Predicting Future Jobs Based on Scholarly Research Articles

Omer Bin Ali Bajubair - Z1905006
Mohammed Abdul Moyeed - Z1912165
Northern Illinois University

Abstract

For this Project, our proposal and area of concentration will focus on measuring the impact of scholarly research articles on job market. Our process includes gathering data from websites like Indeed, LinkedIn to see if an influx in research papers published has a correlation with the jobs being published in the market. We want to predict the future jobs that can be posted based on current research articles published. We are looking at measuring and quantifying the impact of scholarly research articles on Job Market. Ultimately, the question we would like to solve is if there is a correlation between publication of scholarly articles and the jobs being posted by the companies? Furthermore, are we able to predict the effect of an influx of research articles will have on the future job market. Example now due to covid 19 there has been lot of research papers being published in the field of genome sequencing and DNA sequencing so we are trying to find out if it is possible for us to predict will there be influx in jobs like Bio-informatics Engineer or Genetics Engineer roles etc. We will do topic modeling on the data after preprocessing the data and from the topic model analysis, the government, companies and other officials could benefit as they would have a clear picture on what the job market will be in future. There are many approaches to topic modelling that uses the LDA. We will also have various visualizations like dendrograms, heatmap and interactive plots to better understand the correlation between research papers and jobs published.

1 Introduction

The concept of employability has steadily gained importance in recent years. It is worth noting that applied research on employability has, nowadays, an exploratory approach. This is because this research area presents difficulties regarding to have adequate, reliable and updated data, as well as this early status not only prevents agreement on research results are reached, but also poses many questions as to which methodologies and approaches are most appropriate to address these issues. For these reasons, the area is still growing and need to push the outcomes to further research levels. These projects have faced at least two problems: first, the lack of a single, consensual definition of employability and, secondly, the difficulty of obtaining summary indicators to assess it. Indeed, employability is a theoretical construct whose definition varies according to academic discipline and the perspective used, as well as the socioeconomic context to which it refers. There is no clear consensus on the factors that compose or determine it, nor on the employment outcomes to which it leads.

Research papers are a repository of information on the various elements that make up science and technology RD activities. Generating knowledge maps based on research papers enables identification of specific areas of scientific and technical research as well as understanding of the flow of knowledge between those areas. Recently, as the number of electronic publishing and informatics archives along with the amount of accumulated knowledge related to science and technology has proliferated, the need to utilize the meta-knowledge obtainable from research papers has increased.

Hence we are interested in solving if there is a correlation between publication of scholarly articles and the jobs being posted by the companies? Furthermore, are we able to predict the effect of an influx of research articles will have on the future job market.

2 Problem Significance

There is lot of research that has been conducted how to get employed based on the characteristics and various details of students, finding out the Job Market in near future by finding the correlation between the Scholarly articles and current Job market is a relatively tough job for many reasons. The results of our research will help policy makers, investors, and companies make optimal decisions. For example, if we show that scholarly research articles are strongly correlated with a rise in Jobs, this will lead universities and other departments to invest more into research and development.The job market is the market in which employers search for employees and employees search for jobs. The job market is not a physical place as much as a concept demonstrating the competition and interplay between different labor forces. It is also known as the labor market. Job Market has been varying over the years due to many reasons, most recent one is Covid-19, The research on employability has been going on from a long time to predict how each factor has or will affect it. Using Machine Learning for predicting employability bring out a lot of new prospects for improvements on various fronts.There are many studies and publications [4] [5] [6] [7] [8] that find the co-relation between jobs and various factors, in this project we try to find if Scholarly research articles are one of the main factors that effects the Job Market. Scholarly research papers are defined as research being published in various fields.

