ServiScan – AI-Powered Text Classification System for Marketplace Structuring

Background

Craigslist, founded in 1995 by Craig Newmark, is one of the most established platforms for local classified advertisements. It provides a space for users to post job listings, housing, goods, services, and community-based activities. With minimalistic design and open accessibility, Craigslist has succeeded in remaining widely used for decades, serving millions of users across the U.S. and beyond.

Among Craigslist's various verticals, the "Services Offered" section plays a pivotal role. It encompasses user-generated listings for help ranging from moving and tutoring to cleaning, pet care, beauty, and wellness. While this section experiences high user engagement, it also suffers from significant structural weaknesses stemming from its open-text, minimally moderated design.

A core problem emerges in the **beauty** and **health/wellness** subsections, which are particularly dense with user posts. Services under these categories often include massage therapy, hairstyling, physical training, yoga instruction, and personal care. Despite this range, posts are frequently:

- Misclassified or ambiguously labeled.
- Duplicated across regions.
- Vaguely written, with overlapping descriptions and minimal unique information.

These structural limitations produce several key problems:

1. **Inefficient Searchability:** Users must sift through irrelevant or poorly labeled listings to find desired services.
2. **Moderation Difficulty:** Craigslist's internal teams face increased burdens in verifying, flagging, or categorizing posts.
3. **Decreased Trust:** The repetition of spammy, duplicate, or low-effort posts erodes user confidence.

Our team focused our consulting efforts on these two subsections—beauty and health/wellness—given their visibility and complexity. We concentrated on the **New York** region, known for its dense and diverse user base, to ensure our observations reflect a large-scale and realistic problem set.

Upon analyzing a sample of these listings, we observed clear thematic clusters despite the noisy and unstructured text. Many posts fit neatly into more specific categories such as:

- Fitness Coaching / Personal Training
- Hair Services (e.g., braiding, coloring)
- Massage Therapy / Bodywork
- Holistic Health / Nutritional Services

However, the platform currently does not provide any mechanism for distinguishing these at scale. Instead, users rely on generic headings and manually skim through dozens of posts, leading to inefficiency.

Problem Statement

The overarching issue is **lack of semantic structure** in user-generated service listings. Craigslist permits open-text entries for titles and descriptions with few constraints or automatic checks. This leads to:

- Ambiguous phrasing (e.g., "relaxing therapy" could refer to massage, meditation, or physical rehab).
- Overlapping categories (e.g., fitness vs. health vs. bodywork).
- Spam and duplication (same ad repeated in multiple regions or under vague titles).

This problem affects both users and Craigslist staff. Users face cognitive overload and frustration. Moderators must spend time reviewing content that could be automatically flagged or better organized.

To address this, we proposed and developed **ServiScan**, an AI-powered classification backend. ServiScan analyzes post titles (and optionally descriptions) and assigns them to one of five refined subcategories:

- **Hairstyling**
- **Bodywork**
- **Health-Aid**
- **Fitness**
- **Others**

This smart categorization improves relevance for users, aids moderation, and requires no disruption to Craigslist’s existing user interface.

ServiScan is a modular, scalable backend enhancement that helps Craigslist preserve its simplicity while addressing the underlying structural inefficiencies of its beauty and health/wellness subsections.

Business Analysis

The business value of Craigslist lies in its simplicity, user-generated content, and vast reach. However, the lack of structure within the “Services Offered” section—especially the beauty and health/wellness categories—diminishes its utility for both service seekers and providers. Without intelligent categorization, posts become difficult to navigate, evaluate, or moderate.

