Analysis, Insights and Visualization of LinkedIn Profile Data

Samanvitha B[1], Hanita Jain[2], Dr.Venkata Ramana[†]
Department of Mathematics
Vellore Institute of Technology
Vellore, India

***Abstract*—LinkedIn is a social networking site, often used as a professional platform to gain knowledge, develop one's skillset and also to grab opportunities from across the world. The motive of this paper is to analyse a set of people in LinkedIn who are from the Data Science industry. We want to see which field of Data Science is growing, what skills are needed to grow in a particular field, where to look for data-driven opportunities, which company we can approach for our chosen field, whom to follow with respect to our interests and skills, etc. If we look at the statistics, we come to know that 80 percent of the text data collected by any organization or an individual is unstructured and so is our dataset. We have analyzed our unstructured text data by involving steps such as – data cleaning, text mining and visualizing those results with the help of various tools and software like MS excel, Power BI, Tableau and Python. We categorize the LinkedIn dataset into different fields in data science in order to see what field the majority of people are into. We have made use of bar-charts, bubble-plot, treemaps, donut- chart, network GI graph and wordclouds to analyze and present the relationship between our data profiles and the variables. This paper's aim is to help the students in the field of Data Science by giving them information about the industry through a social platform LinkedIn.**

***Index Terms*—LinkedIn, Data Cleaning, Visualization, NLP, Python, PowerBI, Tableau.**

I. INTRODUCTION

LinkedIn, is one of the biggest online professional platform to share and gain knowledge on various content around the world. In this paper, we will be analyzing the data extracted from 120 profiles that are based on one's connections in the field of Data Science. We have collected the data from site in excel sheet by adding an extension to the browser. We will be doing the analysis by using the parameters name, location, summary, organization, organization description, experience, education, university, skills, interests, industry, followers, and mutual connections. In the above mentioned parameters, name column has the full name of the user, location has the city names of organizations that user is currently working at. Organization has the company names, followed by the description of user's role in the company and one's experience in the industry. Industry contains what domain the user belongs to. Education has the highest degree earned by the use and the respective university. Skills and interests columns are the areas of expertise of the users. Mutual connections columns are the ones that show relation between all the connections in the data set. The reason we have chosen LinkedIn data is because it is one such global service that brings people of all ages and background together to give and take information as much as one can from each other. Also, it is an authentic system with very low percent of fraudulent cases and very high percent of sincere cases. We have used tools, techniques and programming languages like Excel for data cleaning, Python and Power BI for data analysis, Tableau for visualization and also to bring out all possible, useful results that can help the ones interested in the field of Data Science.

II. LITERATURE SURVEY

As we have mentioned before, usually 80 per cent of any collected data is unstructured. And to analyze such data, we first need to clean it so that it is noise-free, has no missing values, and has relevant as well as consistent parameters to get some meaningful insight during the analysis part. In one such paper by Wangikar and Deshmukh [1], we came across the current approaches and algorithms used for data cleaning and they explained how each data is of a different kind and one type of cleaning procedure may not be ideal for a different data. We could relate to this paper in the sense that we had to use different ways to clean different parameters of our data. Once the text data is organized a bit, we could start with text pre-processing to remove the noises. Kannan and Gurusamy [2], in their research paper walks us through the steps of pre-processing the text data by incorporating techniques like removal of stopwords, tokenization, stemming, etc. And once we have our data in the required form, we jump to analyze it. Some techniques for data mining and their applications are explained by Talib et al. [4]. They discussed how to extract relevant information from text data along with the challenges that arise during the process. After analyzing the results, we proceed towards data visualization, which is a technique of presenting the data in the forms of graphs or pictures to make the results more comprehensible and effective, as discussed by Sadiku et al. [5]. According to them, the visualization helps to communicate the complex findings in a more accurate, clear and efficient way. One such visualization tool that we have used is Power BI, and as discussed by Krishnan et al. [6], it provides a space for organizations and individuals to create reports, have your data

Fig. 1. WordCloud for 'Summary' Parameter

analyzed automatically within a matter of a short time with minimum efforts.

