

Using web data to study the informal childcare sector in Canada

Jose Morales Vidales* Monica Alexander†

Demand for non-parental childcare has increased in recent decades, including informal home-based childcare offered by nannies. However, traditional data sources lack detailed information required to understand the experiences of nannies seeking work. In this project we utilize a popular website in Canada to study patterns and trends in the informal childcare sector in Canada. We scraped almost 10,000 unique online user profiles of individuals seeking work, and utilized keyword- and embedding-based approaches to classify their demographic characteristics based on free-text data. Initial results show differences in experience and advertised salary based on migration status and country of origin.

1 Introduction and motivation

Demand for non-parental childcare has increased substantially in recent decades, largely driven by an increase in labour force participation of women. In Canada, the use of non-parental childcare covered more than half of families in 2022 (Charters and Findlay 2023). As well as centre-based and licensed or regulated childcare providers, an important part of the childcare landscape in Canada is unlicensed home-based childcare, provided by baby-sitters and nannies. Demand for this informal home-based childcare has also recently increased, potentially because of the relative flexibility of hours compared to childcare centres, and in response to health concerns surrounding the Covid-19 pandemic (Crook 2020). Compared to employees in regulated childcare options, nannies may experience less job security, more uncertainty, and poorer working conditions (Milkman 2023).

Despite the increasing importance of workers in the informal childcare sector in countries such as Canada, little is known about the characteristics, working conditions, and employment

*Statistical Sciences, University of Toronto

†Statistical Sciences and Sociology, University of Toronto

outcomes of this sub-population. Data collected in censuses and surveys is limited; informal childcare workers are often hard to distinguish between care-workers in general, and information about working conditions and place of work is also limited.

To address this gap, in this project we utilize a large, rich, dataset constructed from public web profiles on a nanny advertising website to better understand the characteristics and systematic inequalities in the informal childcare sector in Canada. We are particularly interested in studying differences in work expectations and outcomes by country of birth and migrant status. We extracted almost 10,000 unique online profiles over three months from the website canadiannanny.ca. These profiles contain a range of different types of information, including advertised pay rate, qualifications, experience, as well as a free-text biography. We then used a variety of text-based methods to extract information on key demographic characteristics, including country of origin, age, and citizenship status. In the remainder of this abstract, we describe the data extraction process, methods used to retrieve information on demographic characteristics, initial results and future work.

2 Data

We obtained information on characteristics of nannies seeking work in Canada from the website canadiannanny.ca. This website, which is used by over 200,000 potential nannies across Canada, contains incredibly rich data on nannies' characteristics and experiences on their profiles.

We used webscraping tools in R (primarily relying on the `rvest` package) to extract information on nannies from publicly available profiles. The profiles were scraped using an R script at multiple points of July, August and September 2023. So far more than 16,000 entries have been scraped from the site, of those around 9,700 are distinct users.

Demographic variables are usually reported by the user in the text-based sections of their profiles, which include the short blurb right under the profile picture, a "Reasons to Hire Me" section, and a longer description in the "About Me" section. We also retrieved the users' name, url, location, reported years of experience, hourly rate, last time active on the site, number of reviews, star rating out of five, bullet points under the "I can work:" subsection (part-time, full-time, summer, etc.), children ages the user has experience with (infant, toddler, newborn, etc.), number of children they can look after, experience with children with medical conditions (diabetes, disability, epilepsy, severe allergies, etc.), transportation requirements (close to transit, has driver's license, etc.), qualifications (first aid, CPR, languages, etc.), and services they can provide (housekeeping, cooking, groceries, swim supervision, etc.).

3 Retrieving demographic characteristics from profiles

We are interested in studying differences in user information by key demographic characteristics, particularly age, migrant status (in particular, if the user requires sponsorship), and country of origin. However, these characteristics are not reported in profiles in a fixed field, but rather the information is often contained somewhere in the free-text user-written biographies (see below for an example). As such, we use several information retrieval approaches.

Free-text biography example.

Hi there, my name is Erin (she/her), (23 years old)!

I am a recent (2022) graduate from the University of Waterloo with an Honours degree in Therapeutic Recreation. I spent four years nannying for two families (that I connected with on this website), while completing my degree.

I moved to Toronto in the Fall of 2022 and reside in the West End, Roncesvalles neighbourhood. I have nannied for two families since being in Toronto, one who I am currently working with, and can connect you with both past and present families for recommendations! ...