3 Related Work

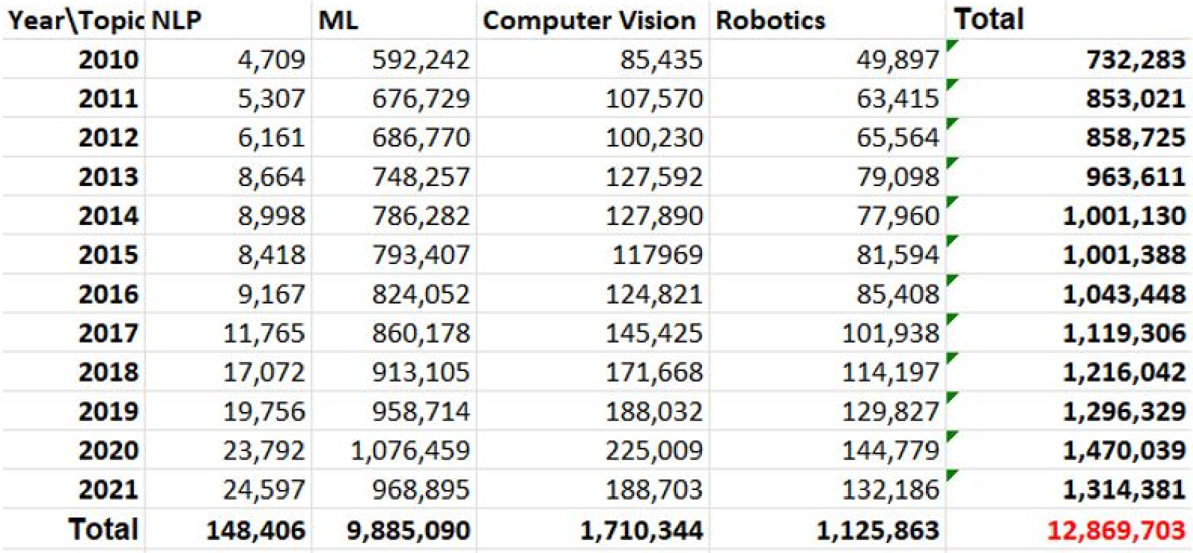
There were a lot of previous publications that focusses on the employment and what factors were more prominent in employability. Mark C. Berger [9] has predicted future earning based on the five broad fields of study, conditional logit model of various choice that incorporates alternative predictive earning variables are specified and estimated. G. Jason Jolley [10], discusses North Carolina's job market in 2020 and draws on national and state level industry and occupational employment projections. It also considers the shortcomings of these projections and suggests actions the state can take to capitalize on the projected employment growth areas. [11] The American Health Information Management Association (AHIMA) conducted a study to assess the future needs of the health information workforce. The study was intended to define the current reality of HIM within the healthcare industry, how the market is shifting to meet future needs, and what knowledge, skills, education, and credentials will be necessary to perform successfully as an HIM practitioner in the future. The study consisted of a survey of HIM and related stakeholders and multiple focus groups. This article summarizes key findings from the survey. [12] predicts the skill shortage in labor market based on a machine learning Approach. This research provides a robust data-driven approach for predicting and analyzing skill shortages, which can assist policymakers, educators, and businesses to prepare for the future of work. [13] Educational Data Mining (EDM) uses various data mining tools and methods for different applications in the field of education. EDM applications and techniques follow both pure and practical research objectives to enhance the learning process and to improve and develop learning quality. Educational data mining helps in forecasting the future patterns to make the organizations or institutions provide quality-based education to the learners. Educational institutions still struggle with graduation and employment toll.

[21] proposes and test a machine learning approach to research about employability and employment. To understand how the graduates get employed, researchers propose to build predictive models using machine learning algorithms, extracting after that the most relevant factors that describe the model and employing further analysis techniques like clustering to get deeper insights. In this researchers have built a model that could predict if a person will get employed or not, showing also what are the factors that affect more the result. After that, researchers could use these factors to filter information, generalize knowledge across the dataset, extract what values on these factors lead to get employed or not, etc. As an example, using these most relevant factors, researchers could clusterize students to gain deeper knowledge about what are the main characteristics between those who get an employment and those who do not. Despite of the complexity of the data, and the different issues with the dataset, researchers have been able to get a predictive model with a 0.71 precision score (0 the worst score, 1 the best one). This result opens the possibility of keep working in this approach to enhance the results and try deeper analysis. Regarding other scores achieved like, F1, recall, etc., should be outlined that the model built performs poorly in detecting the real "False" values (not employed students), so it could lead to bad predictions regarding students without employment. In this case, researchers are confident in that managing better the empty values and applying other kind of strategies in data cleaning and wrangling will

allow to get better predictive models and outcomes. The main results achieved have been quite promising and encourage authors to continue the labor of improving the generation of predictive models for employability and employment. The nature of this kind of problems is extremely complex and varies on the time, but with this kind of algorithmic and automated processes could address it better than the traditional approaches. Based on the results, the authors are committed to continue developing the approach to get better results and improve the process until it could be applied successfully in further research works. Topic Modelling is gaining growing attention in various text mining fields. The LDA is becoming a common tool. As a result, it has been applied to variety of ways and a range of extensions.

4 Dataset

Data for this was not readily available but we had developed a web scraper which can Scrape the Indeed website with the help of its API and will provide us Jobs based on given parameters. We had started collecting data for few fields of interest like NLP, Computer vision, Machine Learning and Software Engineering, Robotics and we have collected around 4000 jobs which were posted for a period of 2 years.



