De La Salle University
AKI
Angelo King Institute
for Economic and Business Studies

DLSU-AKI Working Paper Series
2024-10-095

Comparing Goldin and the Philippines: Analyzing the Research Efforts in Women and Labor Economics Using Basic Text Mining Techniques

By:

Arianna Liza Ortillan
Carlos L. Tiu School of Economics
De La Salle University

This paper was awarded the third prize in the 2023-2024 essay-writing competition organized by the Angelo King Institute among the students of the Carlos L. Tiu School of Economics, on "The Economics of Professor Claudia Goldin."

Comparing Goldin and the Philippines: Analyzing the Research Efforts in Women and Labor Economics Using Basic Text Mining Techniques

By Arianna Liza Ortillan

Claudia Goldin has worked extensively in many fields, but was acknowledged last year via the Nobel Prize in Economics for her work and findings in labor economics and economic history. More specifically, she combined historical analysis and empirical work to create an in-depth forensic understanding of women's participation in the labor market. This includes, but not limited to, changes in how women perceive work, the gender wage gap, the contraceptive pill's role in women's labor force participation, and an extensive understanding of the history of women in the labor market. Most of her work is limited to the context of the United States, with a detour to other countries in some works (Goldin, 1994). Nevertheless, many would agree that she had created the foundational work on how we could potentially look into gender and labor economics.

Although Goldin herself had created a legacy in this field of study, this does not mean that people from other countries have not as well. The idea of gender and labor economics has been a well-studied field with many interests coming from the government, policymakers, and the academe. Even an individual that would garner enough curiosity to stop and observe how their household, their school, their workplace, and other surroundings operate may realize that there may be differences between men and women in general. As work and gender are essential parts of our daily lives, it is natural that we desire to understand our experiences relating to them.

With that in mind, there are several perspectives that researchers would look into concerning this topic. In that regard, Goldin's works are extensive in many senses. She was able to look into several variables such as the contraceptive pill and the labor force participation of women, how an individual's time in school would factor in their building of career and family, and how women's

perception of joining the workforce have changed over different generations (Goldin & Crane, 2021; Goldin, 2006). These three examples are only the tip of the iceberg. She has also looked into historical accounts as the basis of her research. Examples of these records would include legal documents concerning monumental trials that concerned the civic rights of women to text analysis of newspapers and books in determining how people's perception of sex discrimination have evolved over decades (Goldin, 2023). It is undeniable that she took into account and studied any perspective she could look into to give a comprehensive understanding of the historics and intricacies of women in the labor force. One may call her rigorous. Others may call her creative.

Other researchers could take inspiration in her work, specifically in the methods and way of thinking she undertook to come to her conclusions. As established earlier, she was not limited to just empirical or historical work; she used a combination of the two and utilized creative ways of thinking into how she approaches a specific question. Goldin had to intensively dig old files and archived documents to gather her data. In works of historical analysis, feats of hard work are necessary to get hands on the information needed to complete the study.

Nevertheless, this does not constitute a lack of motivation to pursue this field of research. Using her work and ethic as a framework, we could employ her methods in understanding the labor market in the Philippines. Just as Goldin had done, we need creativity and rigorous work to answer our questions despite the relatively limited data available.

Goldin inspired me to take a different approach in methodology and analysis. Combining the idea of using her work as a basis and using a rigorous methodology, I looked into the collective work of researchers and authors under the scope of the women in the Philippine labor market and compared that to Goldin's works. More specifically, this paper answers the question of what are the topics that researchers have focused on studying in relation to the aforementioned field and if there are

differences in research interests between them and Goldin. Recommendations would be based on the analysis of the findings and methodology used.