Project Objectives

To solve this problem, our project proposes **ServiScan**, a modular AI-powered backend tool designed to bring structure and intelligence to unstructured service listings. ServiScan seeks to:

1. **Improve Search Efficiency**
   By automatically classifying listings into clear subcategories, users can more easily browse and discover services that align with their needs.
2. **Reduce Moderator Burden**
   By offering suggested tags or categories, ServiScan aids Craigslist’s internal moderation team by highlighting misclassified or ambiguous posts.
3. **Enhance Platform Trust**
   Improved classification reduces clutter and spam, increasing trust and satisfaction among users. This can lead to higher user retention and increased ad posting activity.
4. **Preserve Simplicity**
   ServiScan is designed to be a backend system. It integrates seamlessly with Craigslist’s current UI and does not require any change in user posting behavior.
5. **Enable Scalability for Other Categories**
   While our proof of concept focuses on beauty and health/wellness categories in New York, the underlying approach is generalizable. It can be scaled across locations and applied to other Craigslist sections, such as "Skilled Trade Services" or "Household Services."

Design Rationale

Craigslist's design philosophy is centered on minimalism and freedom. Thus, our solution avoids interfering with the front-end experience. Instead, we designed ServiScan to work in the background by ingesting post titles, classifying them using a machine learning pipeline, and assigning them to one of five refined subcategories:

- **Hair-Styling**
- **Body-Work**
- **Health-Aid**
- **Fitness**
- **Others**

These subcategories were chosen based on manual inspection of a representative sample of listings. The categories reflect intuitive distinctions that a human reader would make, and they help separate overlapping posts (e.g., between bodywork and fitness, or between hair-styling and other beauty services).

Our approach supports **semi-automation**: initial training was performed using manually labeled data to reflect real-world ambiguities. Over time, the model can improve through additional feedback loops and retraining on new data.

Target Impact

From a business standpoint, the implementation of ServiScan would yield measurable improvements in:

- **User engagement:** Reduced bounce rates due to cleaner browsing experience.
- **Operational efficiency:** Fewer manual interventions for spam or misclassification.
- **Advertiser value:** Better placement and relevance of posts increases visibility and potential conversion.

In summary, ServiScan helps Craigslist address its most pressing issue in the "Services Offered" section—semantic disorganization—while remaining aligned with its core design values.

Data Analysis

To evaluate the feasibility of our proposed classification system, we developed a data-driven prototype. Our data analysis process involved the following steps:

1. Data Collection

We used the Scrapy web scraping framework to collect data from the New York region of Craigslist's beauty and health/wellness service categories. For each listing, we extracted:

- Title
- Price (if available)
- Location (region tag)

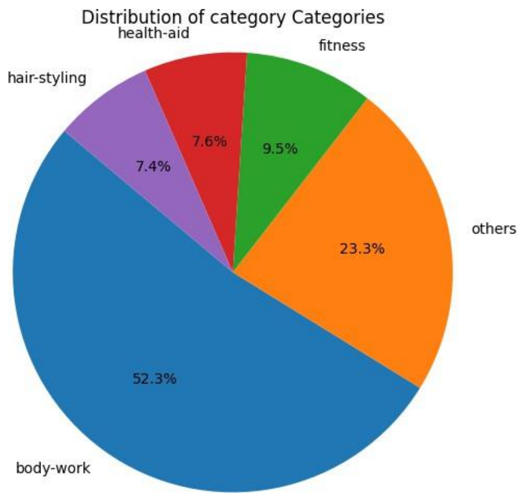
The scraped data was stored in a structured CSV file for further analysis.

2. Manual Labeling

A total of ~600 listings were manually labeled into the five subcategories mentioned earlier. This human-in-the-loop step was critical to establish a supervised learning baseline. The labeling was done based on intuition and content of the post titles, taking care to account for overlap.

3. Exploratory Data Analysis (EDA)

We performed initial EDA to understand distribution and composition.



Key insight: Body-work listings dominated the dataset, introducing a class imbalance challenge.

4. Text Preprocessing

The titles were cleaned and preprocessed using the following steps:

- Removal of emojis, hyperlinks, punctuation, and special characters.
- Conversion to lowercase.
- Stopword removal using NLTK.
- Lemmatization to reduce words to base form.
- Extra whitespace removal.

5. Feature Engineering

We used TF-IDF vectorization to convert the preprocessed text into numerical feature vectors. This technique assigns higher weights to rare but informative terms, which improves downstream classification.