Year\Topic	NLP	ML	Computer Vision	Robotics	Total
2010	4,709	592,242	85,435	49,897	**732,283**
2011	5,307	676,729	107,570	63,415	**853,021**
2012	6,161	686,770	100,230	65,564	**858,725**
2013	8,664	748,257	127,592	79,098	**963,611**
2014	8,998	786,282	127,890	77,960	**1,001,130**
2015	8,418	793,407	117969	81,594	**1,001,388**
2016	9,167	824,052	124,821	85,408	**1,043,448**
2017	11,765	860,178	145,425	101,938	**1,119,306**
2018	17,072	913,105	171,668	114,197	**1,216,042**
2019	19,756	958,714	188,032	129,827	**1,296,329**
2020	23,792	1,076,459	225,009	144,779	**1,470,039**
2021	24,597	968,895	188,703	132,186	**1,314,381**
Total	**148,406**	**9,885,090**	**1,710,344**	**1,125,863**	**12,869,703**

Figure 1: Statistics : Count of Research Papers from Dimensions

We preprocessed the data for our analysis and did stemming and tokenization for making it workable data and remove unnecessary columns. We didn't wanted all the columns for our analysis so we dropped some of the columns and took only necessary columns like abstract text etc into a new dataframe and did analysis on it. We have also removed stop words and made it all lower case for analysis. The Scholarly Research articles was collected from dimensions API for the same number of years as the jobs are being scraped for and was used to find the correlation and then recent papers was downloaded from the same source and was used for predicting the future jobs based on it. We have collected 50k papers for the mentioned topics for the year 2016 and another 50k papers for the year 2020-2021 combined in the field of NLP, ML, Computer Vision and Robotics. We have also gathered data which has the statistics about the count of research topics published from the year 2010 to 2021. The dataset file is 1 GB in size.

