Analyzing Entrepreneurial Challenges and Opportunities for Women in Canada: A LinkedIn-Based NLP Study

Content

1. Introduction

1.1 Background

Since the pandemic, Canada's labor and entrepreneurial environment has undergone significant changes. Intermittent lockdowns, labor shortages, and the quick influx of large-scale immigration have all intensified the shifts in this environment. Understanding the role that women entrepreneurs have played during this period, along with the opportunities and challenges they have encountered, offers valuable insights.

This project aims to use LinkedIn, a prominent social media platform for professionals, to explore the challenges and opportunities encountered by Canadian women entrepreneurs. By focusing on LinkedIn posts and other user related attributes, we expect to gain insights into the entrepreneurial experiences shared by women in this region.

1.2 Main Objectives:

The project included four main objectives as per following:

1. To explore topics related to challenges Canadian female entrepreneurs shared on Linkedin.
2. To identify topics related to opportunities Canadian female entrepreneurs shared on Linkedin.
3. To explore whether the topics of posts have changed over recent years.

2. Dataset Description

Feature	Description	Type
Post_ID	Unique id for each post. Eg, 1, 2, 3.	Nominal
Post_Contents	Textual contents of posts. They may include emoji code and unrecognizable languages.	Categorical
User_ID	Unique id for each user. Eg, 1,2,3.	Nominal
User_Name	Username.	Categorical
User_Position	User position in the company. 3 categories: Founder, Cofounder, CEO	Categorical
User_Company	User company name.	Categorical
Province	User province.	Categorical
NAICS_Category	North American Industry Classification System for user industry.	Categorical
NAICS_Code	Code for user industry.	Nominal
Year	The year of each post.	Categorical

Key Features:

Posts_Contents is used for NLP analysis.

Province, NAICS_Category, NAICS_Code. Year are used for EDA.

3. Method

3.1. Data Collection

3.1.1 Semi-Automatic method

We use a Chrome extension, Simplescraper, to scrape the textual contents of the posts.

The workflow: we clicked into a user's posts page and repeatedly scrollscrolled down the webpage until we loadloaded all contents of posts we need. Then we ran the Simplescraper. Ideally, all textual information will be captured and can be saved into a CSV file.

There are two problems with this method. First, if a user posted in large numbers, it may take more than half an hour to manually load all his/her posts. During the process, webpages often crash, and the work needs to be started all over again. Second, even though all information has been successfully loaded, the Simplescraper may crash when running on this large amount of work. To solve this problem, we load, scrape, and save the work nodes multiple times for one user.
In conclusion, while this semi-automatic method is reliable and saves us a lot of energy, it still works inefficiently when handling large amounts of textual information.

3.1.2Manual method

We scraped data regarding people manually. Since Linkedin is highly restrictive in access to users' information, many users' information is not open to the public. The only way to work around this problem is that we first used key words such as 'women', 'entrepreneur', 'owner', 'leader', 'female', 'founder', 'co-owner' and 'co-founder' to get information of some public figures. Then we can click into these public profiles and get 'recommended' people. In this way, we collected relevant information from 192 users with roughly 30,000 posts.

We scraped time data completely manually.

Simplescraper may only work on static text stored in HTML. It is likely dynamic timestamps are generated after HTML is loaded which cannot be captured by Smiplescraper. Therefore, we need to manually fill in every timestamp.

Again, the manual method is reliable but lacks efficiency. It takes approximately 80 hours only to complete timestamps.

3.2. Data Preparation

We used nltk and gensim library to convert text to lowercase, remove punctuation, stopword, and convert words to root forms.

3.3. EDA

We conducted exploratory data analysis (EDA) on the number of posts by users, their locations, industries, and the distribution of posting times.

3.4. Topic modelling

3.4.1 Creating sub-dataset

We create two sub-datasets in topics of ‘challenge’ and ‘opportunity’. Instead of just using the keywords, we use ‘en_core_web_sm’ pre-trained model from library ‘spacy’. The model can vectorize text and be used to find texts with similar meaning by comparing word vectors.

By using the model, we create challenge dataset with 2083 rows and opportunity 3363 rows.

3.4.2. Topic grouping

In the challenge dataset, we first use Latent Dirichlet Allocation (LDA) to group 50 topics, each of which contains 5-6 words. Most words are key words like ‘women’ or ‘business’, which are too general to make any meaningful analysis.

To deal with above issue, we made a dictionary of the resulting 50 topics, as word in the topics as key, and the frequency of corresponding word as value. In this way, we compare which word might be highly frequent in the topics of ‘challenge’. This will give up a direction we can look deeper into.

Finally, we topic modeling again, but with the newly found frequent keyword. As a result, we can find some meaningful aspects that answer our objective question.