In that regard, this paper aimed to achieve the following objectives:

- Determine the most common classifications in research efforts and interests regarding women in the labor market in the Philippines
- Compare Goldin's works and the research effort under women in the labor market in the Philippines
- Create recommendations on future research paths according to Goldin's work

Methodology

I used a basic text mining technique, specifically word frequency and text classification, as the foundation of data collection and analysis (Silge & Robinson, 2017; Feldman et al., 1998). The first 1000 titles in Google Scholar under the search term *"women labor Philippines*" were scraped using Apify actors as the source of the working dataset. These titles are sorted by relevance, according to Google Scholar's standards. After cleaning duplicate titles and irrelevant search matches, there are now a total of 787 observations. To start the text mining process, the observations underwent tokenization, corpus analysis, and common English stop words removal to work with cleaner text data. From this, I text mined the most frequent words from all the titles, and erased terms that are too broad in nature (i.e. Philippines, women, labor, and gender). Some terms are treated with collocation, where two or more words are often found beside each other (i.e. labor market, labor force, domestic work). After recovering the word frequency, titles were assigned to categories based on these frequent terms.

Data from scraping was able to recover the following relevant variables: title, author, year, and search match. Title and search match matters here the most, as this is what Google Scholar relates to

the search term. Using its own algorithm, it searches both the title and readable content to match the search term. Due to the limitations of Google Scholar and Apify, the abstract was not recovered.

Google Scholar and Source of Data

A discussion regarding how Google Scholar operates is necessary for transparency. Google Scholar (2024) describes its own system as "*to rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature*." This means that it has its own system that ranks titles based on its own standards based on the content, the publisher, the author/s, and citations. Nevertheless, it does not publicly disclose how exactly these variables are weighted to show the ranking by relevance of each title (Halevi et al., 2017).

However, Beel and Gipp (2009) tried to reverse engineer Google Scholar's search engine to determine exactly how it ranks the results. Their findings show that highly cited works are more likely to appear at the top results. In that sense, a mainstreamed article with more citations would show up more likely than an article that argues with opposing views. They have also cautioned about the Matthew Effect, where articles at the top are more likely to get more views and thus further citations. Although, there are frequent cases where it will also show a title with not many citations. They argued that this would depend on the words used to generate results.

Google Scholar also does not have an "advanced search" option, limiting the user solely based on the keywords they input in its search bar. In this case, the search term "*women labor Philippines*" is used to limit the results specifically on those three words. Adding another term such as "*economics*" or "*participation*" would further limit the range of the search results. Titles and works that are not subject to open access, such as a university's repository, would also not be under Google Scholar's reach.

Hernandez and Hidalgo (2020) used a similar, but more advanced approach, in finding seminal papers and its citations, utilizing data mining. They used and compared data from Google Scholar and

Scopus, concluding that the two sources of sample size would yield vastly different results. In that sense, using a different database would most likely give different results from what I was able to gather.

With that said, Google Scholar is the academic database that has the most third party support in terms of scraping web results. Other websites, such as JSTOR, Econlit, or Scopus, would require tools not as easily accessible. For the convenience it has given, I used Google Scholar as the source of my working data for this instance.

Topic Classification of Terms

The categories used in this paper are based on the recovered word frequencies. As such, words were put under umbrella terms. The table below indicates all the words and terms used under a specific category.[1]

Table 1. Classification of Titles Based on Recovered Word Frequency

Category	Words Associated with Category
Migration	migration, migrants, overseas, overseas filipino workers
Geographic Location	Cebu, Quezon, Mindoro, Davao, Bataan, Ilocos, Bukidnon, Negros, urban, rural, upland, coastal, province
Labor Market	labor market, labor force participation, labor force, employment, unemployment, labor supply
Household	household, family, child, mother, wife
Agriculture	rice, farming, plantation, sugarcane, fisheries, agrarian, agroforestry
Social	social landscape, social dimensions, social mobility, social construction, rights, culture
Education	education, educational, schooling
Earnings	earnings, income, wage, salary
Politics	political, political

[1] Technicalities of grammar and spelling are not included to simplify the table. Words here were put in its plural and singular forms and upper and lower case to account for possible permutations.

Health	health, nutrition, maternal
Fertility	fertility, fertile, post-menopausal, menstrual, pregnancy
Policy	policy
Industry	industry
Domestic Work	domestic work
Others	anything not covered by these words were put into these category

I would like to note that many titles under *Others* have used general terms that are not in any of these categories. Such terms would be "*women*" and "*economic*". There are also many instances where a word would be unique in one or two titles. In this case, having access to the abstract would have been more beneficial in this type of procedure.

