# WHERE ARE THE WOMEN ENTREPRENEURS?

## A Case Study on Standard Chartered's SocialAI Model to Identify Women-Owned SMEs in Bank Portfolios

Standard Chartered is a leading international cross-border bank headquartered in London with a presence in over 50 of the world's most dynamic markets. The bank has taken steps to advance financial inclusion and gender equity through its work with women entrepreneurs. As part of its broader social sustainability agenda, the bank has sought to improve access to finance and economic participation for underserved groups, including women-owned and women-led businesses, recognising their role in driving inclusive growth, job creation and community resilience.

Many banks on similar journeys to elevate women entrepreneurs struggle to segregate women-owned and led small and medium businesses ("WSMEs") as a distinct segment, as the banks do not have the data (and therefore the resulting analytics) to understand how women entrepreneurs can be best served. Even when banks update systems to capture sex-disaggregated data for new clients, they often face the much larger hurdle of how to identify women-owned businesses in their existing portfolios. For Standard Chartered, the question became, what are the ways it could overcome these challenges to segment existing WSME clients around the world without an exhaustive and expensive new data collection exercise?

In order to try to answer this, the Sustainability Data team piloted an innovative approach to assess feasibility of disaggregating its SME portfolio by sex, with over 90% accuracy, without the collection of private personal data. The team developed an AI model using a hybrid approach to explore feasibility of identifying whether enterprise owners are male or female using the first name, country-adapted, and public and internal datasets.

This Case Study, targeted to Financial Service Providers, aims to provide an overview of the model Standard Chartered explored, including the processes for developing such a model, the challenges encountered, and potential business implications.

## The Benefits Of Sex-Disaggregated Data Collection

Sex-disaggregating data will allow Financial Service Providers (FSPs) to better understand and serve the WSME segment. In the financial services industry, data allows companies to estimate the size and growth of market segments, and to understand their behaviour and characteristics. An intentional analysis of the women entrepreneurs market is critical to create tailored products and services, to develop strategy based on performance and profitability of different segments, and to understand how to market to groups of customers. For an FSP, the benefit of high-quality sex-disaggregated entrepreneur data is two-fold. First, an FSP can expand its customer base among women entrepreneurs in lending, investing, or banking services, and improve their own bottom lines (see ABN AMRO 2022; Women's World Banking 2020). FSPs can also meet corporate citizenship or social sustainability goals by helping to promote financial inclusion for women entrepreneurs, who are crucial to equality within societies, to economic growth, and to innovation.

**Authors:** Aurica Balmus (We-Fi), Evangeline Zhao (Standard Chartered), and Xu Jun (Standard Chartered) | **Editors:** Wendy Teleki (We-Fi) and Elizabeth MacBride (We-Fi) | **Research:** Felicia Siegrist (We-Fi) | **Reviewers:** Anna Tabitha Bonfert (World Bank), Frederic Meunier (World Bank), and José Felix Etchegoyen (International Finance Corporation)

Sex-disaggregated data is also useful for policymakers and regulators looking to encourage greater financial inclusion of women-owned and led businesses. Regulators can use global and national data to establish baselines to understand how much financing is going to women and women entrepreneurs. Sex-disaggregated data also helps them to assess whether their financial inclusion strategies are working and where constraints and bottlenecks remain. As a result, FSPs that invest in equitable data systems, including those that sex-disaggregate data, may be able to access pools of development or sustainable investment capital.

## Social Sustainability at Standard Chartered

Standard Chartered is widely recognized for its ambitious sustainability objectives. The bank aims to mobilize USD 300 billion in sustainable finance by 2030 and is working across its business and functions to operationalize its target of reaching across net zero in its financed emissions by 2050. It is also applying its innovative mindset to find new ways to unlock the areas where capital isn't yet moving at scale, helping contribute to the advancement of the broader sustainability ecosystem. Standard Chartered also aims to drive social impact through both its commercial and philanthropic activities. This reflects its commitment to delivering sustainable, inclusive growth and prosperity across the markets in which it operates.