Information Retrieval (IR) is the process of obtaining any type of media based on user information needs. The resulting IR system is often called a search engine (Jurafsky and Martin, n.d.). The IR task that we consider returns a list of ordered sentences, taken from profile descriptions, based on a query. The IR architecture uses the vector representations of queries and sentences, which are then ordered based on a similarity function like cosine or dot product. We tested two main IR methods to extract demographic information of interest. Below is a brief description of both methods considered.

3.1 BM25Okapi

BM25 stands for “Best Matching 25” (i.e the 25th iteration of the function), and Okapi makes reference to the first IR system that used it. This is a bag-of-words model in which query and document vectors are based on unigram word counts, and each word is considered independently of its position. This function is similar to TF-IDF (term frequency - inverse document frequency) weighting, but it adds two parameters: k , which adjusts the balance between term frequency and IDF, and b , which controls the importance of document length normalization (Brown 2022). The BM25 score of a query Q containing keywords q_1, \dots, q_n and a document D is:

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \times \frac{tf(q_i, D) \cdot (k + 1)}{tf(q_1, D) + k \times \left(1 - b + \frac{b |D|}{|D|_{\text{avg}}}\right)}$$

Where $tf(q_i, D)$ is the number of times the keyword q_i appears on document D , $|D|$ is the length of the document and $|D|_{\text{avg}}$ is the average document length. The inverse document frequency $IDF(q_i)$ is often computed as:

$$IDF(q_i) = \log \left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1 \right)$$

Where N is the total number of documents and $n(q_i)$ is the number of documents containing the keyword q_i . To implement this method we are using the python package `rank_bm25` which sets $k = 1.5$ and $b = 0.75$ by default. It also sets keywords with negative IDF to the average of the non-negative IDFs multiply by 0.25.

A major disadvantage of the BM25 scoring function is that the query’s keywords have to appear in the documents that it’s searching through. So if the author used a synonym or expressed it in another way the model has no way of knowing that that is a relevant document.

3.2 Sentence Embeddings

Our second approach was to compute vector representations of sentences, or embeddings, using transformer models. We used the `SentenceTransformers` package in Python, which was developed for semantic search tasks similar to our goal here (Reimers and Gurevych 2019). We mainly used the `multi-qa-mpnet-base` bi-encoder model, which was specifically trained on 215 million question-answer pairs from sources such as Yahoo answers, StackExchange and Google and Bing search queries. Once the embeddings of both the query and sentences are computed we can use the cosine or dot product similarity score to rank them.

The main advantage of this approach is that we don’t need to have the query keywords appear on the sentences; the model should retrieve sentences that don’t contain any of the words but have synonyms or similar ones. This could be both an advantage and disadvantage, however, because the model might select certain sentences that although are similar, aren’t really relevant for that specific query. For example, when feeding the query “Do you need visa sponsorship?”, the model would selected sentences such as: “I have a driver’s license”.

3.3 Queries used and retrieval results

3.3.1 Immigration Status

All three variables were extracted with both approaches and two different queries: a keyword-based query and a question. With regards to retrieving users who needed sponsorship, Table 1 shows the queries and thresholds used, as well as, the number of users and true positives

(users who actually required sponsorship as reviewed manually afterwards) retrieved. The BM25 model, with the keyword-based query, retrieved the most true positives from the texts; even though, the bi-encoder model had a better ratio of true positives, it’s unclear if this holds when reducing the threshold to allow the retrieval of more users. Therefore, we consider that the BM25 model is the most appropriate for this specific task.

Table 1: Users that require visa sponsorship retrieved by IR models.

Model	Query	Threshold	Total Users Retrieved	True Positives
BM25	Do you need sponsorship?	5.0	248	84
	immigration visa sponsor sponsorship	0.0	213	183
multi-qa-mpnet	Do you need sponsorship?	0.4	198	87
	immigration visa sponsor sponsorship	0.4	97	84

3.3.2 Country of origin

The two queries we tested with in order to retrieve users’ country or nationality were: “I am from” and “Where are you from?”. However, almost every word in these queries are included in the stop words lexicon that we are using before feeding the queries into the BM25 model, so it’s not able to retrieve any valid sentence (future iterations will not remove stop words from the text so a proper comparison will be made). The transformer model is able to retrieve 243 users with valid nationalities using the keyword-based query.