We use the same logic to our opportunity dataset.

3.4. Word Cloud

Word Cloud is regarded as a data visualization technique that displays text data where the size of each word indicates its frequency or importance. This method provides an intuitive visualization of key themes and terms within a text corpus.

It is quite necessary to calculate the frequency words in LinkedIn Users’ post content to find out the real situation they are facing.

To get the satisfying final visualization in Word Cloud, we plan to find out 20 most frequency words through the processes of encoding, cleaning and then analyzing. For the analyzing tool and libraries,we used Wordcloud, Matplotlib, Pandas, Numpy. For the text processing, we used Tokenization and Frequency calculation. We set parameter and made bar chart of top words, and printed keywords according to frequency.

3.5. TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical metric used to evaluate the importance of a word within a document compared to a collection of documents. Term Frequency (TF) measures how often a word appears in a document, indicating its importance within that specific document, while Inverse Document Frequency (IDF) assesses how unique the word is across all documents, giving higher scores to rarer terms. Common words such as "the" or "and" receive lower scores, as they appear frequently across many documents. The overall TF-IDF score is the product of TF and IDF, highlighting words that are both common within a document but rare across the collection, making them essential for distinguishing that document.

This method is useful in various applications, such as search engines, document classification, text summarization, and content recommendation systems. In a specific context, each individual post can be treated as a separate "document" while the collection of posts represents the entire corpus. Applying TF-IDF to this structure helps in identifying key terms within each post in relation to the entire set of posts.

4. Results

4.1. EDA results and interpretation

From the number of posts distribution, we found that most users are moderately active, with fewer very active or minimally active users. Therefore, the textual data is not biased by some very active users.

From location and NAICS distribution, we found that user activity is heavily concentrated in Ontario and British Columbia and user base is heavily concentrated in professional services and retail sectors.

From the posts timeline analysis, we found that users are disproportionately active in recent years. In addition, not only the number of challenge related posts, not peaked in covid period, but increased over time, the proportion of those posts shows the same pattern. (see JupyterNotebook Section 3.2, Page 15)

The reason for the disproportional distribution over time: first, linkedin platform is more popular in recent years. second, only recent active users are recommended to us in the data collection stage, which biased our data.

The reason for the continued increase challenge posts after covid: first, people may still struggle after covid. second, people in real challenge may be reluctant to share on Linkedin.

4.2. Topic modelling results and interpretation

After first round of topic grouping in the challenge dataset, we found that 'support' showed up in half of the 50 topics we created. This led us to ponder: what kinds of support are needed when people talk about support and challenge? Therefore, we further use 'support' as keyword to group topics in the challenge dataset. Finally, we found some frequent keywords that are helpful.

'community', 'life+work', 'awardees', 'learn', 'financial', 'investment' (see JupyterNotebook Section 4.2, Page 21)

A simple explanation for every key word:

Community: user expressed community support essential to their success.

Life + work: these two words show up together in one topic, implying support regarding life and work (balance) when facing difficulty.

Awardees: not common in daily use but very frequent in Linkedin posts. The word implies business owners' need for support from third party's recognition to increase credibility.

Learn: support for skill and knowledge.

Financial and Investment: These two are not grouped in one topic, but both imply that financial support may be an essential issue when talking about challenge.

We used the same method as ‘challenge’ to explore key words for opportunity. However, the result doesn’t show much meaningful insight.

4.3. Word Cloud results and interpretation

For challenge, after COVID (2022-now), the key words didn’t change radically according to the results of the word cloud, still focus on women and business, however, we noticed the word *Global merged on the list,* which indicating there are more collaboration and traveling across countries for Canadian women entrepreneurs.

For opportunity, during COVID, the most frequent words related to opportunities are as follows, our, opportunities, women, business, etc. After COVID, the most frequent words didn’t see a radical change, which is the same situation as ‘challenge’.

4.4. TF-IDF results and interpretation

The TF-IDF analysis provides insight into the words most closely associated with the keyword's "opportunity" and "challenge" in the LinkedIn post data. For "opportunity," the top related words include broad themes like "women," "business," and "us," as well as more specific concepts like "great," "new," "canada," and "team." This suggests the discussions around "opportunity" often involve topics related to gender, industry, geography, and collaborative work environments.

The words linked to "challenge" paint a different picture, highlighting more global and empowering themes such as "women," "global," "un," "empower," and "world." This indicates the challenges discussed often have to do with broad, systemic issues that require collective action and advancement, particularly for underrepresented groups. The presence of words like "business" and "us" also suggest challenges faced within organizational and professional contexts. Together, these results give some context around the types of opportunities and challenges being discussed in the LinkedIn posts.