Gender equality and inclusion have long been part of this sustainability vision. In 2016, Standard Chartered signed the UK HM Treasury Women in Finance Charter, pledging to have women occupy 35% of senior leadership roles by 2028. It also has received an EDGE Certification in six markets for its work on sex equality and takes part in the Bloomberg Gender-Equality Index. Data, analysis, and reporting are crucial to these workplace initiatives, which seek to include more women in decision-making in the financial sector.

## Challenges In Collecting High-Quality Sex-Disaggregated Data

In 2023, the bank faced a number of data challenges as it looked to apply its sustainability strategy across its global operations. These ranged from diverse data protection and privacy regulations across countries to vast differences in data collection practices in each country of operation. On the operational side, Standard Chartered faced three main challenges in disaggregating data by sex. First, categorizing the bank's existing portfolio by the sex of the business owner was extremely difficult, as this information had not been collected when the bank initially established relationships with customers or potential customers. Second, the bank needed to determine an appropriate definition of what constitutes a woman-owned business. Finally, it faced the challenge of developing one solution for both new and existing customers that would be replicable across various geographies and cultural contexts.

Standard Chartered executives were not alone in realizing that their systems needed to change. Few banks can systematically identify women entrepreneurs or women-run companies in their portfolios, whether they are sole proprietors or women-led SMEs with multiple owners (World Bank Group 2020; GBA and Data2X 2018). Despite the significant benefits that such data could provide to promote financial inclusion, foster equitable economic growth, and expand markets served by FSPs, such data remain scarce.

Many FSPs encounter internal challenges when categorizing customers by sex (GBA, Data2X, and IDB 2019). In most cases, internal bank data collection systems are designed to disaggregate information by product type. When management information systems allow for segmentation analysis, it is rare that demographic factors like sex or age have been tagged. As a result, any decision to collect sex-disaggregated data requires new enhancements to internal systems, which require financial investment and employee time.

Even when efforts are made to collect sex disaggregated data, quality can be poor (CGAP 2022). There are often challenges related to the definition of what constitutes a women-owned or women-led business. An FSP may find it straightforward to identify the sex of a sole proprietor, but the process becomes more challenging for any SMEs with multiple owners or chief executives. According to some definitions, an enterprise may be categorized as a WSME based on either the ownership structure or the management roles within the firm. More broadly, the absence of nationally accepted definitions for terms like "women-owned" or "women-led" can create inconsistencies in the data collected over time, or by different financial institutions within the same country, or by operations in different countries.

This lack of standardization also makes it difficult to build comparable national datasets and draw meaningful insights.

Beyond inconsistent definitions, socio-cultural factors also affect the data collection. For example, ownership shares may not accurately reflect whether a woman genuinely controls an enterprise registered in her name or conversely if she effectively runs a business formally registered in, for example, her husband's name. These challenges underscore the need for locally informed definitions and robust verification processes to ensure that the data collected is accurate and reflects local customs and realities.

If FSPs do not currently tag their clients by sex, they typically need to define what will be counted as a woman-owned business and then develop methodologies to tag both existing clients and new clients (also see ADB 2023) as male or female. For new applicants, they should establish systems to categorize individuals who are owners or senior managers of SMEs, ensuring sex-specific data is collected from the outset. For existing clients, FSPs may need to develop a methodology to retroactively disaggregate data within their existing client databases, allowing for a comprehensive understanding of their current portfolios. These tasks can be daunting and time-consuming, but machine learning technology may be able to be leveraged to predict whether business owners are men or women subject to appropriate data security and responsible AI considerations.

Currently, the determination of whether an SME is women-owned or women-led is carried out during onboarding of clients. Standard Chartered has adopted a universal definition for a woman-owned business: (a) For women-owned business: (i) at least 51% owned by woman or women; or (ii) the business is founded by a woman (i.e. 50% shareholding); or (b) For women-led business: (i) at least 20% owned by woman or by women; and (ii) with a woman as CEO, COO, President or Vice President or where a woman is a signatory of the business's legal documents and financial accounts; and (iii) has a board of directors at least 30% comprise of women, where a board exists.