III. DATA ANALYSIS

A. Data Cleaning

We begin the analysis with removing the unnecessary columns from the downloaded excel file like first name, last name, URL of user's profile, address, birthday, phone number. Once we clear the data set of incompetent parameters, we are set to look at the left over parameters and analyze them, in order to get the required results.

B. Most Prevalent Branch of Data Science

The first outcome that we are looking for is what branch of data science is mostly prevailing among the users of our data science in the industry. For achieving this, we will be using the parameters 'summary' and 'organization descriptions' of all the 120 users. 'Summary' parameter contains the detailed explanation of user's journey in the industry. 'Organization descriptions' contain the explanation of user's job in the organization and the areas he has worked on.

For getting to know which branch and what the users are mostly working on in the industry, we have used Python and some of its libraries. We have used the library pandas which is built on numpy package. We used this in order to load our data set table from excel to Python and perform the essential manipulations to get our answer. We use the NLTK package which means 'Natural Language Toolkit' for natural language processing, which is a branch of AI that helps the computers in understanding, interpreting and manipulating human language.

First we remove all the punctuation marks, numbers and special characters from the text and change all the words into lowercase. Next we import the module stopwords from nltk.corpus package which is helpful in removing meaningless words like is, am, the, etc. from the text we loaded. Next we load the module RegexpTokenizer from nltk.tokenize in order to extract token (sequence of letters) from string using the regular expression. For example, we have the string 'Sky is blue.' Here the regular expression extracts every token as follows: 'sky', 'is', 'blue'. Following this, we install the package wordcloud in order to visually represent the text of the data we load in different sizes and colors, based on the count (number of times the word is repeating). Once we install all these libraries and packages in Python, we can call them and get the results. For our data set, we need the answer for the fields in data science that are most popular among the users. After the data undergoes the cleaning and extraction, we get the results as shown in figure 1 and figure 2.

Fig. 2. WordCloud for 'Organization Description' Parameter

C. People's Interests and Universities they come from

Once what branch is highly prevailing among the users is answered we wanted to know people from which backgrounds have chosen what kind of interests. We had people from different universities with different interests. For knowing in detail about this, we loaded the parameters 'universities' and 'interests' in Power BI which is a business analytics tool introduced by the Microsoft. We have analyzed which university is maximum attended and what are the common interests of users that attended those institutions in form of donut chart which is similar to pie chart showing which university is mostly attended by the users and their respective interests relationship. The results are shown below in figure 3.