Results and Discussion

Figure 1 shows the 20 most frequent words among the titles. This is after removing the words that are too broad (i.e. *Philippines* and *gender*) or terms that are technical (i.e *evidence* and *study*)[2]. After that, the most frequent term used among the 787 observations is *migration* with 94 repetitions. This is followed by the *labor market*, which was repeated 46 times. Right after that is *rural* with 41 repetitions. The fourth most frequently seen word is related to the first, *migrant*, with 33 repetitions. *Rice* is also repeated the same number of times. Other most frequently used words in descending order are *education, labor force, family, farming, employment, care, household, industry, domestic work, health, households, political, wage, community, and fertility*.

[2] Unedited word frequency table can be found in the Appendix section, showing the 100 most frequent words

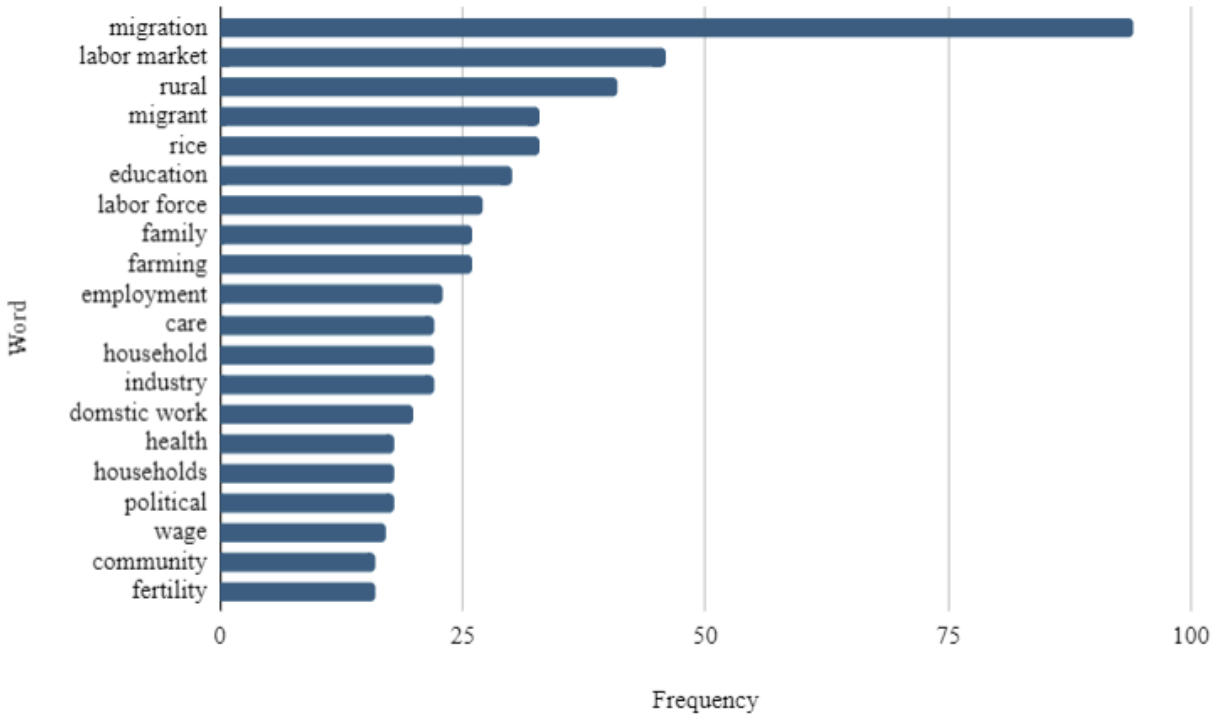

Figure 1. 20 Most Frequently Used Words among the TItles

By analyzing these, we can infer that there is a good amount of research within the context of women in the Philippines' labor market in the context of migration. There is also a significant amount of research allocated into understanding the structures of the labor market and labor force. There are also titles under the umbrella of agriculture, as seen from the words of *rice* and *farming*. Other terms vary from each other and were used as the basis of simple classification.

