Final Project Report

Project Title:
Classifying Foreign-Invested Firm Production in China and Its Environmental Impact

Student Names
Ingrid Morales — 76088415, ingriam1@uci.edu
Abiramashree Adaikalam — 15468472,aadaikala@uci.edu
Viveka Agrawal — 63483720, vivekaa@uci.edu
Minjin Li— 49769412, minjinl1@uci.edu
Brian Lezama-Monterrosas — 44676688, blezamam@uci.edu
Chloe Florit — 83589814, cflorit@uci.edu

Link to Shareable Resources:
MDS 2025 Capstone

Introduction and Problem Statement

Foreign‑Invested Enterprises (FIEs) are central to China's growth, but existing registry and industry codes do not clearly describe what these firms actually do in the production process. Broad labels like "manufacturing" and "services" fail to capture where firms land in the production stage. This lack of granularity limits our ability to study how different types of foreign production activities shape environmental outcomes. To address this gap, we leverage the Foreign-Invested Enterprises in China (FIEC) database (≈600k firm records) with Chinese language business operation descriptions, supplemented by a web scraper targeting firm data from the Chinese Ministry of Commerce (MOFCOM) database. These free-text descriptions are messy, making them challenging to work with but also information-rich for determining firms' actual roles in the production pipeline.

Furthermore, our project aims to classify each firm's production activity by uncovering the latent semantic structure of these business descriptions using a combination of topic modeling (LDA), BERT-based embeddings, and supervised machine learning methods. This framework allows us to extract meaningful production categories and ultimately evaluate how different forms of foreign production activity influence environmental outcomes.

Related Work

Our project builds on three types of unsupervised NLP (Natural Language Processing) prior work: topic modeling, sentence embeddings, and Large LLM (Language Model) based pseudo-labeling, combined with supervised classification models.

Topic modeling: LDA (Latent Dirichlet Allocation) is a widely used probabilistic topic modeling that represents each document as a mixture of latent topics and each topic as a distribution over words. LDA was introduced by Blei et al. (2003) and has since become one of the most common choices for

discovering interpretable structure in large text based data. In our context, LDA provides a low-dimensional representation of foreign firms' business descriptions that can later be clustered and interpreted as a production stage pattern. However, its bag of words assumption discards word order and internal connections, which can be a potential limitation in understanding the actual production context of the foreign firm.

Sentence embedding: Beyond LDA's bag of words assumption, sentence level neural embeddings capture semantic similarity between texts in a continuous vector space. The Sentence-BERT adapts BERT model architecture into a siamese network trained on sentence pair objectives to produce high quality sentence embeddings that are well suited for semantic similarity, clustering, and downstream classification tasks (Reimers & Gurevych, 2019). In our context, we use a sentence BERT model text2vec-base-chinese that's focusing on Chinese text embedding to extend this idea and train our firm business operations to generate high dimensional embeddings which later reduced by UMAP (*Uniform Manifold Approximation and Projection*).

LLM-based pseudo-labeling: Recent research has increasingly explored the use of large language models (LLMs) as high-quality pseudo-labelers, especially in domains where labeled data is limited or costly to obtain. Studies have shown that LLMs can generate reliable weak labels for text classification tasks, often matching or surpassing traditional heuristic or rule-based methods. Approaches such as zero-shot and few-shot prompting enable LLMs to infer category boundaries without explicit supervision, while iterative self-training methods refine pseudo-labels by incorporating model feedback. Recent works also highlight the benefits of combining LLM-generated labels with lightweight supervised models, where the LLM provides semantic richness and the downstream model contributes scalability. Overall, LLM-based pseudo-labeling is emerging as an effective strategy for improving classifier performance, and reducing annotation effort in real-world applications.

Data Set

Our analysis is based on the firm level registry information from foreign-invested enterprises in China (FIEC) dataset, compiled by our project sponsor through web scraping of publicly accessible Chinese government records. The compiled dataset itself has not yet been formally published but the source is from the Chinese Ministry of Commerce (MOFCOM) under https://wzxxbg.mofcom.gov.cn/gspt/ (see Figure 12).

The full dataset contains 600,349 records of foreign firms, it contains 18 features including company name, enterprise id, start date, industry, registered capital, location, country of origin, and business operations. But for our analysis plan we only focused on the business operation description feature (businessoperations), which is a plain text feature and would be the sole input to our NLP pipeline. The other features are ignored in this project because our primary goal is to construct a production stage label that best represents a firm's business operations, rather than to perform a general classification task.

In the raw dataset, coverage of the business operations column is generally complete, we have 1.2% of the record missing in that case. For expectation the dataset behaves like a near census scale rather than a sample of population, providing access to the true population behaviors. However, the information

content of these descriptions varies widely. Among all firm records, the average description length is about 126 Chinese characters and the median is about 96 characters. But roughly 10.9% of firms have extremely short descriptions with less than 10 characters (Figure 1), which provide very little detail about the firm's actual activities and are therefore challenging to classify reliably.

Exploratory analysis

Exploratory analysis reveals strong recurring patterns in business operations. After segmenting each business operation within the description with Chinese punctuation, we compute the term frequency for each business segment. Figure 2 shows the Top 10 frequent operation keywords, the most common operation is technical consulting, appearing in 22.9% of firms, closely related operations such as "technical services" (19.9%), and "technology development" (18.2%) together indicate that a very large proportion of foreign-invested enterprises in China involved in technology oriented activities. The other cluster of segments such as "goods import & export" (13.4%) and "technology import & export" (12.6%) indicate the foreign trade oriented business operations.

At the character level analysis, the top 10 frequent characters in Figure 3 points to economic activity and regulatory context. The most common characters correspond to "operate" (67.8%) and "manage / operate" (63.7%) indicating that language about operating or managing a business is common in most firms. Another cluster of frequent characters "product / goods" (63.3%) and "item / project" (59.0%) reflecting the product based activities in many firms. The "batch / wholesale" (63.2%) and "develop" (56.6%) also indicate other directions of firms on marketing activities and development activities.

Technical Approach / Methods

Our analysis follows a multistep pipeline that transforms Chinese raw descriptions to structured production labels. The full workflow is described below.

Data preprocessing and Feature engineering

First, our workflow begins with cleaning and structuring messy raw Chinese operation descriptions from the FIEC database. We first clean the text by removing boilerplate legal phrases, and normalizing formatting. This includes removing punctuation, white space, leading and trailing blanks. Many Chinese descriptions have long lists of regulatory disclaimers that carry no semantic value for classification. We stripped these using a curated list of common legal phrases (See more details in section 4 of appendix). This step substantially improves tokenization quality before segmentation, LDA modeling, and BERT-embedding.

