pp. 7987–7994, Dec. 01, 2014. doi: 10.1016/j.eswa.2014.06.044.

[11] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective," Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.03402

[12] V. Crescenzi, A. A. A. Fernandes, P. Merialdo, and N. W. Paton, "Crowdsourcing for data management," Knowledge and Information Systems, vol. 53, no. 1, pp. 1–41, Oct. 2017, doi: 10.1007/s10115-017-1057-x.

[13] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in Data Crowdsourcing," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 4. IEEE Computer Society, pp. 901–911, Apr. 01, 2016. doi: 10.1109/TKDE.2016.2518669.

[14] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced Data Management: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9, pp. 2296–2319, Sep. 2016, doi: 10.1109/TKDE.2016.2535242.

[15] D. McDuff, R. el Kaliouby, and R. Picard, "Crowdsourced Data Collection of Facial Responses," in Proceedings of the 13th International Conference on Multimodal Interfaces, 2011, pp. 11–18. doi: 10.1145/2070481.2070486.

[16] R. M. C. Mccreadie, C. Macdonald, and I. Ounis, "Crowdsourcing a news query classification dataset," in Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, 2010.

[17] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in Conference on Human Factors in Computing Systems - Proceedings, May 2017, vol. 2017-May, pp. 2334–2346. doi: 10.1145/3025453.3026044.

[18] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding," 2016, pp. 510–526. doi: 10.1007/978-3-319-46448-0_31.

[19] C. Shan, N. Mamoulis, G. Li, R. Cheng, Z. Huang, and Y. Zheng, "A Crowdsourcing Framework for Collecting Tabular Data," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 11, pp. 2060–2074, May. 2019.

[20] W. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," Journal of Engineering and Applied Sciences, vol. 12, pp. 4102–4107, Jun. 2017, doi: 10.3923/jeasci.2017.4102.4107.

[21] S. S. Bhatti, X. Gao, and G. Chen, "General framework, opportunities and challenges for crowdsourcing techniques: A Comprehensive survey," Journal of Systems and Software, vol. 167, Sep. 2020, doi: 10.1016/j.jss.2020.110611.

[22] S. Mohtaj, B. Roshanfekr, A. Zafarian, and H. Asghari, "Parsivar: A Language Processing Toolkit for Persian," in Proceedings of the Eleventh International Conference on Language Resource Evaluation, 2018.