With these parameters in place, Standard Chartered leveraged fully public, open-source technology to approximate whether the existing owners of a set of SME clients were male or female. This has allowed the bank to explore the feasibility of understanding the share of women-led businesses within the existing SME lending portfolio. As part of this pilot, strict guidelines were put in place as to who can see the outputs of the model, in compliance with data protection regulations.

## Components Of The Standard Chartered SocialAI Model

The Sustainability Data team that developed the Standard Chartered SocialAI model used a hybrid approach to pilot identification of the likely sex of enterprise owners using the first name, country-adapted and internal datasets. It is a custom-built model to pilot the ability to gain insights into the existing portfolio of clients to better serve client needs without additional data collection. The model integrated multiple modules and machine learning algorithms into a unified framework of prediction to achieve a better performance than a single process. The SocialAI model, is capable of predicting with 90% accuracy on the tested dataset, without relying on any actual sensitive personal data.

Implementing the model required significantly more effort compared to a typical AI/ML project, encompassing aspects such as data approval, personal data impact assessment, and responsible AI. The most challenging hurdle was obtaining data approvals, as the data was considered part of personal data even though it couldn't be traced back to an individual. Numerous approval queries to other teams within the bank focused on ensuring proper customer consent, data and process management to prevent issues related to process integration, predictive nature, and gender discrimination.

The SocialAI model is highly configurable and allows detection via a process that includes using 1 digit in national ID and name database lookup. Specifically, the SocialAI model leverages integration of name datasets, Bidirectional Long Short-Term Memory (Bi-LSTM) deep neural network and ID specifiers for detection and prediction based on the person's first name and country, which improves the model's prediction power and accuracy. By default, the model either detects information from 1 digit from IDs if available or performs a name lookup in databases (providing statistical results) and predicts the sex of an enterprise owner. The output provides a prediction for each individual (e.g., female or male) along with the confidence level.

The SocialAI model includes four major components:

(See Annex for more detail)

**1. Name Database with Lookup Module:** This module takes an individual's first name and country as input to query databases, which forms the basis for making sex predictions.

**2. Name Sex Prediction Using Machine Learning:** The SocialAI model applies two algorithms to encode names as numerical vectors specific to each country. This step leverages deep learning (bi-LSTM) and NLP techniques (text tokenisation and embedding), utilising both character-level and token-level features to predict sex based on name and country.

**3. Special Feature Module:** This component uses additional information, such as specific digits embedded in national IDs that indicate the sex of an individual. This applies in countries where national IDs include such indicators.

**4. Rule Engine:** A configuration module that defines the priority order of detection/prediction results from the three modules. For example, ID-based results take the highest priority, followed by database lookup, and finally, name prediction.

## Lessons Learned From The SocialAI Model At Standard Chartered

The development of the SocialAI model at Standard Chartered provided the bank with new insights into the complexities and challenges of using machine learning to produce sex-disaggregated data in a global banking context. This experience highlighted the importance of legal compliance, data quality, tailored approaches, and fairness in building models that are accurate and ethically sound in different cultural settings.

One key lesson learned was the variability in prediction accuracy across different countries. This variation was primarily driven by cultural, linguistic, and naming conventions unique to each region. These differences highlighted the difficulty of applying a uniform data model globally and emphasised the need for region-specific adaptations to improve accuracy. For example, "Kiran" and "Amit" are traditional male names in India, while they might be sex-neutral in the West and Middle East. "Amit" is predominantly masculine, meaning "infinite" or "Limitless" in Hindi and Sanskrit, while it is more neutral in Hebrew, meaning "Friend" or "companion". The situation is similar for the Malay name "Zulkifli" and "Khairul" (usually male) or "Nurul" and "Zulaikha" (usually female). As some names are not common in the global West, the general-purpose predictor without country-specific databases cannot provide a high-confidence answer.