3.3.3 Age

For this task we used the queries: “How old are you?” and “I am years old”. There was no need to select a threshold when retrieving users’ age, since out of the top sentences we just extract the numbers and performed the subsequent transformations mentioned before. The BM25 with the keyword-based query was able to retrieve 237 users with apparent valid ages and 223 with the question based one, while the bi-encoder performed better with the question query it only manage to extract 195 users with valid ages.

Table 2: Average value of quantitative variables by province. The statistic in parenthesis is specified in the header.

Province	N (prop.)	Rate (sd)	Years Ex. (sd)	No. Children (sd)	Age (sd)
Alberta	1674 (0.17)	19.45 (4.55)	8.38 (6.46)	3.03 (2.06)	26.93 (7.85)
British Columbia	1592 (0.16)	20.98 (4.18)	7.6 (6.18)	2.78 (1.11)	26.03 (6.76)
Manitoba	232 (0.02)	16.79 (3.53)	6.7 (5.44)	2.78 (0.94)	25 (4.55)
New Brunswick	98 (0.01)	16.44 (2.78)	8.72 (6.96)	3.31 (2.36)	52 (NA)
Newfoundland and Labrador	49 (0.01)	17.49 (2.97)	6.25 (6.74)	3.12 (0.83)	NaN (NA)
Nova Scotia	198 (0.02)	18.44 (3.49)	9.13 (7.32)	4.2 (4.45)	22.67 (4.04)
Ontario	5085 (0.52)	19.66 (4.86)	8.19 (6.89)	2.99 (2.71)	27.6 (8.72)
Prince Edward Island	15 (0)	16.01 (2.11)	8.6 (8.08)	5 (NA)	NaN (NA)
Quebec	466 (0.05)	17.93 (3.78)	7.05 (6.06)	2.7 (1.27)	25.64 (7.76)
Saskatchewan	309 (0.03)	16.32 (3.29)	7.19 (6.97)	3.83 (2.37)	22.4 (6.69)
Territories	13 (0)	22.18 (3.99)	7.25 (3.3)	3 (NA)	NaN (NA)
National	9732 (1)	19.52 (4.65)	8.02 (6.64)	2.98 (2.35)	26.85 (8.03)

4 Initial results and observations

4.1 Summary characteristics by province and age

Table 2 shows summary statistics by province of specific characteristics reported by users. Ontario is by far the province with the most users with over 50% of these located there, followed by Alberta and British Columbia, which are both well over 1k users and above 15% of the total. Nationally, nannies advertise a mean hourly rate of \$19.51 CAD with a standard deviation of \$4.65, the province with highest rate in average was British Columbia with \$20.98, followed by Ontario at \$19.66. The province with the lowest average rate was Prince Edward Island with \$16.01. With regards to the years of experience the national average was around eight years.

Figure 1 shows the distribution of age over all users and by province. The median over the whole population is 25, and the provinces' median ranges from 20 to 25 years old. It's clear that the distributions of the total users and for each provinces are right-skewed.

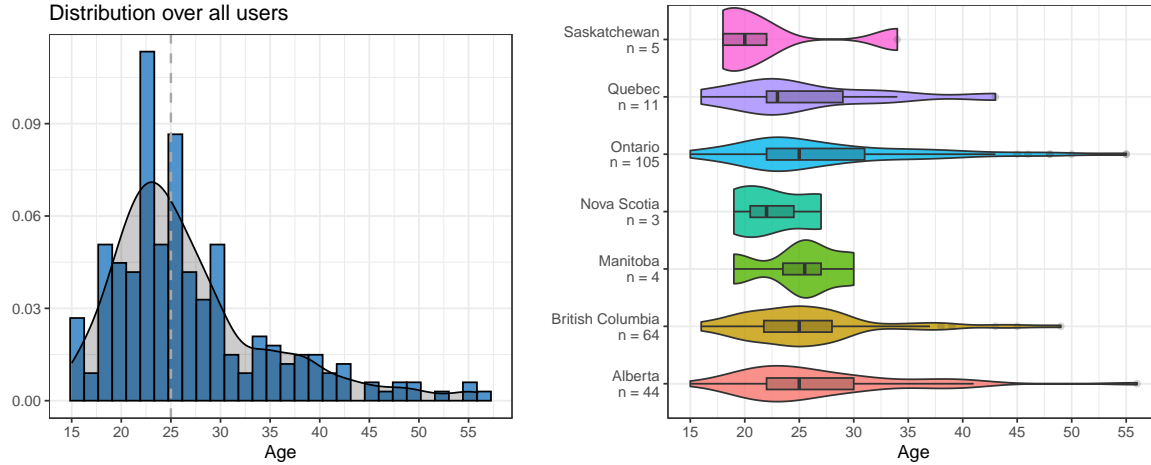


Figure 1: Age distribution of users and by province (selecting those with at least 2 users with valid ages).