After cleaning, we preprocess each business-operation description using a multilingual tokenization pipeline tailored for mixed Chinese English text. This step converts unstructured text into meaningful tokens suitable for topic modeling. We start by loading English stopwords (NLTK) and constructing a Chinese stopword list from four large open-source lists (SCU, HIT, Baidu, and Goto456). These provide broad coverage of high-frequency, low-information function words. We then extend the stopword lists with:

- Domain-specific boilerplate words (e.g., "经营范围", "项目", "服务")

- Topic-2 “legalese” stopwords identified from our initial LDA run, which primarily captured regulatory approval language rather than substantive business activity
- Additional English legal terminology (e.g., *license*, *approve*, *special access*) suggested during model inspection

This allows the tokenizer to suppress regulatory noise and retain only semantically meaningful production terminology (See more details in section 4 of Appendix). Furthermore, for topic modeling, we convert the processed text into standardized Genism structures such as a dictionary (maps each token to an integer ID), and a corpus (Bag-of-Words representation for each document). These become the inputs for all downstream LDA topic models.

Modeling approach

We model firms’ production activities using unsupervised semantic structure learning. Our goal is to uncover coherent, interpretable groups of business-operation descriptions that correspond to different production roles.

Latent Dirichlet Allocation (LDA)

We train a Gensim LDA model on the cleaned Bag-of-Words corpus to learn a low-dimensional topic representation of each firm. We evaluated the following topic numbers: k = [2, 4, 5, 6, 7, 8, 10, 15, 20]. For each k, we compute coherence and perplexity. See more details under the evaluation section for the final results. The model outputs a topic word distributions and firm level topic mixture vectors. These topic vectors capture the dominant production activities present in each firm’s text (e.g., “manufacturing”, “technology services”, “chemical materials”).

Topic Embeddings for Clustering

Each firm is represented as a k-dimensional topic vector (one value per topic). These become structured numerical features for clustering. We assign each firm a dominant topic label, defined as the argmax of the topic mixture. This creates a discrete cluster ID for each firm. However, solely relying on the dominant topic can obscure meaningful variation among firms with similar but not identical topic distributions. We apply clustering to capture this finer structure. Clustering groups firms based on their overall semantic structure, producing interpretable clusters. We use t-SNE to project the topic vectors into 2-D space for visualization. This allows us to assess whether firms form coherent groups and whether the clusters produced under different choices of k align with meaningful structure.

BERT Embeddings

To represent each firm’s cleaned text in a continuous semantic space, we used *text2vec-base-chinese*, a Sentence-BERT model optimized for contextual representation in Chinese. Each document was encoded into a **768-dimensional embedding**, where distances reflect similarity in business activities rather than shared keywords. This embedding-based approach holds substantial advantages over bag-of-words or LDA: it handles short text, captures semantics, and does not require pre-specifying the number of topics.

Dimensionality Reduction with UMAP: Directly clustering in 768-dimensional space often leads to unstable results. We applied UMAP (*Uniform Manifold Approximation and Projection*) to reduce embeddings to 5–10 dimensions while keeping local neighborhood structure. This ensures that firms describing similar processes (e.g. machining, logistics, or software services) remain close in the reduced space. UMAP also improves HDBSCAN's ability to detect irregular clusters.

Clustering with HDBSCAN (BERTopic): We used HDBSCAN as the core clustering algorithm within the BERTopic framework because it can identify:

- clusters of unequal density and size,
- arbitrarily shaped clusters, and
- the presence of noise points.

Unlike LDA, HDBSCAN does not require specifying the number of topics. A minimum cluster size of 50 was imposed to ensure interpretability. We ended up with 89 clusters and one noise cluster, each corresponding to an industry theme (e.g. electronics manufacturing, apparel, food services, logistics, medical devices, local policy services).

Topic Consolidation into Manufacturing Stages: Because 89 clusters are too granular for economic interpretation, we grouped them into 10 manufacturing-pipeline stages using a large language model. These categories capture upstream, midstream, and downstream roles including component manufacturing, assembly, material processing, logistics, resource extraction, and R&D.

Supervised Learning Model

For the supervised learning component, we fine-tuned a multilingual transformer model, *XLM-RoBERTa-base, a multilingual masked-language model* pretrained on 2.5TB of CommonCrawl text, chosen for its strong performance on Chinese semantic tasks and robust cross-lingual representations.

Since no human-labeled dataset existed, we constructed a training set using LLM-generated pseudo-labels, produced through few-shot prompting aligned to the semantic structure identified by LDA and KMeans.

The input to the model consisted of preprocessed Chinese business descriptions, tokenized using the XLM-R tokenizer with a fixed maximum sequence length.

The tokenized inputs were converted into tensors containing:

- input_ids
- attention_mask
- token_type_ids (unused, as RoBERTa does not use segment embeddings)

The model was fine-tuned using a standard cross-entropy loss with class-weight normalization to account for imbalances in pseudo-labeled classes.

We employed an 80/20 train–validation split, AdamW optimization, linear learning-rate warmup, and early stopping to prevent overfitting. Model evaluation used accuracy and validation loss, achieving ~80% accuracy despite pseudo-label noise. This supervised stage serves as the scalable backbone of the system, enabling high-throughput, consistent classification across the full 600,000-firm dataset.

Web Scraper

The web scraping system implements a stateful pipeline designed to extract structured company information from the MOFCOM Foreign Investment Enterprise database. The overall system combines preprocessing of firm lists, CAPTCHA-protected web requests, structured response parsing, and systematic data storage into an integrated workflow. The final scraper represents a significant redesign of the original version, emphasizing modularity, fault tolerance, and reproducibility. The full processing pipeline consists of four major components: data ingestion and preprocessing, HTTP session and CAPTCHA management, API query execution and response parsing, and data post-processing and export (see Figure 13).

The scraper loads multiple Excel files using pandas.read_excel() and extracts the firm-name column for use in API queries. Names are normalized through whitespace trimming and UTF-8 encoding (implemented via urllib.parse.quote()), ensuring that Chinese characters are processed correctly. To increase the accuracy of the given scraper, we incorporate an additional step in which each firm is assigned an estimated "FIE likelihood" category (e.g., Very High, High, Medium) based on simple rule-based name features (see Figure 14). This classification determines which firms are prioritized during scraping.

Key Classification Features:

1. **Linguistic Patterns:** English names, foreign company suffixes (Ltd., GmbH, etc.).
2. **Geographic Indicators:** References to Hong Kong, Taiwan, or foreign countries.
3. **Organizational Codes:** Unified Social Credit Codes beginning with "91"
4. **Contextual Markers:** Parentheses containing foreign names, "International" keywords.
5. **Domestic Penalties:** Chinese province/city names reduce FIE probability.