Another key takeaway was the critical importance of legal compliance in data usage. Not all public datasets are legally permissible for projects, making it essential to conduct thorough legal reviews and, where necessary, to create internal datasets that comply with all relevant regulations. Ensuring that data acquisition teams are well-versed in data compliance and intellectual property rights is essential to avoid legal pitfalls.



***Key lesson:** Develop a protocol for regular legal reviews of data sources and ensure a well-trained team to avoid legal issues. If external data sources pose legal risks or are not adequately licensed, consider using or creating an internal dataset that complies with all legal requirements.*

The project also revealed the importance of data quality. Large datasets often contain errors, such as incorrect labels, or may include duplicate entries from various sources. The quality of the data for some internal non-mandatory data fields can also vary and be of lower quality. These issues highlight the need for rigorous data cleaning and pre-processing stages to ensure the accuracy and reliability of the data used by the model. Quality control measures, including sample checks and removing duplicate entries, are critical for maintaining the integrity of the dataset.



***Key lesson:** Establish routine data cleaning to handle inaccuracies and duplications, train data teams on quality assurance methods and tools, and develop a comprehensive processing toolkit based on internal business characteristics and the logic of the model.*

Another significant lesson learned was the challenge of training the model and validating the results. Variations in data distribution between training and testing can result in performance inconsistencies, particularly due to the significant differences in naming conventions across countries. For example, some names have a small number of characters (e.g., names based on the Latin alphabet), while others, like Chinese, Japanese, or Korean names, may consist of a larger number of characters. The development process highlighted the need for advanced techniques, such as k-fold cross-validation and domain adaptation, to ensure consistent performance across different data subsets. Moreover, it became clear that different naming conventions require different approaches. For instance, character-level features may work well for names of Latin origin, while languages with complex character sets, such as Chinese or Japanese, benefit more from token-based vector methods.

This insight led to the development of a modular AI system, allowing preprocessing and feature extraction techniques to be tailored to the specific characteristics of the input data.



***Key lesson:** Design and merge datasets to ensure they cover a wide and representative population. Techniques such as k-fold cross-validation and domain adaptation are important to ensure the model performs well across diverse data subsets. Using a shuffled train-test split can be further beneficial to fine-tune the model. Using hybrid methods for different types of names is recommended when there is a large variability in names. Perform continuous research and development to refine methods based on linguistic and cultural nuances.*

The Standard Chartered team also had to decide how to handle imbalanced datasets. The data often exhibited inherent imbalances in the Standard Chartered datasets, with ratios of female to male data varying significantly across different countries. In some countries, the ratio of female to male could be as large as 20% to 80%, leading to significant disparities. This imbalance can create issues in developing AI models, as algorithms might become biased toward the majority group, resulting in inaccurate predictions or unfair outcomes. For example, if the data heavily favors one sex, the model may struggle to make reliable predictions for the underrepresented group. To address this, the team implemented techniques like resampling, where data from the underrepresented group is duplicated or the overrepresented group's data is reduced to create balance. They also explored advanced methods such as algorithmic adjustments to ensure fairness, like weighting the data to give more importance to the underrepresented group.



***Key lesson:** Implementing robust evaluation metrics that specifically measure performance of the AI model in imbalanced scenarios is crucial, along with continuous monitoring and adjustments to refine AI model performance in such cases.*

Finally, the project underscored the importance of fairness and inclusivity. To avoid biases and improve accuracy across diverse demographic groups, the model had to incorporate a wide range of naming conventions from various cultures. The built-in transparency in the model's decision-making process was an important step that allowed Standard Chartered to understand how predictions were made and allow the assessment of potential biases. This transparency, combined with the intentional inclusivity in data collection, enhanced the model's utility and fairness in global applications.