Figure 2 illustrates the titles into categories. In descending order, the categories are *Others, Migration, Geographic Location, Labor Market, Household, Agriculture, Social, Education, Earnings, Politics, Health, Fertility, Industry, and Domestic Work*. Overlapping categories are necessary due to the multifaceted nature of economics studies. For example, a study dealing with female migrants' income falls under both *Migration* and *Earnings*. Another example would be a study dealing with the role of wives in a farming area that fits into both *Households* and *Agriculture*.

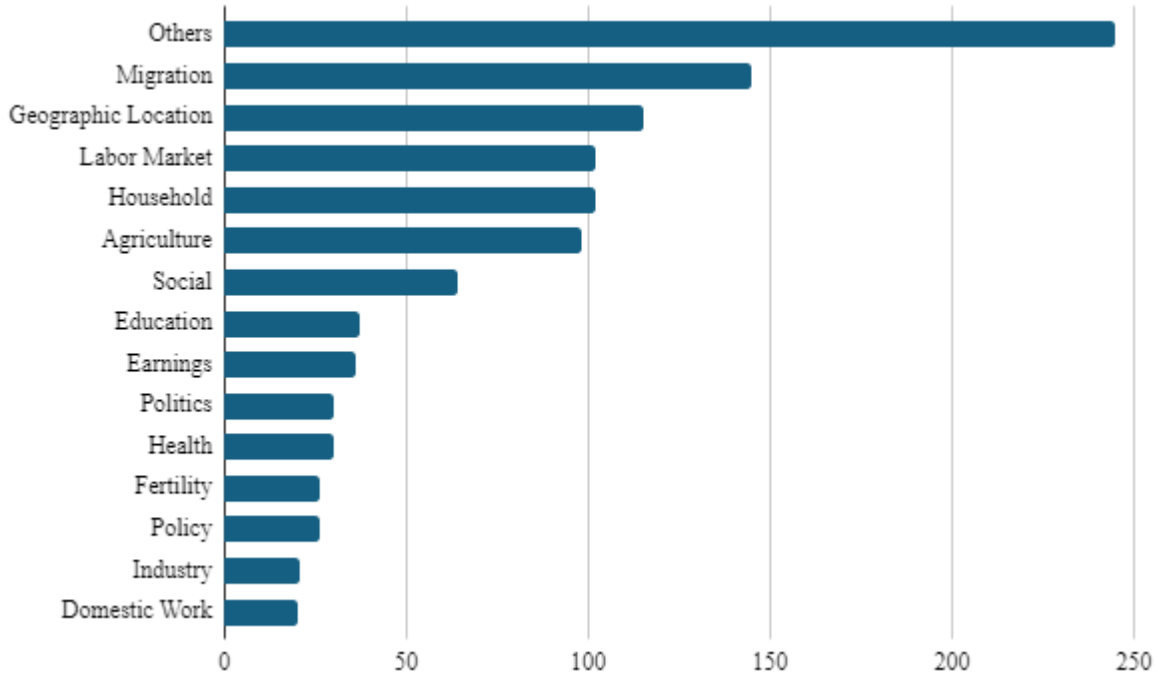

Figure 2. Putting Titles into Categories (with overlapping)

From the scraped database, *Migration* is the category with the most titles This would mean that a significant portion of research effort has been put into the migration of women. The number of studies under *Geographic Location* also indicates that there are several studies that limit itselves to specific areas, provinces, or cities. Factors such as social context or data availability may differ from location to location, thus necessitating a need for location-based research. This is followed by the *Labor Market*, where the structures of the labor market such as the labor force participation, employment, and unemployment are tackled. *Household* would refer to titles that deal with the structure of the household, such as the wife or the children, or household as a unit of observation. *Agriculture* tackles the area of farming, crops, fisheries, and forestry.