Table 3: Statistics by immigrant status.

	Sponsorship		No sponsorship	
	Mean	Std dev.	Mean	Std dev.
Rate	19.474	4.254	19.517	4.659
Years Exp.	6.723	5.219	8.070	6.682
Age	28.400	6.653	26.780	8.094
No. Children	3.962	9.303	2.950	1.663

4.2 Nannies needing sponsorship

Table 3 shows users who require visa sponsorship charge a similar amount to those who don't. The distribution of reported years of experience of nannies who require sponsorship is much less spread out than those who doesn't. Figure 2 shows the number and proportion of users requesting sponsorship by province. Unsurprisingly Ontario has the most foreign users, but provinces such as Saskatchewan suggest higher proportions.

4.3 Country of origin

The majority of foreign users are originally from the Philippines, followed by Mexico with 21 users. Figure 3 shows the differences in rate distribution and reported years of experience by country of origin. Nannies from the Philippines are willing to take a substantially lower pay than others, with a median hourly rate of less than \$16 CAD and some users even going down

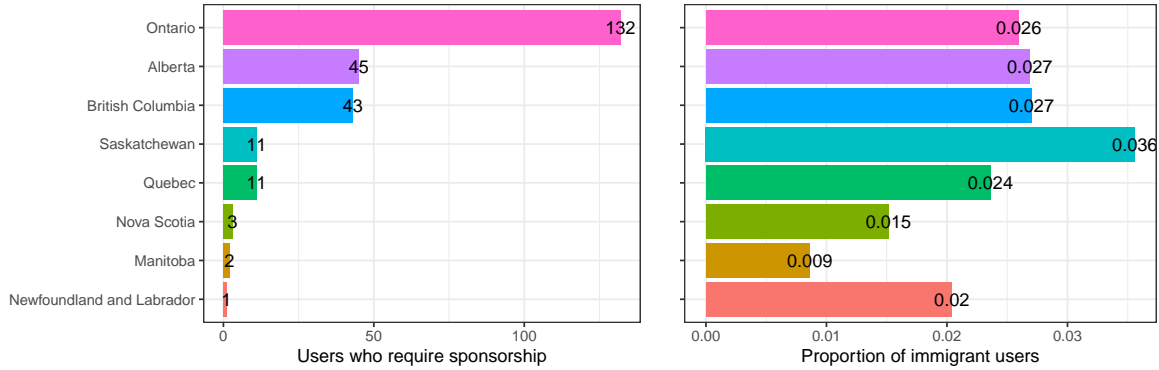


Figure 2: Number and proportion of users who require sponsorship by province or territory.

to \$12 per hour. Users from Mexico, Japan, Colombia or Brazil have an hourly rate much closer to the global median at \$20 an hour.

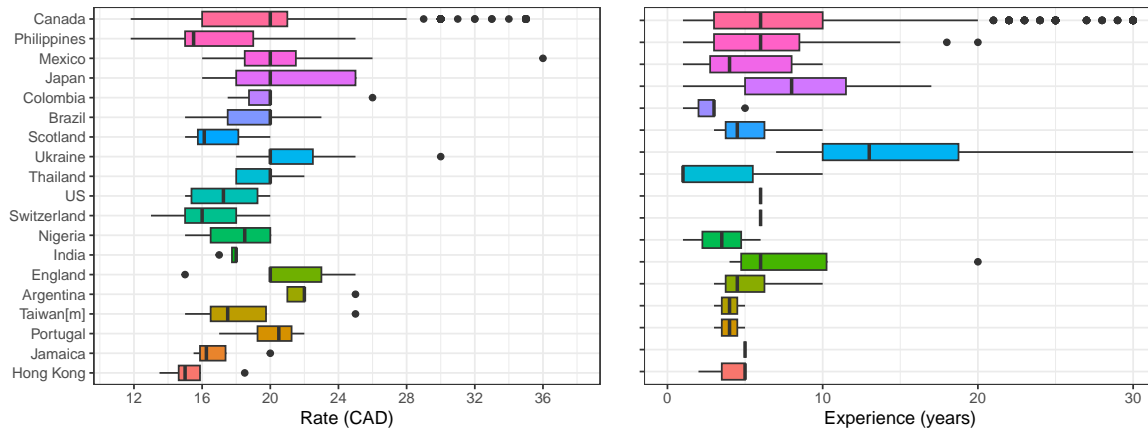


Figure 3: Distribution of rate and years of experience by country of origin.