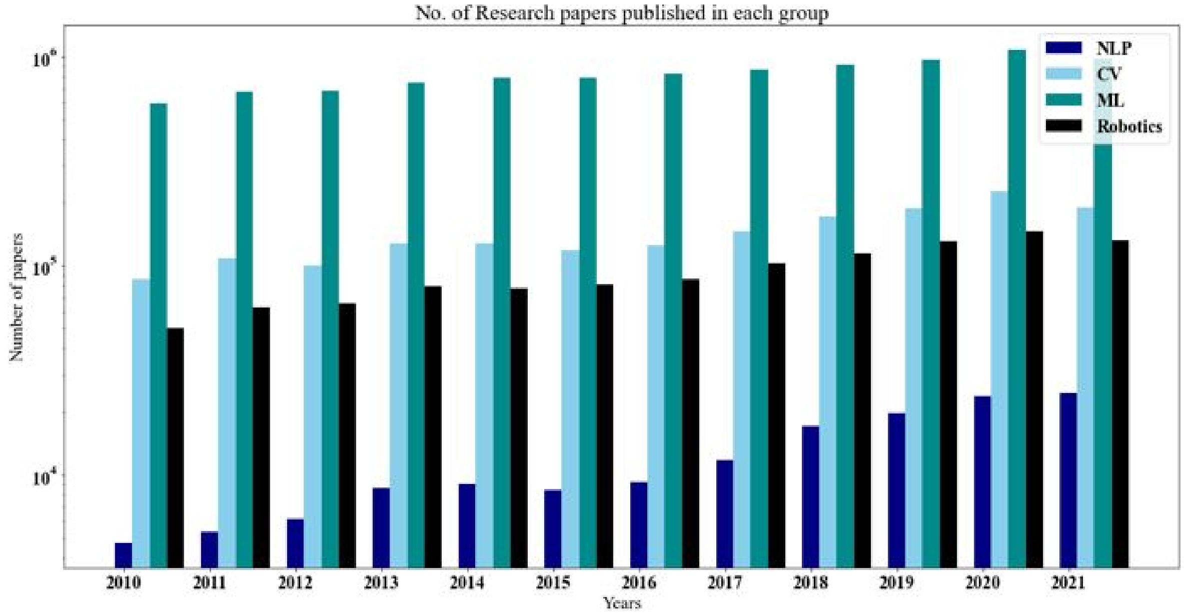


Figure 2: Research Papers published Each year

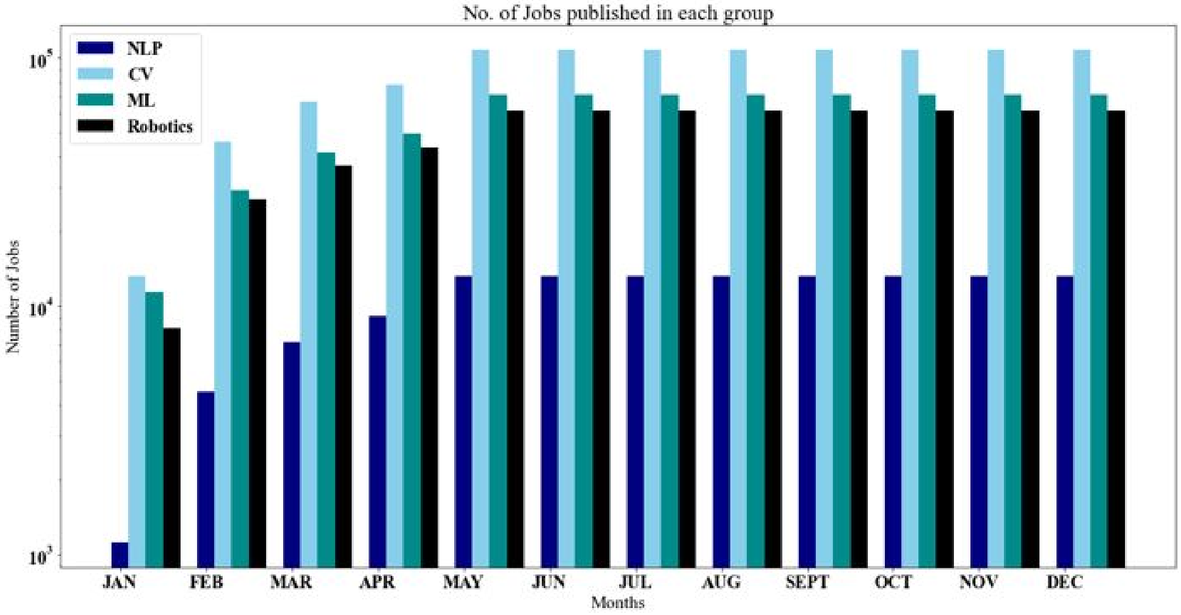


Figure 3: Jobs in Last Year

From this visualizations we understand that NLP was very less compared to other topics in the beginning but ML was increasing. So we would like to link these papers with external topics like jobs to make predictions. When we see the counts of jobs most of them were posted in 2021 and NLP was increasing. Also one point to note is Research papers published from the last 10 years but jobs are coming in now for that specific domains. So its taking some time for the jobs to be available in the market. Abdul Rahman has helped us in getting some of the data from Dimensions. So with the help of this dataset we will find if there is a correlation between publication of scholarly articles and the jobs being posted by the companies? Example say like now quantum computing is booming and lot of research papers are being published now so we want to predict will there be any quantum computing jobs in the future by doing correlation on the research articles published and the jobs posted in that field.

---

5 Methods

We have gone through various scholarly resources and there is no direct connection with the works that had been done before with our work, there are many publications that refer to employment based on various factors like [18] predicts the future job market based on Job Ads, whereas [19] predicts job insecurity from background variables and [20] tries to find the impact on future job market based on E-learning system. Since we are considering the future Job market based on research articles currently being published, our work becomes different from others.

Using novel methods and techniques like LDA could open new possibilities and ways to work in assessing employment based on research paper publications. Also, these methods could unleash new ways for manage huge amounts of information as a whole but considering each factor in a weighted way within the predictive models to be built. Previously some projects use basic statistics due the difficulty of handling big datasets in a non-automated mode, but in our way, the same procedure can be used to crunch all data related jointly. So our approach is Once we have collected the required Data (Research papers [ML,NLP,Computer Vision and Robotics] and Jobs), we will preprocess it with Pandas and remove any irrelevant, duplicate, or unhelpful data. Then we intend to do topic modelling (LDA [17]) on the Job description and find if the current job is related to the field for which it was fetched, as Jobs were collected from keyword search, some jobs of other fields might match based on keyword match. Then we plan to do topic analysis (LDA [17]) on research papers collected to filter the unwanted results and get the topics of the papers. We the preprocessing is finished we will compare the topics generated from each dataset to find the co-relation between them. The difference between both the LDA models are analyzed though heatmaps with topics of jobs and papers on each side, The colors show how closely each topic is related to other item's topics. Hovering on each section shows what terms make the correlation good or bad for that section. We have also implemented a interactive LDA analyzer that can help in understanding the topics generated and find if the model is generating good topics or not.Visualizations are important tools that help us understand the co-relation better than numbers, so we have relied heavily on them.Besides Doing LDA, we have also created vectors of both the papers and jobs and then found the similarity between them to be able to recommend closest jobs based on given paper. Dendogram was used to analyze the similarity.

---

6 Results and Evaluation

We have observed that the Jobs for the topics Natural Language Processing, Computer Vision, Machine Learning and Robotics have gained a boost in the year 2021 itself, before there weren't many jobs related to these fields on Indeed. The Scholarly articles are increasing year by year. While working on the subset of the vast data available, we found that the Scholarly Articles for various topics has shown different results on correlation using LDA with jobs, Using the Heatmaps, we were able to analyze the level of correlation, the dark red shows high correlation, whereas the dark blue shows least correlation. Based on these, the NLP model has shown slight correlation with jobs in 2021 with 2016 articles or 2020 and 2021 articles, Whereas the Machine Learning, Computer Vision and Robotics has shown better correlation with 2016 data rather than the newer 2020 and 2021 data with 2021 jobs.