Social, in this context, are outside forces that shape a person's daily experiences and that directly and indirectly affect behavior (Burke et al., 2009). More specifically, these are the social contexts that could potentially affect a woman's day-to-day life. This is followed by titles involving *Education*, *Earnings*, and *Politics*. There are also several studies that tackled *Health* and *Fertility*. Despite both

categories being about the body, *Fertility* refers specifically to reproductive health. *Policy* then refers to titles that concern itself with plans of what to do in specific situations that have been agreed to officially by a group of people, in this case the government (*Policy*). *Industry* covers the participation of a specific type of individuals in different industries. Lastly, *Domestic Work* refers to any type of labor that concerns maintaining the household. This is relatively different from the category of *Household* as that refers to the structure of a household, rather than the labor put into maintaining one.

As mentioned in an earlier section, the *Others* category is for anything that the top most word frequency procedure was not able to capture. A different paper under Beel and Gipp (2009) discusses that researchers finding "gems" in Google Scholar that are not subject to citation weight is also possible. It is likely that many of the scraped titles, many belonging to the *Others*, may be a gem, or papers that are not considered mainstream considering the words they may have used. To create a more effective classification, getting the abstract or summaries of these titles would be the ideal option rather than simply relying on titles.

Comparing to Goldin's Work

Goldin's work was awarded the Nobel Prize under the notion of "providing the first comprehensive account of women's earnings and labor market participation through the centuries." She was also credited for her work regarding the causes and changes of the remaining wage gap. Despite that, her works were not limited to these alone, and there are much more insights we can recover from the rest of her writings (The Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2023, 2023).

Aside from her work regarding women's labor force participation, their earnings, and the wage gap, she had also written about the diffusion of contraceptive pills and its effect on career and marriage alongside the civil formation of policies of women's rights. Outside of the scope of gender, she had also made significant work on human capital and schooling (Goldin & Katz, 2007).

Using the statistics recovered from word frequency and simple classification, the research effort regarding women in the labor force in the Philippine's context shows a slightly different direction from Goldin's. Although it is evident that there are a lot of studies about earnings, households, and the labor market, the Philippines have its unique set of traits that differentiate it. For instance, migration and agriculture are a matter of importance in the Philippines' economy. Goldin rarely focused on these topics, perhaps due to a lack of interest or the different set of circumstances found within the United States. The same case can be made for domestic work, as she had only passed through it in her writings in some of her historical analysis (Goldin, 1994).

Although there is a considerable difference in the direction of research effort between Goldin and the collective work in the Philippines due to the differences in context, there are still several similarities that can be observed. Studies on education, earnings, fertility, and the household are important variables in understanding labor economics and its associated market. Perhaps, the key difference between these similarities is the extent and scope of work. Goldin was acknowledged for her work that utilized centuries worth of data and was able to construct a historical analysis backed up by empirical evidence. Many of the titles scraped had only worked with data given a relatively smaller time frame.

Conclusions and Recommendations

Methodology

I would recommend using a better source of academic articles and books that can provide not only the titles, but also the abstract and summaries. This would enable a less biased dataset to text mine while also enlarging the scope of analysis to the content of the papers themselves. I would also recommend analyzing the trends of the word frequencies within a time period. This may give a more in depth overview of the trends in research regardless of what field one may want to look into.

Research Paths

It would be hard to recommend what Goldin did, that is using extensive data to analyze women's labor participation over a long period of time. If this was the case, someone else would have done this type of research already. But she was able to develop the foundational work that would enable future researchers to think of how to approach the labor market in different perspectives. As such, in topics that deal with microdata that requires rigorous and forensic work, we have to think creatively in how we approach our questions and methodology.

References

Beel, Joeran, & Gipp, B. (2009, July). Google scholar's ranking algorithm: An introductory overview. Academia.edu. https://www.academia.edu/24683245/Google_Scholars_Ranking_Algorithm_An_Introductory_Overview

Beel, Joran, & Gipp, B. (2009). Google scholar's ranking algorithm: The Impact of Citation Counts (an empirical study). 2009 Third International Conference on Research Challenges in Information Science. https://doi.org/10.1109/rcis.2009.5089308

Burke, N. J., Joseph, G., Pasick, R. J., & Barker, J. C. (2009). Theorizing social context: Rethinking behavioral theory. Health Education & Behavior, 36(5_suppl). https://doi.org/10.1177/1090198109335338

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, J., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. Practical Aspects of Knowledge Management.