***Key lesson:** Strategically select fairness metrics, such as equal opportunity or predictive parity, in evaluating the fairness of systems. Choose metrics that align with the project's specific ethical and operational goals. Additionally, define an acceptable range of fairness to guide model adjustments and ensure the AI system treats all groups equitably.*

FSPs looking to adopt this approach could develop a similar model based on the provided key logic. This includes:

- **(a) looking up names in an existing database to identify sex of enterprise owners with set probabilities,**
- **b) using pre-developed models to train and predict sex of enterprise owners based on different character sets (e.g., Latin, or Chinese), and**
- **(c) extracting sex of enterprise owners from national IDs based on specific digits, where applicable. However, this option may not be available in all countries.**

## Implementation Considerations for Broader Adoption

The development of SocialAI also surfaced practical challenges that financial institutions may face when seeking to adopt AI-enabled approaches to sex-disaggregated data at scale. First, broader implementation requires dedicated resourcing, senior sponsorships and cross-functional coordination across business, technology, legal, compliance, data governance and responsible AI teams. This is particularly important where the tool is applied across diverse markets with different naming conventions, regulatory requirements and data availability, as this requires continued testing, validation and local adaptation to ensure the model remains robust and appropriate for different contexts.

A second challenge is related to the assurance and auditability of AI-generated outputs. While assurance providers may review the underlying code, controls and model governance, AI-generated classifications may not carry the same evidentiary status as directly reported client information. This requires clear internal alignment on the appropriate use of the model's outputs, particularly where they are used to support portfolio analytics, strategic planning and the identification of legacy data gaps, rather than as a direct substitute for externally assured client-reported data. This may also create challenges in external disclosure of this data.

Finally, Standard Chartered has already begun strengthening direct gender-data collection for new SME clients through its onboarding processes. This created an important governance consideration on how to apply AI-enabled retrospective classification to existing clients in a responsible and transparent manner. In particular, the bank has to ensure that any approach in potential use of the model for legacy portfolio tagging was proportionate, well-controlled, and did not create the perception of artificially expanding reported women-owned SME figures.

# ENDNOTES


ABN AMRO. 2022. “The value of inclusivity in banking.” Amsterdam: ABN AMRO.

ADB. 2023. “Steps for Integrating Sex-Disaggregated Data in a Financial Institution.” Manila: Asian Development Bank.

CGAP. 2024. “Supply-Side Sex Disaggregated Data for Advancing Financial Inclusion.” Washington, DC: World Bank Group.

GBA and Data2X. 2018. “The Way Forward: How Data Can Propel Full Financial Inclusion for Women.”

GBA, Data2X, and IDB. 2019. “Measuring Women's Financial Inclusion: The Value of Sex-Disaggregated Data.”

IFC. 2017. “MSME Finance Gap. Assessment of the Shortfalls and Opportunities in Financing Micro, Small and Medium Enterprises in Emerging Markets.” Washington, DC: World Bank Group.

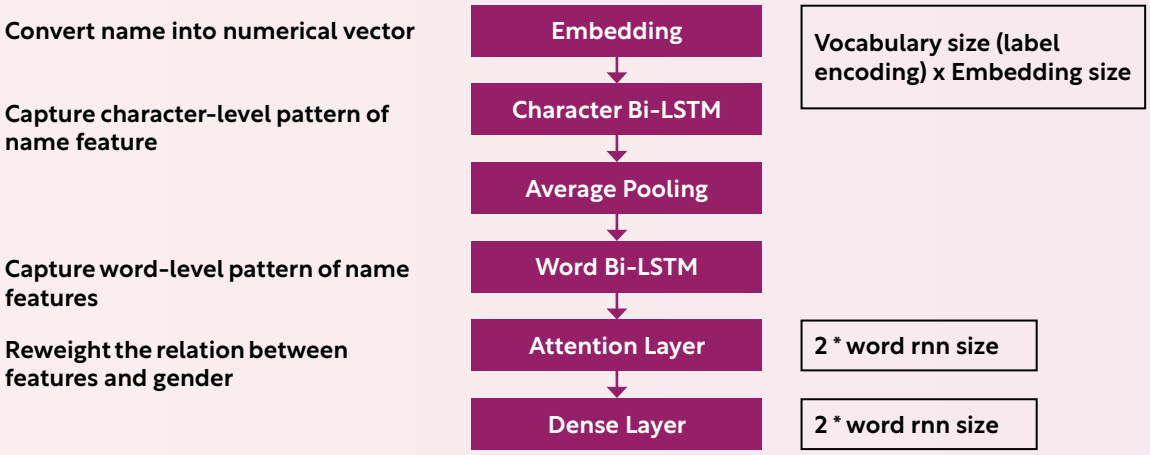
Women’s World Banking. 2020. “Empowering MSMEs: Creating a Better Banking Experience for Women-Led Micro, Small, and Medium Enterprises in Kenya.” New York City, New York: Women’s World Banking.