The MOFCOM website requires two manually obtained session cookies: JSESSIONID and insert_cookie. These must be extracted from a browser's developer tools and inserted directly into the scraper. The scripts do not automatically obtain new cookies or refresh expired ones. When the server begins rejecting CAPTCHA submissions (detected via repeated decoding failures), the scraper prints "Cookie expired" and terminates the current run, requiring manual cookie replacement before resuming.

Each firm query requires solving a CAPTCHA image downloaded in JPEG format. The scraper applies a multi-stage decoding process:

1. **Image preprocessing:** improves OCR accuracy through contrast enhancement, sharpening, and binary thresholding.
2. **Dual OCR attempts:** runs ddddocr once on the preprocessed image and once on the original JPEG.
3. **Format validation:** checks whether results satisfy expected MOFCOM CAPTCHA constraints (3-8 alphanumeric characters).
4. **Selection heuristic:** chooses the most frequent valid result among attempts; invalid results are discarded.

The scraper retries decoding several times, using time.sleep(1) between attempts. After decoding a firm's CAPTCHA, additional random delays (random.uniform(2, 4)) simulate human behavior before the next firm is processed. These measures reduce rate limiting and improve stability during long runs. It also introduces exponential backoff for failed HTTP requests, implemented with wait_time = min(2 ** attempt, 30), and maintains a counter for consecutive failures; when thresholds are exceeded, the script assumes server-side blocking or expired cookies and stops safely rather than corrupting downstream output. Results are aggregated into a pandas DataFrame and written to Excel in 100-firm chunks. Each chunk produces a file named: firm_info_{start_index}_{end_index}.xlsx. The scraper tracks progress in memory, allowing the user to restart from the last successfully completed firm after updating cookies.

Software and Codebase

We mostly used python for the implementation. We developed preprocessing scripts responsible for boilerplate removal, Jieba segmentation, and stopword curation. Semantic embedding generation was implemented using the sentence-transformers library, while UMAP and HDBSCAN were run through the BERTopic package. Additional tooling included:

- pandas / numpy for data manipulation
- regex for text cleaning and firm classification
- umap-learn for dimensionality reduction
- hdbscan for density-based clustering
- lightgbm, scikit-learn, and PyTorch for downstream supervised evaluation

One practical challenge was ensuring reproducibility across multiple cleaning stages; minor changes in tokenization or stopword design significantly altered cluster structure. Another difficulty involved tuning HDBSCAN to avoid over-fragmentation, which required experimentation with min_cluster_size and UMAP parameters.

Evaluation/Validation/Experiments

We evaluate our multi-stage classification pipeline using quantitative metrics, controlled evaluation setups, baseline comparisons, ablations, and qualitative inspection. Since the system includes components ranging from topic modeling to supervised classification and web scraping, each module is assessed with appropriate methods

Classification Metrics and Setups

- **LDA Topic Modeling**
    - **Perplexity:** Measures held-out likelihood. Lower perplexity indicates a better generative fit.
    - **Topic Coherence (C_v):** Measures semantic similarity of top words. Higher coherence indicates more interpretable topics.
        - Evaluated perplexity and coherence across $\mathbf{k} \in \{2, 4, 5, 6, 7, 8, 10, 15, 20\}$.
        - Performed hyperparameter search over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to maximize coherence.
        - Experimented with removing these legalese words directly in the preprocessing step
- **KMeans Clustering**

- **t-SNE Separation:** Used for visual validation; well-separated clusters indicate meaningful structure.
- Tested the effect of down-weighting legal/regulatory topics to reduce noise and improve compactness.

➢ **Supervised Classification Model**

- **Accuracy:** Measures proportion of correctly predicted labels.
- **Evaluation Loss:** Cross-entropy loss; lower values indicate better generalization.
- **Weighted F1:** Balances precision and recall while accounting for class imbalance.
  - Trained using an **80/20 train–validation split**
  - **Monitored training loss for convergence and overfitting**

➢ **Qualitative Checks**

- **Manual Spot Checks:** randomly sampling a subset of outputs by the group members and Sponsor

Results & Visualisation: Figures 4-11

Error Analysis

Overlapping regions captured hybrid or multifunction firms.

Limitations

Misclassifications primarily arose in clusters where vocabulary overlaps across stages (e.g. "加工" appears in both material processing and assembly contexts). Very short descriptions also tended to fall into the noise cluster due to insufficient semantic information. These limitations show the noisiness of business-scope text.

Web Scraper Evaluation

➢ **Metrics**

- End-to-end scrape coverage used as the primary metric (firms successfully extracted from MOFCOM).
- Baseline accuracy: **2%**; Enhanced scraper accuracy: **14%** across all six files.
- The enhanced scraper extracted **7× more valid records** than the baseline.

➢ **OCR Experimentation**

- **ddddocr (baseline method):** extremely low solvability, consistent with ~2% end-to-end accuracy.
- **Tesseract:** tested as alternative; not reliable enough for automation.
- **Tesseract + OpenCV preprocessing:** slight improvement (contrast, thresholding, sharpening) but still insufficient for automated CAPTCHA solving.
  Overall: OCR remains the dominant limiting factor.

➢ **Manual-Input Validation**

- Manually typing CAPTCHA into search requests yielded **100% success**.
- Verified detail retrieval and JSON field mappings were correct.
  Confirms that system failures stem from CAPTCHA decoding, not API logic.

➢ **Evaluation Setup**

- Direct comparison of baseline vs. enhanced scraper.

- Tested on all **six Excel files** of firm names (~2,000 high-probability FIE candidates per file).
- Processed in **100-company batches** to analyze chunk-level performance.
- Executed entirely on **Google Colab** to ensure consistent computing conditions.
- Each firm allowed **up to 10 attempts**, each with **up to 8 CAPTCHA retries**.

➢ **Results**

- **Baseline:** ~2% accuracy; coverage dropped from ~4% to <1.5% as session cookies expired after ~2 hours.
- **Enhanced scraper:** ~14% average accuracy; stable for **4+ hours** with no degradation.
  - Batch-level accuracy: **5%–25%** per 100-company segment.
  - Filtering firm names by foreign-investment probability further improved success rates.

➢ **Error Analysis**

- Major bottleneck: **CAPTCHA solver accuracy**.
  - CAPTCHA images differ between search API and CAPTCHA endpoint → structural mismatch limiting OCR performance.
- Some failures caused by firms absent from MOFCOM entirely (input limitation).
- Network issues contributed minor, sporadic failures.
- Enhanced scraper still significantly reduced total CAPTCHA-related errors.

Team Member Participation

The project was completed through coordinated contributions from all team members. Ingrid Morales focused on LDA tuning and clustering analysis, while Minjin Li led the preprocessing pipeline and contributed to the LDA component as well. Abiramashree Adaikalam developed the LLM-assisted transformer model as well as the LDA, and Chloe Florit implemented the BERT-based classification model. The web scraper was jointly developed by Viveka Agrawal and Brian Lezama, who each contributed approximately half of the scraper design, implementation, and testing.