Goldin, C. (1994). The U-Shaped Female Labor Force Function in Economic Development and Economic History. https://doi.org/10.3386/w4707

Goldin, C. (2006). The quiet revolution that transformed women's employment, education, and family. American Economic Review, 96(2), 1–21. https://doi.org/10.1257/000282806777212350

Goldin, C. (2023). Why Women Won. https://doi.org/10.3386/w31762

Goldin, C. D., & Crane, N. (2021). Career and family: Women's century-long journey toward equity. Princeton University Press.

Goldin, C. D., & Katz, L. F. (2007). The race between education and technology the evolution of U.S. Educational Wage Differentials, 1890 to 2005. National Bureau of Economic Research.

Goldin, C., & Katz, L. (2000). The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions. https://doi.org/10.3386/w7527

Google Scholar. (2024). About google scholar. Google. https://scholar.google.com/intl/en/scholar/about.html#:~:text=Google%20Scholar%20provides%20a%20simple,universities%20and%20other%20web%20sites.

Halevi, G., Moed, H., & Bar-Ilan, J. (2017). Suitability of google scholar as a source of scientific information and as a source of data for scientific evaluation—review of the literature. Journal of Informetrics, 11(3), 823–834. https://doi.org/10.1016/j.joi.2017.06.005

Hernández, A. B., & Hidalgo, D. B. (2020). Findings seminal papers using data mining techniques. Open Journal of Social Sciences, 08(09), 293–305. https://doi.org/10.4236/jss.2020.89023

Policy . Cambridge Dictionary. (n.d.). https://dictionary.cambridge.org/us/dictionary/english/policy

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media.

The Sveriges Riksbank Prize in Economic Sciences in memory of Alfred Nobel 2023. NobelPrize.org. (2023). https://www.nobelprize.org/prizes/economic-sciences/2023/press-release/

Appendix A

word	n	word	n	word	n
philippines	515	impact	19	world	12
labor	239	transnational	19	agricultural	11
women	222	economy	18	agriculture	11
philippine	139	health	18	feminization	11
gender	117	households	18	fishing	11
women's	95	political	18	income	11
migration	94	experience	17	management	11
workers	60	wage	17	sector	11
market	45	community	16	status	11
participation	44	effects	16	trafficking	11
rural	41	fertility	16	violence	11
evidence	37	filipina	16	approach	10
study	36	gendered	16	children	10
female	33	export	15	city	10
migrant	33	migrants	15	comparative	10
rice	33	rights	15	crisis	10
economic	31	based	14	feminist	10
global	31	cebu	14	growth	10
social	31	communities	14	movement	10
education	30	food	14	national	10
force	30	issues	14	perspective	10
development	29	politics	14	research	10
international	29	markets	13	time	10
family	26	policy	13	villages	10
farming	26	province	13	asia	9
filipino	26	sex	13	asian	9
domestic	25	supply	13	change	9
employment	23	trade	13	culture	9
production	23	child	12	determinants	9
care	22	empowerment	12	politics	14
household	22	implications	12	markets	13
industry	22	labour	12	policy	13
analysis	21	outcomes	12	province	13
role	20	policies	12	sex	13
globalization	19	roles	12	supply	13

word	n	word	n		
politics	14	villages	10		
markets	13	asia	9		
policy	13	asian	9		
province	13	change	9		
sex	13	culture	9		
supply	13	determinants	9		
trade	13	families	9		
child	12				
empowerment	12				
implications	12				
labour	12				
outcomes	12				
policies	12				
roles	12				
world	12				
agricultural	11				
agriculture	11				
feminization	11				
fishing	11				
income	11				
management	11				
sector	11				
status	11				
trafficking	11				
violence	11				
approach	10				
children	10				
city	10				
comparative	10				
crisis	10				
feminist	10				
growth	10				
movement	10				
national	10				
perspective	10				
research	10				
time	10				