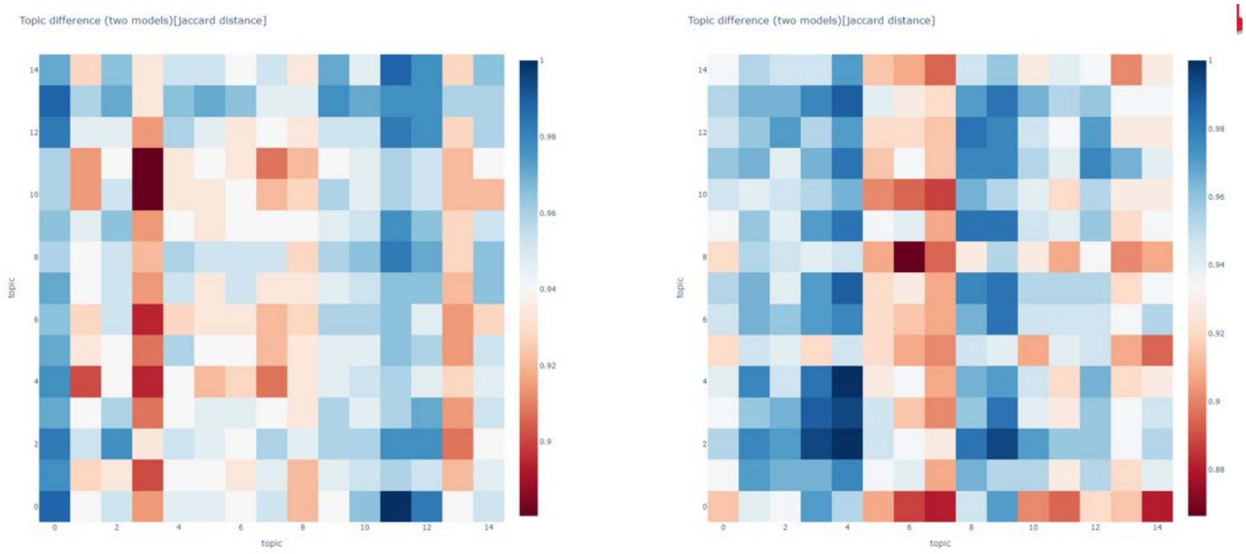


Figure 4: LDA Comparison - NLP
Left side 2016 articles with 2021 jobs Right side 2020 and 2021 articles with 2021 jobs

The LDA comparison visualizations for each topics does not show huge correlation but it goes to some extent to hint that correlation exists, but the correlation will not be same for all topics, for example some fields may have correlation with few years older papers as those methods or topics mentioned in paper might take time to be available for implementation,

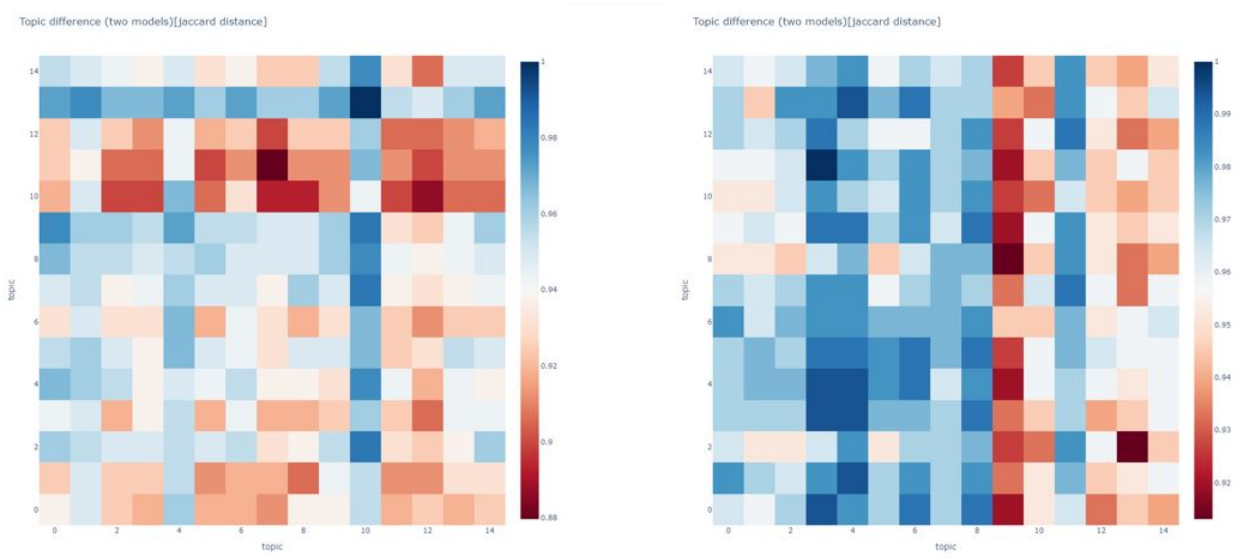


Figure 5: LDA Comparison – Computer Vision
Left side 2016 articles with 2021 jobs Right side 2020 and 2021 articles with 2021 jobs

whereas few fields might be Industry ready as the article is submitted. The Dendrograms also shows good clusters with cosine similarity on the data. With this hint of correlation we can further work with collecting more data.