Member	Tasks Worked On
Ingrid Morales	LDA tuning (40%), Clustering (100%)
Minjin Li	Preprocessing, LDA (30%)
Abiramashree Adaikalam	LDA(30%), LLM assisted Transformer Model (100%)
Chloe Florit	BERT Classification Model (100%)
Viveka Agrawal	Web Scraper (50%)
Brian Lezama	Web Scraper (50%)

Conclusion

Our project aims to answer a simple but difficult question: what do foreign invested firms in China participate in the production process? The MOFCOM dataset provides near census coverage of Foreign-Invested Enterprises (FIEs), but its free-text business operation descriptions are messy and don't specify the broad production stage labels. By combining unsupervised topic models, embedding based clustering, LLM-based pseudo-labeling, and supervised learning, we were able to translate more than 600k Chinese descriptions into interpretable production stage labels with reasonably high consistency and produce production stage classification pipelines for incoming foreign firms.

Although the web scraper was given, providing it to be functional for mass production showed us the challenges of data collection. We began by doing research into session mechanics, HTTP requests, and OCR tools. Much of the work became about reverse-engineering the website behavior, understanding dependencies and experimenting with OCR libraries. While attempting to web scrape this specific government website we were introduced to such limitations that include: low OCR accuracy, session flows are brittle, and the slightest errors in headers could lead to unknown failures. This experience taught how complex real-world scraping environments can be and how certain constraints can interrupt the entire pipeline.

Despite the scale and richness of our dataset, our analysis is still constrained by several structural limitations of the dataset. The business operations feature is self reported administration text, it's not an objective source for the actual production activities of the firms, and it provides a potential quality issue for such descriptions. About 11% of firms reported a description with less than 10 characters, making it difficult to produce a precise label for its role in the production pipeline.

On the modeling side, our labels and evaluation are based on unsupervised cluster labels and pseudo-labeling through LLM rather than human labeling. This means our results depend heavily on the researchers' interpretation and on the reliability of LLM labeling. In addition, our assumption is that each firm operates in a single production stage, but in the real world many firms participate in multiple stages of the production process. Which our current single label framework cannot fully capture.

One of the main difficulties we faced in this project was working with a Chinese text–based dataset even though most of our team does not speak Chinese. This made every preprocessing and modeling choice more challenging. A second major challenge was that the dataset initially had no ground truth labels, which pushed us toward a fully unsupervised and semi-supervised workflow. Instead of training a classifier on a labeled dataset, we first had to discover structure using LDA and embedding based clustering, then interpret those clusters as meaningful production stages, and only afterward build supervised models on top of these labels.

If we have more time, our first priority would be to build a small human labeled dataset for production stages to provide a group truth labeling set so we could quantify and compare the accuracy of our different models and act as a test set for our supervised model. Looking beyond the immediate scope of this capstone, a direction of future research could be to link our production stage labels to environmental and spatial datasets to study the environmental impact of foreign invested firms in China.