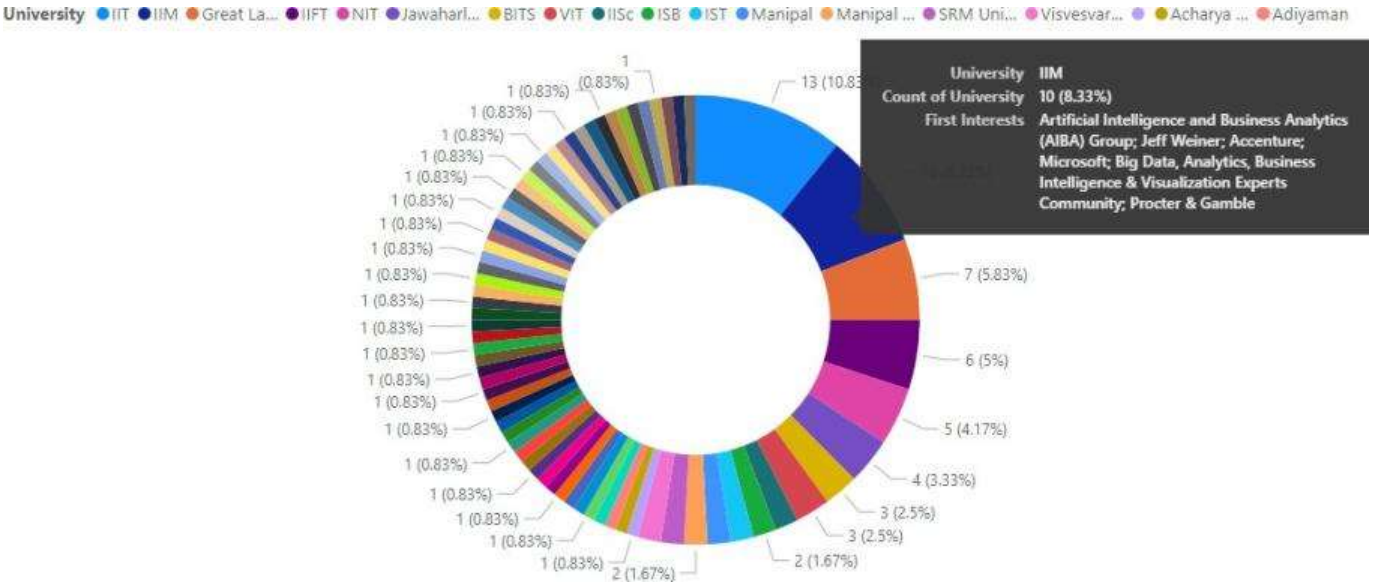

Fig. 3. Frequency Distribution for Universities attended by people and their Interests

D. Locations with most Data Science Recruitments

Once we got to know which university most of the users have attended, we wanted to know in which city the users were currently settled in. For that we have used the parameters 'location' in Tableau, which is a data visualization software that helps in analyzing the data loaded in it. For the purpose of knowing the city in which the maximum users are settled in, we have chosen the bubble-chart visualization which displays the data in the form of clusters of circles based on the frequency of users residing in that city. The results are as shown in figure 4.

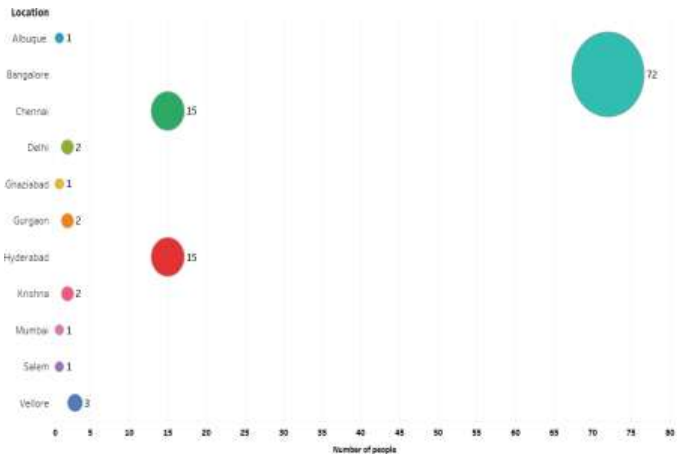

Fig. 4. Bubble Chart showing frequency of people living in that location

E. Educational Background and Skills

After the question of which city is answered, we got more curious. We wanted to know what sort of degree we needed, what sort of skills one needs in order to get in to the company one desires. For this purpose, we have taken the parameters 'location', 'organization', 'university', 'Education degree', 'Industry', and 'skills'. We chose Python platform to get a hold of this question. We have loaded the libraries pandas, numpy along with matplotlib that'll help us in analyzing and plotting our data. After these, we load the module plotly.express. We use this specific module in order to analyze and display our data in the form of treemaps. Treemaps charts help us visualizing hierarchical data in the form of nested rectangles. On choosing a particular rectangle, it shows you the information loaded in the rectangle. Considering the above parameters, we arrange the rectangles in the treemap as city in which the user is currently working in, and then we have the sub rectangles with all the organizations present in the cities. Following these we give the skills, degree, university, industry is displayed. For example, if one wants a job in Hyderabad under Microsoft, when you point at Hyderabad city with Microsoft rectangle, you will be able to see the skills, university, degree that one needs to be placed in that company. The results are as shown in the figure 5.