4.4 Comparison to census data

We collected data from the 2011 census from IPUMS on women working in home care and educational support occupations. Table 4 shows statistics by province. We found 7,652 women in Canada and 34% were from Ontario, significantly lower than what was found on the canadiannanny site, where 52% reported being located in this province. The discrepancy might be due to a lack of Québécois nannies on the site, only 5% reported Quebec as their location, which doesn't match true population proportions. Census data assigns 15% of people to Quebec. The reasons for this might be that, even though users can write their description in

French or any language they want, the site itself is in English with no French version.

Table 4: Census statistics by province.

Province	N (prop.)	Non-citizens (%)	Employed (%)	Mean Age (sd)
Alberta	917 (0.12)	182 (20%)	788 (86%)	41.166 (13.41)
British Columbia	1086 (0.142)	200 (18%)	918 (85%)	43.022 (14.14)
Manitoba	379 (0.05)	29 (8%)	330 (87%)	40.343 (13.79)
New Brunswick	258 (0.034)	5 (2%)	220 (85%)	44.736 (13.51)
Newfoundland	230 (0.03)	0 (0%)	175 (76%)	46.43 (13.16)
Nova Scotia	255 (0.033)	7 (3%)	211 (83%)	42.506 (13.37)
Ontario	3006 (0.393)	593 (20%)	2515 (84%)	42.17 (13.88)
Prince Edward Island	48 (0.006)	2 (4%)	39 (81%)	41.938 (13.85)
Quebec	1207 (0.158)	108 (9%)	1032 (86%)	44.535 (13.7)
Saskatchewan	237 (0.031)	12 (5%)	205 (86%)	40.958 (14.11)
Yukon	29 (0.004)	3 (10%)	23 (79%)	43.069 (13.41)
National	7652 (1.00)	1141 (15%)	6456 (84%)	42.643 (13.85)

Average age range from 40 to 46 years old with a standard deviation of around 13-14 years across all provinces and at the national level. These numbers are much higher than what was observed from the users on the site, where the mean age was 26.85 at the national level and ranging from 22 to 27 across all provinces. This might indicate that a higher proportion of young people are using the site, and they're using it to find part-time or temporary jobs.

5 Summary and future work

In this project we utilize the large amounts of data available online to glean insights into the informal childcare sector in Canada. Our work to date has involved extracting almost 10,000 unique nanny profiles and using information retrieval techniques to extract information on age, country of origin, and immigration status. Initial observations suggest some evidence for systematic differences in advertised rates and other characteristics by immigration status and country of origin. Comparison to census data suggest some aspects of non-representativeness.

Future work will focus on several aspects. Firstly, we are working on improving the information retrieval algorithm to extract more information on potential migrants. Secondly, we will investigate image processing techniques to extract information about age and possibly gender

from user profile pictures. We will also consider methods such as multi-level regression and poststratification to adjust for non-representativeness in the data. Finally, we plan to collect multiple waves of profiles over time, in order to get estimates for the implied turnover rates of nannies seeking employment.

References

- Brown, D. 2022. “Rank-BM25: A Collection of BM25 Algorithms in Python.” Available at https://github.com/dorianbrown/rank_bm25.
- Charters, T., and L. Findlay. 2023. “Characteristics of Child Care Centres Serving Children Aged 0 to 5 Years in Canada, 2021 to 2022.” StatCan Economic and Social Reports. Available at <https://www150.statcan.gc.ca/n1/pub/36-28-0001/2023003/article/00001-eng.htm>.
- Crook, Sarah. 2020. “Parenting During the Covid-19 Pandemic of 2020: Academia, Labour and Care Work.” *Women’s History Review* 29 (7): 1226–38.
- Jurafsky, Daniel, and James H Martin. n.d. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.”
- Milkman, Ruth. 2023. “Stratification Among in-Home Care Workers in the United States.” *Critical Sociology* 49 (1): 11–22.
- Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks.” *arXiv Preprint arXiv:1908.10084*.