Appendix

Figure 1 Distribution of Business Description Length

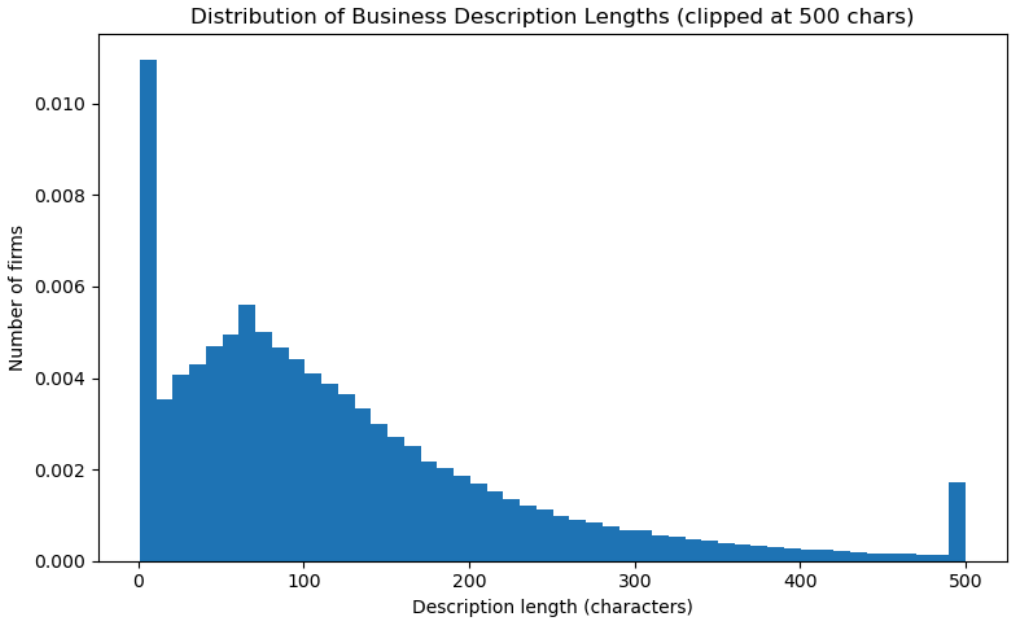


Figure 2: Most frequent business operation

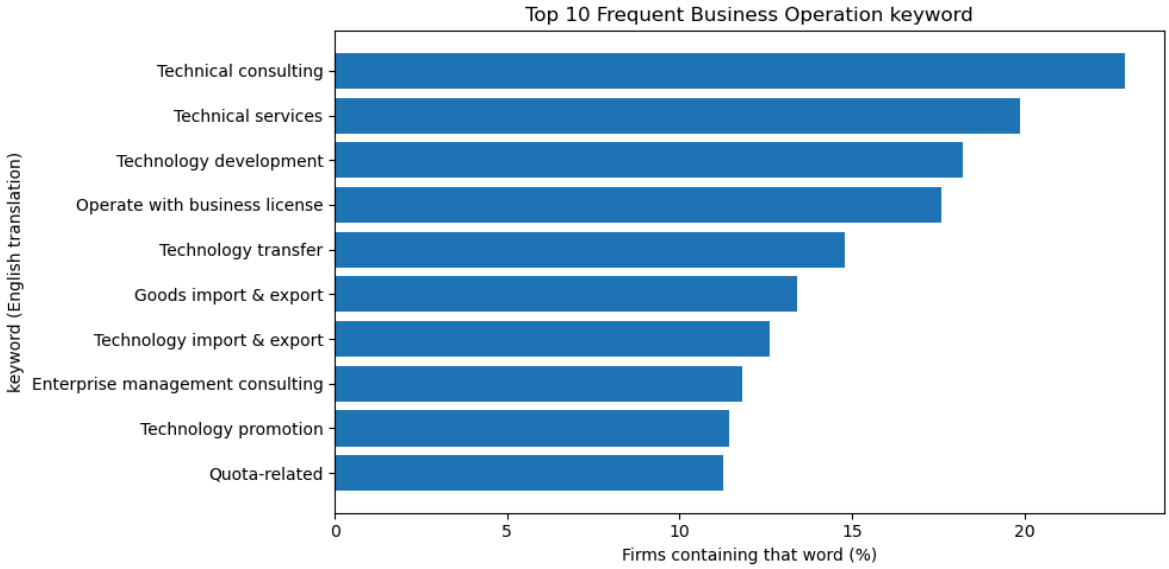


Figure 3: Most frequent translated words

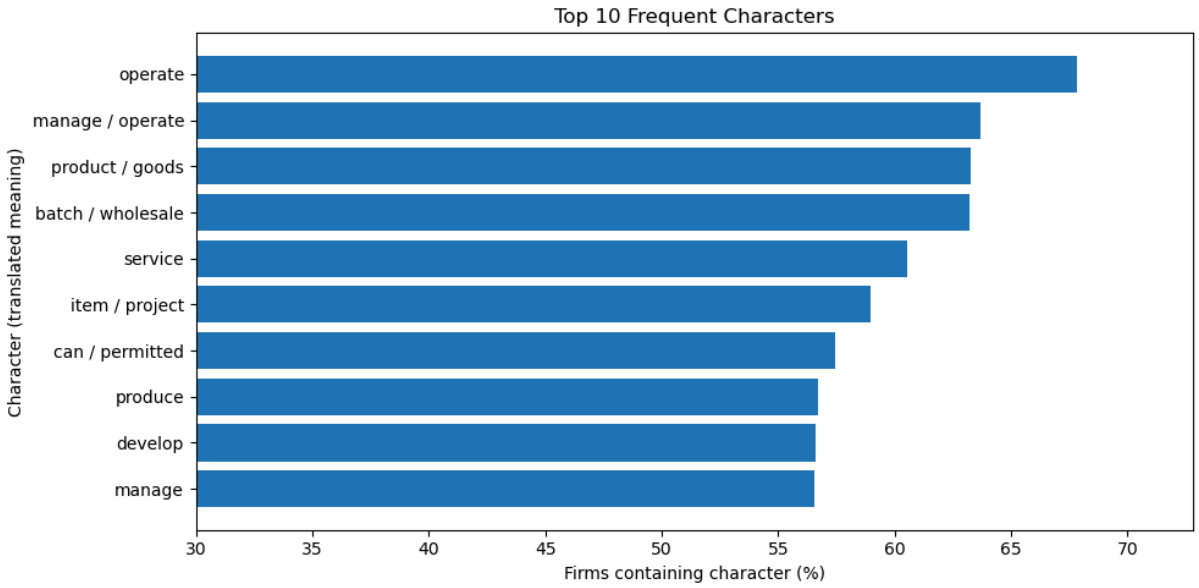


Figure 4: Visualization results of our coherence and perplexity scores. K=6 produced the best results

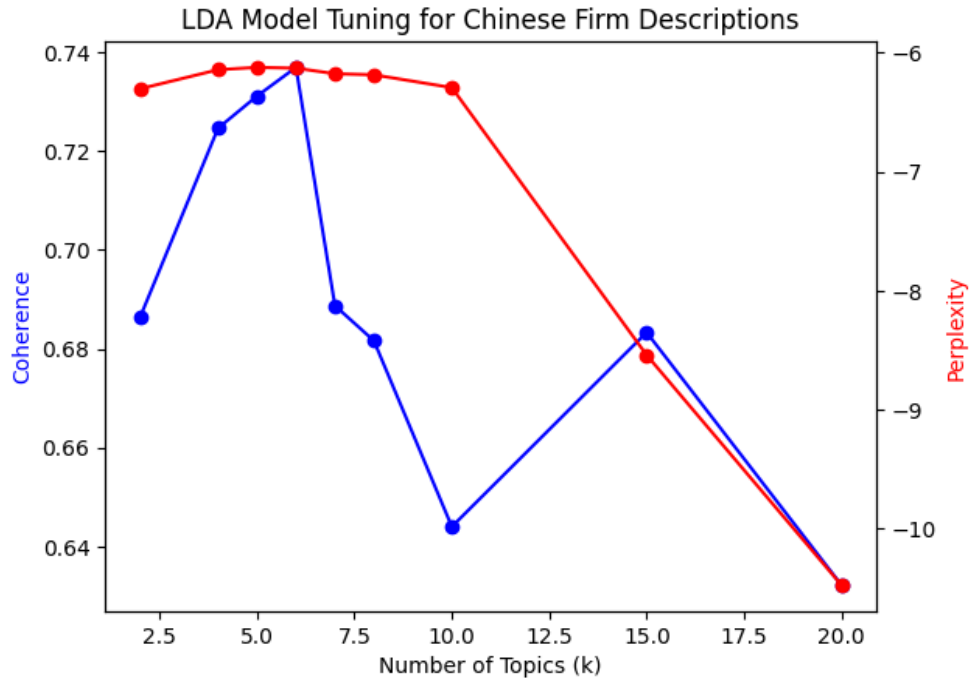


Figure 5: Results of fine tuning α and β

	topics	alpha	beta	coherence	perplexity
0	6	0.01	0.16	0.7152	-6.2455
1	6	0.30	0.16	0.7506	-6.1054
2	6	0.40	0.16	0.7544	-6.1060
3	6	0.40	0.01	0.7559	-24.9852
4	6	0.40	0.60	0.7159	-6.0479
5	6	0.40	0.90	0.7159	-6.0646
6	6	0.50	0.16	0.7390	-6.1144
7	6	0.60	0.16	0.7182	-6.1128