F. Mutual Connections

You see, the data set that we have is based on the connections of one person. We wanted to know if the people from the data set we have are connected, if so, how many of them are mutually connected. For this, we have used GIgraph which can be obtained by adding an extension of GIgraph in the Excel sheet. GIgraph is a network visualization tool that is used to explore the data by comparing, connecting and finding similarities and intensities. Here, we have used the parameter 'mutual connections' in order to see how many

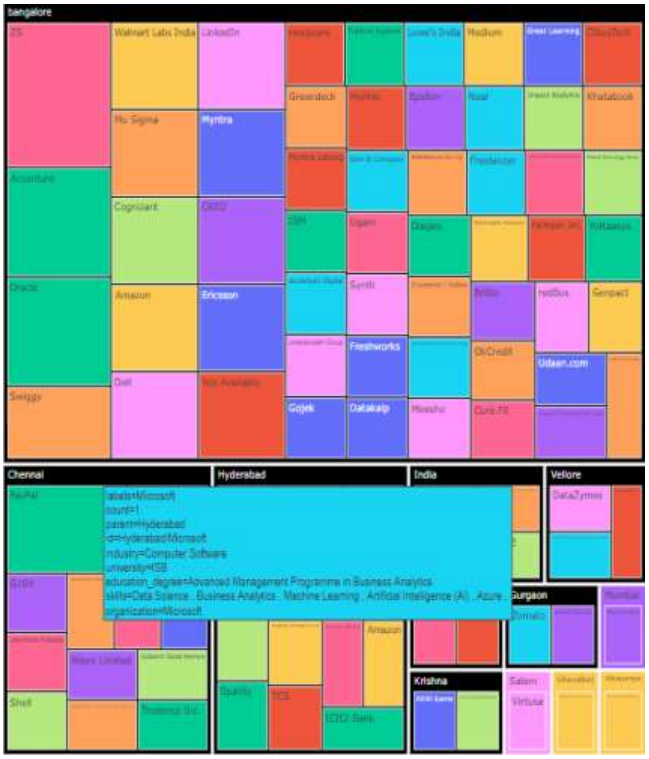

Fig. 5. Treemap presenting the companies based on their locations with people's University, Degree and Skills

of the users are connected to each other and which user is mostly common among the others. The results are shown in the figure 6.

G. Whom to Follow?

Once we know who is connected to whom, we wanted to figure out, which person to reach out to if we have a query or need any sort of discussion regarding an industry that one is interested in. For this, we have used Excel for data analysis and PowerBI for visualization. We have taken the parameters 'industry', 'years of experience', 'followers', and 'interests' into consideration. The column 'years of experience' is drawn from the columns 'organization start' and 'organization end' in the original data set. From these columns, we calculate the total number of days the user has worked in the industry. This is done by subtracting the present working with the very month/first year of the user. This gives us an idea about how long the user has been in the industry. This can be counted as experience. The next parameter we took was total number of followers of the user. We must note that in LinkedIn, we can increase our reach through connections or through followers. In some of the cells of this follower's column, the number is zero indicating it is a connection account. The rest of the accounts are follow accounts. By taking these two columns into account including 'industry' and 'interests', we visualize it using PowerBI. The software loads the data and considers the user with highest experience and followers and shows us which person to reach out in a particular industry. Once you point at the particular industry, the tooltips show you the name of the person with maximum number of followers and their experience in the industry and their interests.

It is a general notion that people consider a person with many years of experience may have higher number of followers or vice-versa, but we can observe from the chart that this is not always true. Years of experience may or may not give you the followers. Even if we have less experience but are skillful, we could gain followers. The results of this are shown in figure 7.