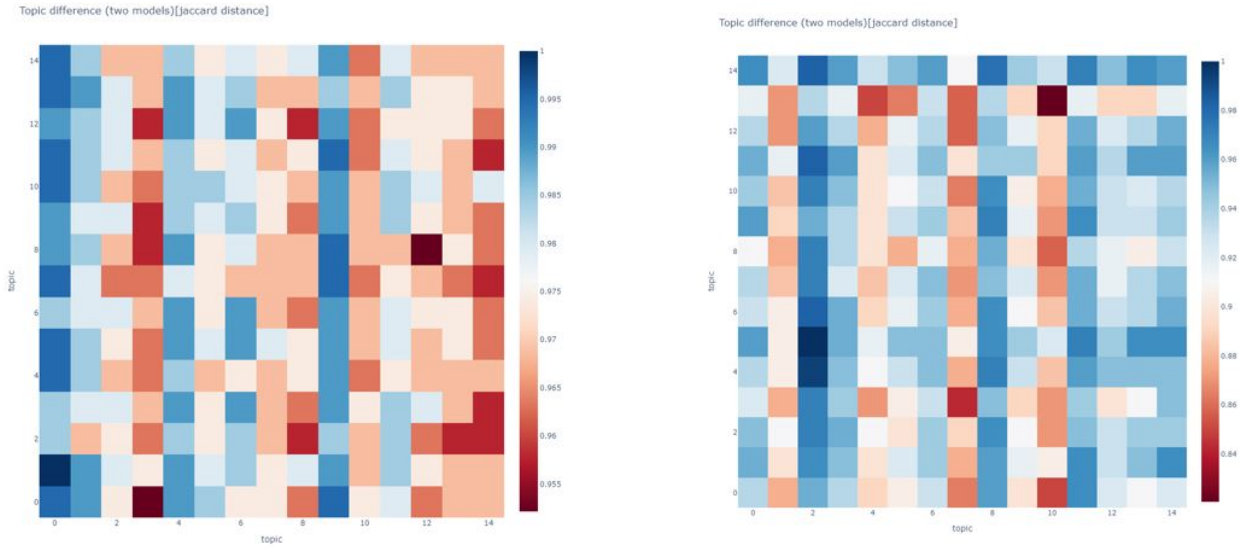


Figure 6: LDA Comparison – Machine Learning
Left side 2016 articles with 2021 jobs Right side 2020 and 2021 articles with 2021 jobs