Figure 6: Results of the experiment of downweighting Topic 2 Legalese label

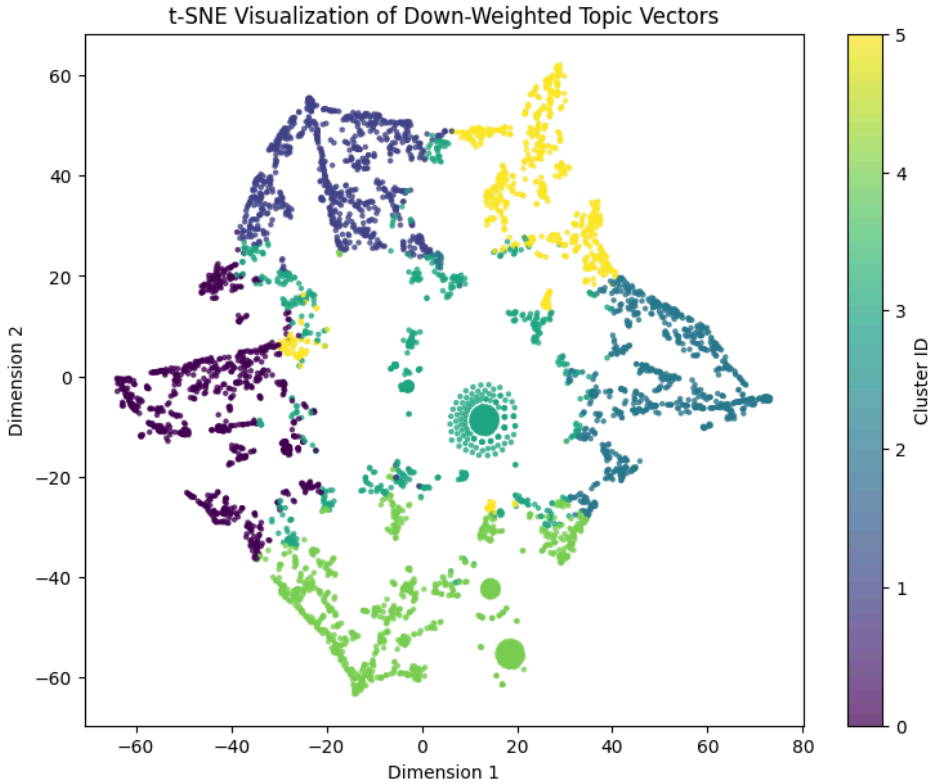
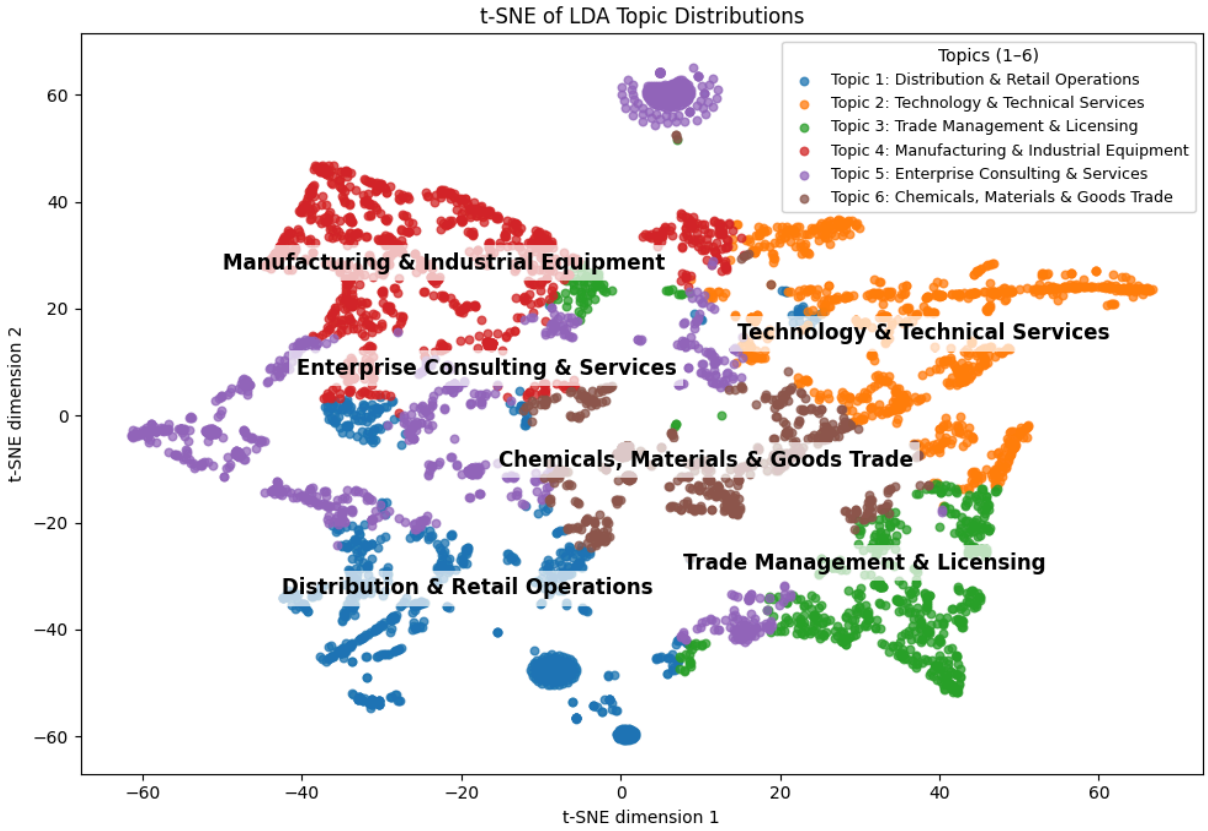
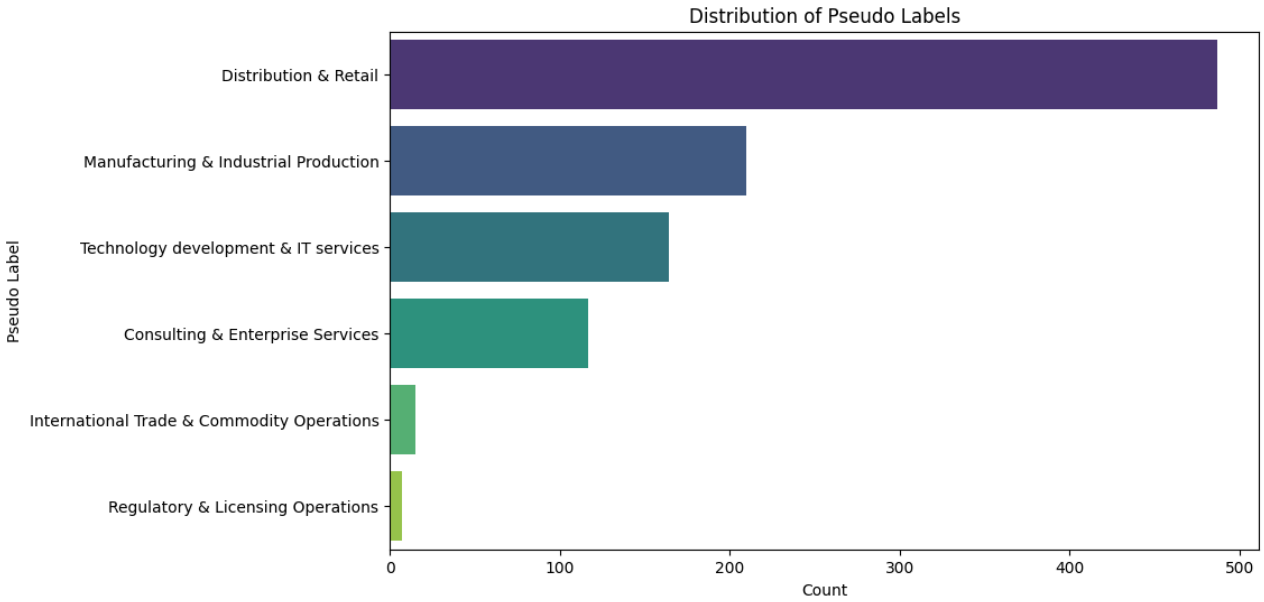