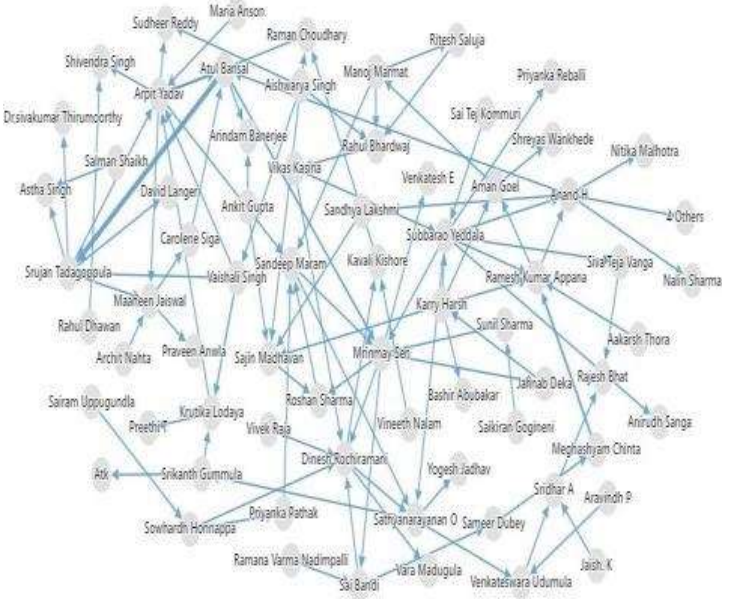

Fig. 6. Mutual Connections network graph

So far, we haven't come across as such analysis of LinkedIn data as discussed above. This analysis has helped us in understanding which fields of data science are prevailing and skills we need to learn in order to make it into the industry smoothly. It also helped us in seeing what skills, interests and qualifications the organizations looking for in the respective industry.

RESULTS

Summarizing the findings in this paper, we can say that according to our data of one person's LinkedIn connections in the field of data science, majority of the users are into the branch of machine learning, data analytics, deep learning, natural language processing, data visualization, computer vision, etc. all with the help of ML algorithms, predictive modeling, programming languages and software like python, Microsoft excel, SQL, C++, etc. While coming to the educational background, we notice most of the people hale from IITs, IIMs, Great Lakes, and NITs with interests and skills in AI, ML, DL, Healthcare, etc. Considering where most of