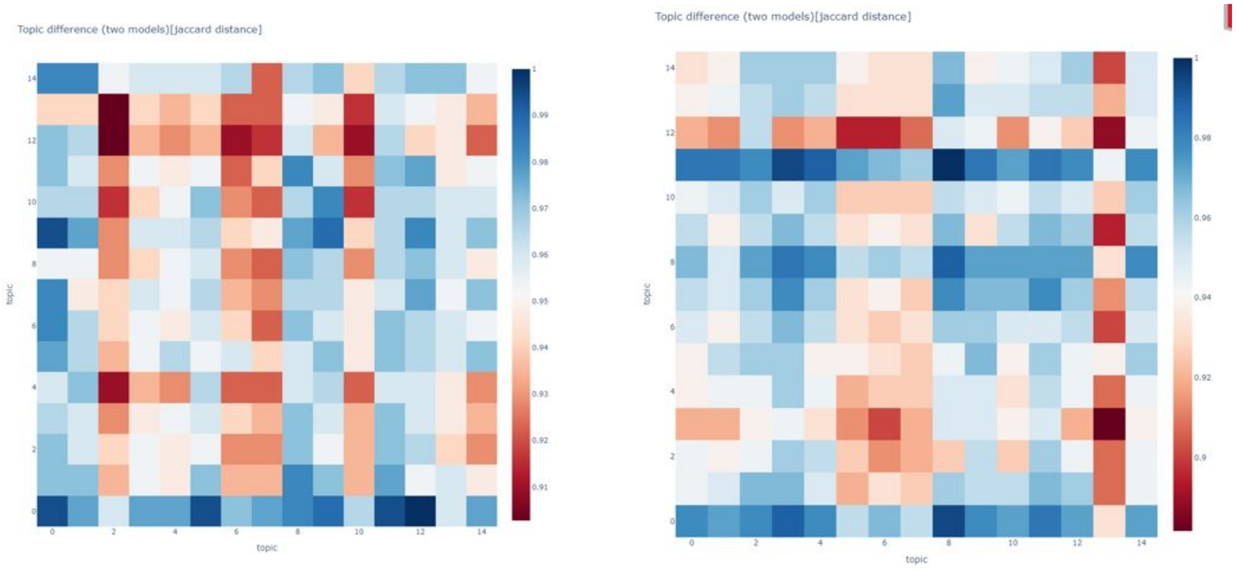


Figure 7: LDA Comparison - Robotics
Left side 2016 articles with 2021 jobs Right side 2020 and 2021 articles with 2021 jobs

7 Conclusion

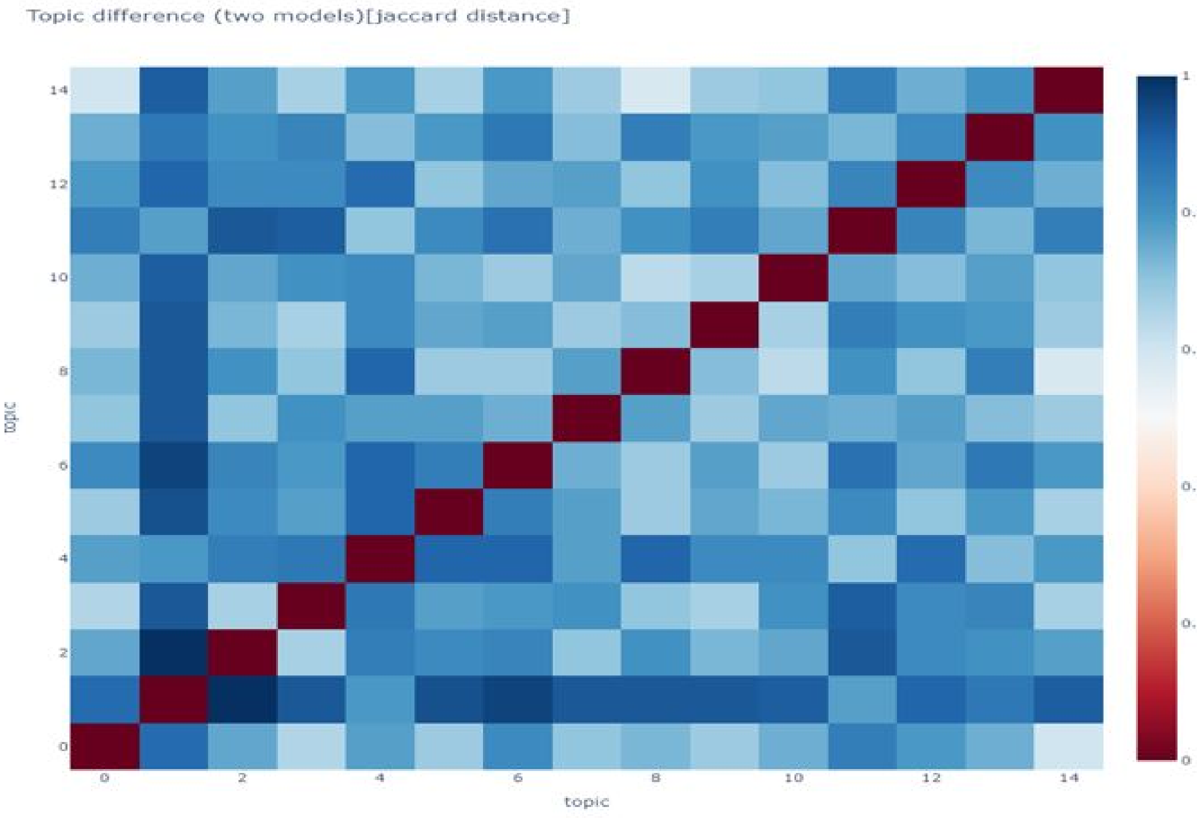


Figure 8: Standard Topic Difference

Job Market is not very predictable, it carries few patterns but might be affected by various reasons, Predicting this volatile market will help employer ans employees better guage the future. Previous works were done on how jobs were affected by various things like Covid, etc. Also few works included prediction of getting a job based on Students circular activity and Coursework or based on field of study. Our work focuses on finding if the job market has any correlation with research articles already published or being published, Would we be able to predict the type of and number of jobs that can be available in future based on scholarly articles. Using LDA technique we were able to find a good amount of correlation between them, Since we are getting this correlation on subset of dataset available, we can expect more finding if all the data is used for analysis. We get correlation for some topics for current research papers while for some it correlates better with older papers. Finally we would like to conclude by stating that mining more data will open doors for more findings.