Figure 7



- Separability after removing Topic 2 phrases

Figure 8



- **Training Loss:** 0.24 → 0.39 (smooth convergence)
- **Eval Accuracy: 80%**
- **Eval Loss:** 0.886

Figure 9: Industry Topic Distribution

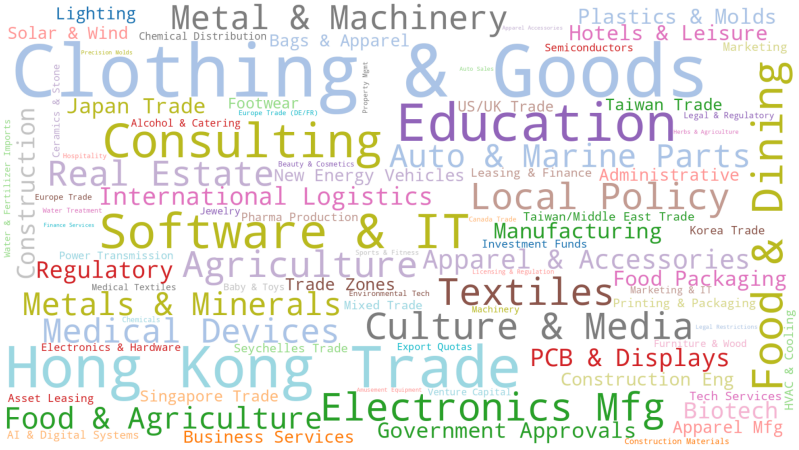


Figure 10: Supervised Models Tested

Model	**Accuracy**	**Weighted-F1**
Logistic Regression	0.63	0.63
LightGBM	0.75	0.74
FastX	0.76	0.76

Figure 11: Results after further removing Topic 2 general phrases

Metric	Before	After stopword removal	Change
Coherence	0.7559	0.7716	↑ Better semantic clarity
Perplexity	–24.99	–26.66	↓ Better model fit

Figure 12: Description of Company-Level Variables Extracted from MOFCOM API

Field (Chinese / Code)	Meaning / Description	Data Type	Size / Format	Dimensionality
公司名称 (search_name)	Firm name entered into the search system	String	5–50 chars	Single value
公司名称 (entp_name)	Official legally registered company name	String	10–100 chars	Single value
统一社会信用代码 / 组织机构代码 (entp_gs_code)	Government-issued unified social credit code	String	18 chars fixed	Single value
状况 (gs_status_name)	Legal registration status (e.g., active, cancelled)	Categorical	~4 Chinese characters	Single category
成立日期 (record date format)	Company establishment date	Date	YYYY-MM-DD	Single value
投资行业 (industryname)	Registered industry classification	Categorical	~50 possible categories	Single category
注册资本 (register_capital)	Company's declared registered capital amount	Numeric	Variable digits	Single value
注册资本单位 (unit_name)	Currency/measurement unit of capital (e.g., RMB, USD)	Categorical	3–4 Chinese characters	Single category
经营范围 (business_scope)	Full business scope description filed with authorities	Text	50–2000 chars	Single text field
地址 (reg_addr)	Official registered business address	String	20–200 chars	Single value
法定代表人 (right_man)	Legal representative (official corporate signatory)	String	2–10 chars	Single value

投资者信息 (investor_info)	Structured list of investors, each with additional details	Structured Text	Pipe-delimited entries	Multi-record (1–10 items)
变更信息 (changes_info)	Historical record of changes to registration (e.g., capital changes, address changes)	Structured Text	Pipe-delimited entries	Multi-record (0–50 items)
年报年度 (year_report)	Years in which annual reports have been submitted	List	4-digit years, pipe-delimited	Multi-value (0–15 years)

Figure 13: System Flow Diagram of the Web Scraper

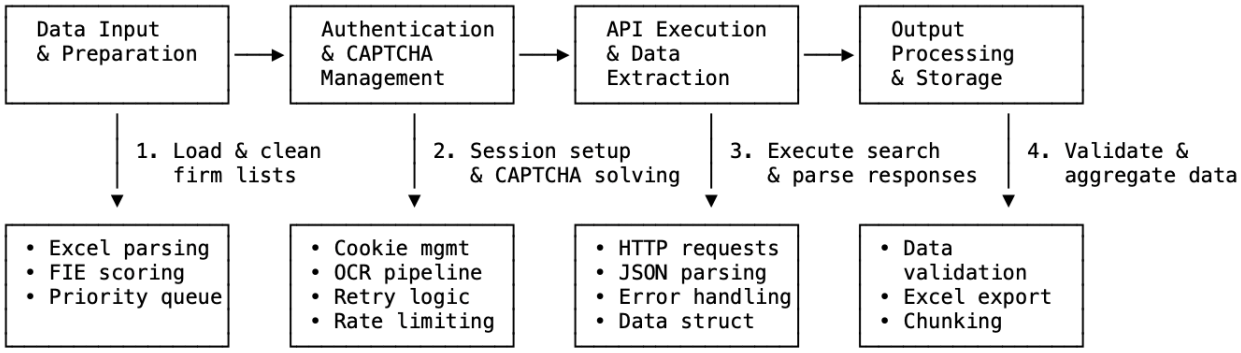


Figure 14: Web Scraper Classification

```
============================================================
FIE PROBABILITY ANALYSIS
============================================================
Total firms analyzed: 20890

Probability Distribution:
  Very High   :   54 firms (  0.3%)
  High        : 10706 firms ( 51.2%)
  Medium      : 7585 firms ( 36.3%)
  Low         : 2183 firms ( 10.4%)
  Very Low    :  362 firms (  1.7%)

============================================================
RECOMMENDATION:
Focus on 'Very High' and 'High' probability firms: 10760 firms
Expected successful captures: ~4304 firms
============================================================
```