World Bank Group. 2020. “G-20: Data Enhancement and Coordination in SME Finance: Stocktaking Report.” Washington, DC: World Bank Group.

# ANNEX

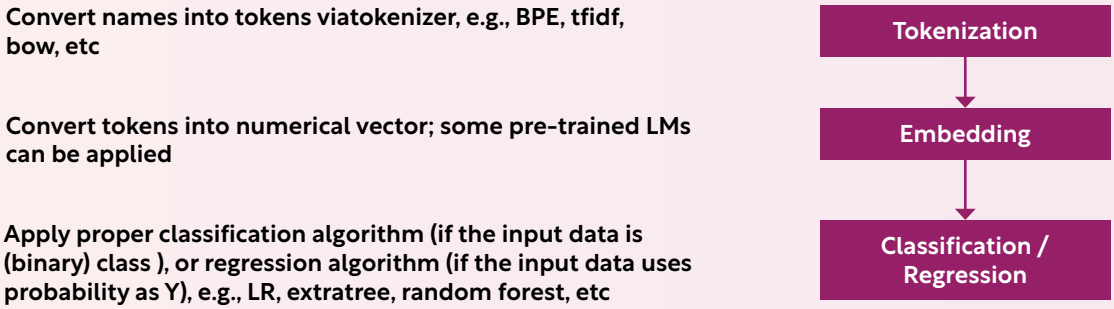
### Algorithm 1: bi-LSTM (bidirectional Long-Short Term Memory)

The deep learning bi-LSTM model captures the pattern of the names regarding sex in both character-level and word-level. This algorithm can be applied to Latin-like characters. CNN (convolutional neural network) algorithm




### Algorithm 2: Natural Language Processing NLP (text tokenization and embedding techniques)

The Natural Language Processing NLP text tokenization and embedding algorithm applies a tokenizer to separate the name string into tokens and convert them into vectors (embeddings). The users can employ the pre-trained (language) embedding models (e.g., word2vec) or train their own language models with the predefined format and sufficient corpus. Once the vectors are obtained, the users can apply any classification (if binary labels)/regression (if probability labels) algorithms to predict the sex. In the code, we provide a random forest-based algorithm with combined vectors of embedding and word counters as features.



# ANNEX

A unified framework for sex prediction works better than single processes to achieve a better performance. A unified model combines all three algorithms to address the shortcomings of each individual module.

| Module/ algorithm | Functionality | Pro | Con |
|---|---|---|---|
| ID Specifier | Provide the exact sex infor-mation directly based on the (national/resident) ID num-ber (either the full number or the specified digits) | Simply rule-based; highly ac-curate when based on author-itative data | Limited to countries where such IDs encode sex; not universally applicable. |
| name database lookup | Identify sex by referencing statistical databases of names | Simple and effective for names present in the data-base | Only applicable to existing names in the database, whose size and coverage can vary |
| ML - character-based | Predict the sex based on the (first) name using the model trained with self-contained character-level features | Able to predict any names theoretically; able to train the model without language model dependency | Performance may decline with longer names or un-known characters due to fea-ture extraction limitations |
| ML - token-based | Predict the sex based on the (first) name using the model trained with pre-trained em-beddings | Able to predict any names theoretically; applicable to process the names with very large number of characters, e.g., Chinese | Dependent on the quality and relevance of the pre-trained language model |