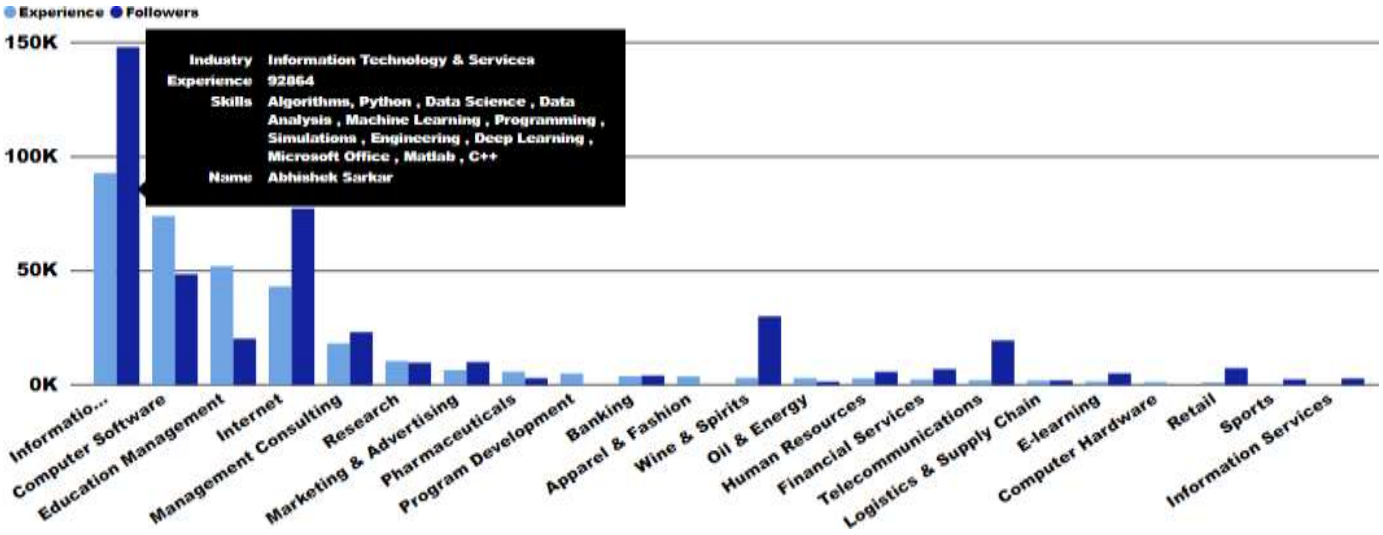

Fig. 7. Bar graph showing the person with maximum followers from each industry with their skills

the users are working in India, we find that Bangalore has the highest number followed by Chennai, Hyderabad and Delhi with IT&Services industry having the highest prevalence. On the basis of location, we can look into each city as a different level and say which organization has hired what number of users and with what educational background along with their skills and their respective industry. For example, in Hyderabad, Microsoft from Computer science industry has hired 1 user who has done Advanced Management Programme in Business Analytics from ISB with skills of ML, AI, Business analytics and Azure. Then we find the users with the maximum mutual connections, followers and calculate the experience in days for each user. We come to know that, high experience does not always guarantee you high number of followers and vice versa. Your skills and interests play a major role in where to look for a job and whom to connect according to that.

LIMITATIONS

One of the important limitations is that, these results are based on the connections of one person. This indicates that the results can be different for every person based on their connections. For example, for above considered person, maximum number of people settled is in the city Bangalore. But if you consider connections of some other person, it can be a different city. The results of mutual connections can also vary based on the person.

CONCLUSION

We have achieved the motto of this paper which is analyzing LinkedIn data to give necessary and useful results such as which domain is mostly being pursued, skills one needs in order to get placed in a particular organization and whom to follow up or ask when one has any query or discussion in a particular industry.

REFERENCES

[1] R. Deshmukh and V. Wangikar, “Data Cleaning: Current Approaches and Issues”, In Conference: IEEE International Conference on Knowledge Engineering At: Department of CS IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad ’01, 2011.

[2] V.Gurusamy and S. Kannan, “Preprocessing Techniques for Text Mining”, In Conference: RTRICS At: Podi, Project: Multilingual Natural Language Processing ‘10, 2014.

[3] A. Kadhim, “An Evaluation of Preprocessing Techniques for Text Classification”, International Journal of Computer Science and Information Security, vol. 16, no. 06, pp. 22-32, 2018.

[4] R. Talib, M.K. Hanif, S. Ayesha, and F. Fatima, “Text Mining: Techniques, Applications and Issues”, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 11, pp. 414-418, 2016.

[5] M. Sadiku, A. Shadare, S. Musa, C. Akujuobi, and R. Perry, “DATA VISUALIZATION”, International Journal of Engineering Research and Advanced Technology (IJERAT), vol. 02, no. 12, pp. 11-16, 2016.

[6] V. Krishnan, S. Bharanidharan, and G. Krishnamoorthy, “Research Data Analysis with Power BI”, In: 11th International CALIBER-2017, Anna University, Chennai, Tamil Nadu, pp. 211-218 ‘08, 2017.

[7] A. Bansal and A.K. Upadhyay, “Microsoft Power BI”, International Journal of Soft Computing and Engineering (IJSCE), vol. 07, no. 3, Jul, pp. 14-20, 2017.