8 Future Work

The results showing some amount of correlation with this dataset, provides promising future in predicting the jobs based on research papers, This can be further extended to understand the correlation better with mining more data, i.e, All jobs that are listed in Indeed with full descriptions and All papers for a particular year with full text, using this, we hope better understanding of correlation can be achieved. Besides just finding the correlation, we can find clear reason why some topics are affected with newer research papers and why some are affected with older papers. Also, Instead of Just recommending few Job titles we might be able to predict the number of jobs, if the data is properly collected. We can agree that the data is most integral part, so to collect data,instead of Just using indeed, we can also collect from other JOB boards and combine them to filter only unique data, this will hep us overcome the API limits and collect more data. If All the papers of a specific topic for a particular year are mined, the Correlation between the number of papers and number of jobs posted could be analyzed with a plot, we used 50K dataset so were not able to achieve this.

9 Acknowledgments

We would like to thank Dr. Hamed Alhoori and Abdul Rahman Shaik (PhD Student in Computer Science). Without their assistance and dedicated involvement in every step throughout the process, this project would have never been accomplished.

10 References

[1] By BRIAN BEERS, https://www.investopedia.com/terms/j/job-market.asp

[2] Mezhoudi, N., Alghamdi, R.,Aljunaid, R. et al. Employability prediction: a survey of current approaches, research challenges and applications. J Ambient Intell Human Computing (2021).

[3] PREDICTING EMPLOYMENT THROUGH MACHINE LEARNING By Linsey S. Hugo (May-2019)

[4] AI, Robotics, and the Future of Jobs BY AARON SMITH AND JANNA ANDERSON

[5] Arthur R (2021) Studying the UK job market during the COVID-19 crisis with online job ads. PLoS ONE 16(5): e0251431. https://doi.org/10.1371/journal.pone.0251431

[6] Termini, C.M., Traver, D. Impact of COVID-19 on early career scientists: an optimistic guide for the future. BMC Biol 18, 95 (2020). https://doi.org/10.1186/s12915-020-00821-4

[7] Deep learning-based prediction of future growth potential of technologies by June Young Lee,Sejung Ahn, Dohyun Kim

[8] Moro E, Frank MR, Pentland A,Rutherford A, Cebrian M, Rahwan I. Universal resilience patterns in labor markets. Nature Communications. 2021;12(1):1–8. pmid:33785734

[9] Berger MC. Predicted Future Earnings and Choice of College Major. ILR Review. 1988;41(3):418-429.doi:10.1177/001979398804100306

[10] G. Jason Jolley, Predicting North Carolina's Job Market in 2020

[11] Sandefer, Ryan; Marc, David; Mancilla, Desla; Hamada, Debra. ”Survey Predicts Future HIM Workforce Shifts: HIM Industry Estimates the Job Roles, Skills Needed in the Near Future” Journal of AHIMA 86, no.7 (July 2015): 32-35.

[12] N. Dawson, M. -A. Rizoiu,B. Johnston and M. -A. Williams, ”Predicting Skill Shortages in Labor Markets: A Machine Learning Approach,” 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 3052-3061, doi: 10.1109/BigData50022.2020.937777 3.

[13] Bhagavan KS, Thangakumar J, Subramanian DV (2020) Predictive analysis of student academic performance and employability chances using HLVQ algorithm. J Ambient Intell-HumanizComputing.

[14] https://www.indeed.com/

[15] https://www.linkedin.com/jobs/

[16] https://www.dimensions.ai/dimensions-apis/

[17] Changzhou Li, Yao Lu,Junfeng Wu, Yongrui Zhang,Zhongzhou Xia, Tianchen Wang,

Dantian Yu, Xurui Chen, Peidong Liu, and Junyu Guo. 2018. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering. In Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1699–1706. DOI:https://doi.org/10.1145/3184558.3191629

[18] C. M. Jaramillo, P. Squires,H. G. Kaufman, A. Mendes da Silva and J. Togelius, "Word embedding for job market spatial representation: tracking changes and predicting skills demand," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 5713- 5715, doi: 10.1109/BigData50022.2020.9377850.

[19] Näswall, K., De Witte, H.(2003). Who Feels Insecure in Europe? Predicting Job Insecurity from Background Variables. Economic and Industrial Democracy,24(2), 189–215. https://doi.org/10.1177/0143831X03024002003

[20] "The Impact of E-Learning on the Future Job Market – Predicting a New Educational Type of Learning Style for The Next Generations." The Impact of ELearning on the Future Job Market – Predicting a New Educational Type of Learning Style for The Next Generations. Vol. 1. N.p., 2019. Print

[21] Proposing a Machine Learning Approach to Analyze and Predict Employment and its Factors