This image shows the probability distribution of firms and how many firms from a single Excel file were classified into the different categories.

```
Sample High Probability Firms:
         firm_name  fie_score probability      indicators
91440300MA5HKQYK1X         60   Very High 英文名, 香港, 统一信用代码
91440400MA55UHKB8E         60   Very High 英文名, 香港, 统一信用代码
91310115MADFHKJK6C         60   Very High 英文名, 香港, 统一信用代码
91440300MA5HKE439A         60   Very High 英文名, 香港, 统一信用代码
91331000MA7DHK4N21         60   Very High 英文名, 香港, 统一信用代码
91440300MA5HKJGQ7T         60   Very High 英文名, 香港, 统一信用代码
91440300MA5HK8344T         60   Very High 英文名, 香港, 统一信用代码
91310000MA1HK0H36A         60   Very High 英文名, 香港, 统一信用代码
91210213MA0YW2HKX0         60   Very High 英文名, 香港, 统一信用代码
91440113MACCHKDE4W         60   Very High 英文名, 香港, 统一信用代码


Sample Low Probability Firms:
      firm_name  fie_score probability                           indicators
   天津艺虹印刷发展有限公司        -10    Very Low                    domestic_province
      义乌市金肯贸易商行        -10    Very Low                  domestic_city_start
   上海斗牛士牛排馆有限公司        -10    Very Low                    domestic_province
   上海静湖酒业有限责任公司        -10    Very Low                    domestic_province
 福建省福新房地产开发有限公司        -10    Very Low                     domestic_province
 佛山市达米克医疗用品有限公司        -10    Very Low                   domestic_city_start
    清远市尚霖布艺有限公司        -10    Very Low                  domestic_city_start
   广东华源盛业贸易有限公司        -10    Very Low                    domestic_province
重庆市梁平县宏美达纺织有限公司        -20    Very Low domestic_province, domestic_city_start
 重庆市永川区积流贸易有限公司        -20    Very Low domestic_province, domestic_city_start
```

This image shows some examples of what is labeled as high probability versus low probability in terms of being a foreign invested firm and why. If the firm name is solely in Chinese, it is presumed to be a domestic firm, thereby lowering its probability of being a foreign firm. If the firm name in the file starts with a number, it has a higher chance of being considered a foreign firm.

Web Scraper Output Comparison

Before (Original)

```
1 processing(1, 500, 'E83B6179B3B89D6872ACFFCE594F6E9D', '22558280', '/content/Firm lists/2025.01.04.xlsx', 100)
```

```
Total number of firms to search: 20891
Processing firms:  50%|█████     | 50/100 [04:40<04:40,  5.61s/it]
Catched: 6 7 Did not find: 44

Network error, 1 life burnt. 1 lives remain

Total number of firms to search: 20891
Processing firms:  15%|█         | 15/100 [01:37<09:10,  6.48s/it]
Catched: 3 4 Did not find: 12

Network error, 1 life burnt. 1 lives remain
```

After (Enhanced)

```
============================================================
Processing chunk: 1901 to 2000
============================================================

Total number of firms to search: 10866
Processing firms: 100%|██████████| 100/100 [14:18<00:00,  8.58s/it]
Processing firms: 100%|██████████| 7/7 [00:44<00:00,  6.29s/it]
================================================
Summary:
Successfully captured: 7
Not found: 100
Errors: 0
Success rate: 7.0%
================================================

Saved: ./firm_info_1901_2000.xlsx

============================================================
FINAL SUMMARY:
Total firms captured: 219/2000
Overall Accuracy: 10.95%
Processing complete!
============================================================
```

As can be seen from the original scraper output, it's not easy to understand the output and what it means in terms of how many firms were captured, but the new enhanced scraper output gives more understanding as to what the percentage of coverage is for each chunk of firms as well as the overall accuracy of a single batch of firms (in this case, 2000).

Technical Approach - Data preprocessing

This is an example of the legal phrases and general words from topic 2 that we removed from our cleaning.

```python
legal_phrases = [
  '一般项目',
  '许可项目',
  '除依法须经批准的项目外，凭营业执照依法自主开展经营活动',
  '依法须经批准的项目，经相关部门批准后方可开展经营活动',
  '具体经营项目以相关部门批准文件或许可证件为准'
]
```

```python
topic2_english = set([
    "business", "business items", "Approval", "According to law", "approve",
    "involving", "nation", "access", "special", "prohibit", "license"
])
topic2_chinese = set([
    "经营", "经营项目" , "审批", "依法", "涉及", "批准", "国家", "准入", "特别", "禁止
", "许可"])
```

Multilingual Tokenization

We apply the following pipeline:

1. Character filtering – Retain only Chinese characters and English letters:
    a. re.sub(r'[^\u4e00-\u9fffA-Za-z]', ' ', text)

2. Lowercasing for English normalization

3. Chinese segmentation using Jieba, which handles multi-character production terms (e.g., “机械加工”, “船舶设备”)

4. Stopword removal for both Chinese and English

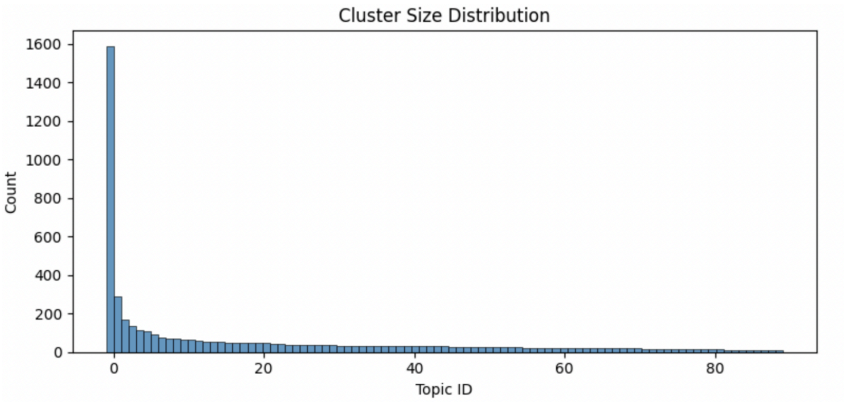
5. Minimum length filtering (tokens longer than 1 character)

Evaluation/Validation/Experiments

Unsupervised Cluster Evaluation

Cluster quality was examined through:

- Cluster size distribution plots, showing a heavy-tailed structure with many small clusters and one large noise cluster.
- Word clouds and top-keyword summaries, confirming that clusters align with recognizable industries (e.g. plastics, apparel, electronics components, logistics).
- Manual inspection, validating that UMAP + HDBSCAN separated firms with distinct industrial roles, even when business descriptions were short.



Supervised Evaluation Using Derived Features

We used the mapped 10-stage manufacturing labels as targets for supervised learning. Three models were evaluated under an 80/20 train-test split.

Bibliography

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Kapadia, Shashank. “Evaluate Topic Models: Latent Dirichlet Allocation (LDA).” Towards Data Science, 24 Dec. 2022, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.