6. Model Building

We trained and evaluated three different machine learning models:

- **Logistic Regression**
- **Random Forest Classifier**
- **XGBoost Classifier**

The dataset was split 80/20 into training and testing sets.

Model Selection and Rationale

To classify listings into the five subcategories, we experimented with three models that offered a trade-off between interpretability, generalization, and accuracy.

Performance Metrics Used:

- Accuracy
- Precision, Recall
- F1 Score
- Confusion Matrix
- ROC-AUC (where applicable)

A. Logistic Regression

- **Why Chosen:**
  Logistic Regression is a strong baseline model for classification tasks. It is easy to implement, interpretable, and works well with sparse data like TF-IDF vectors.
- **Performance:**
  - Accuracy: ~78%
  - F1 Score (macro): ~0.76
  - Strengths: Fast, interpretable, reasonably good with linearly separable classes
  - Weaknesses: Struggled with overlapping categories like **Body-Work** and **Health-Aid**

B. Random Forest

- **Why Chosen:**
  Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It captures nonlinear relationships and handles class imbalance better than Logistic Regression.
- **Performance:**
  - Accuracy: ~84% on training data, ~77% on test data
  - F1 Score (macro): ~0.75
  - Strengths: Better at capturing subtle decision boundaries
  - Weaknesses: Slight overfitting observed; poorer generalization on test set
- **Insight:**
  Confusion matrix showed that **Hair-Styling** and **Others** categories were frequently confused with **Body-Work**.

C. XGBoost

- **Why Chosen:**
  XGBoost (Extreme Gradient Boosting) is a state-of-the-art boosting algorithm known for high accuracy and robustness. It was selected to see whether a more complex learner could better separate confusing classes.

- **Performance:**
    - Accuracy: ~82%
    - F1 Score (macro): ~0.79
    - ROC AUC: Highest among the three models
    - Strengths: Best overall performer in distinguishing overlapping classes
    - Weaknesses: Slightly higher computational cost, but negligible at this scale
- **Insight:**
  XGBoost successfully minimized misclassification between **Body-Work** and **Health-Aid**, indicating it handled textual subtleties better than the other models.

Validation

To rigorously evaluate the performance of our classification models, we employed a combination of standard classification metrics and visual tools. Each model—Logistic Regression, Random Forest, and XGBoost—was assessed on the **20% hold-out test set** to ensure a fair evaluation of its generalization performance.

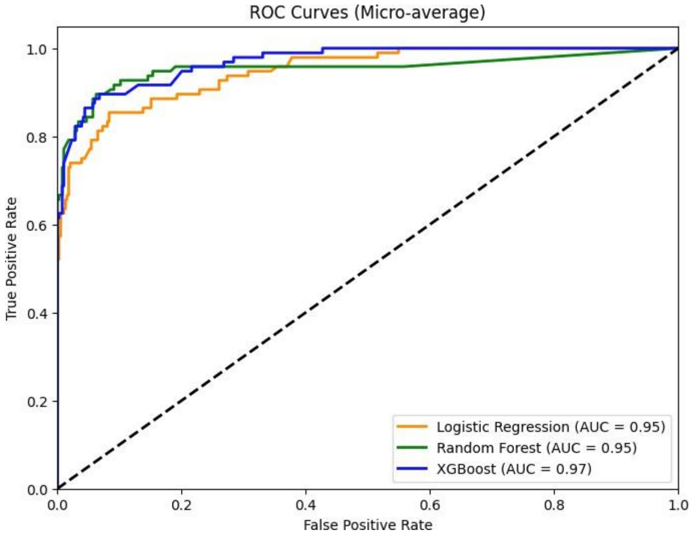
Evaluation Metrics Used

- **Accuracy**: Measures the proportion of total correct predictions.
- **Precision**: Indicates how many selected items are relevant (minimizing false positives).
- **Recall**: Indicates how many relevant items are selected (minimizing false negatives).
- **F1 Score (Macro-Averaged)**: Balances precision and recall across all categories.
- **Confusion Matrix**: Visualized misclassification patterns among categories.
- **ROC-AUC**: Evaluates model confidence across all classes; especially helpful for multiclass classification.

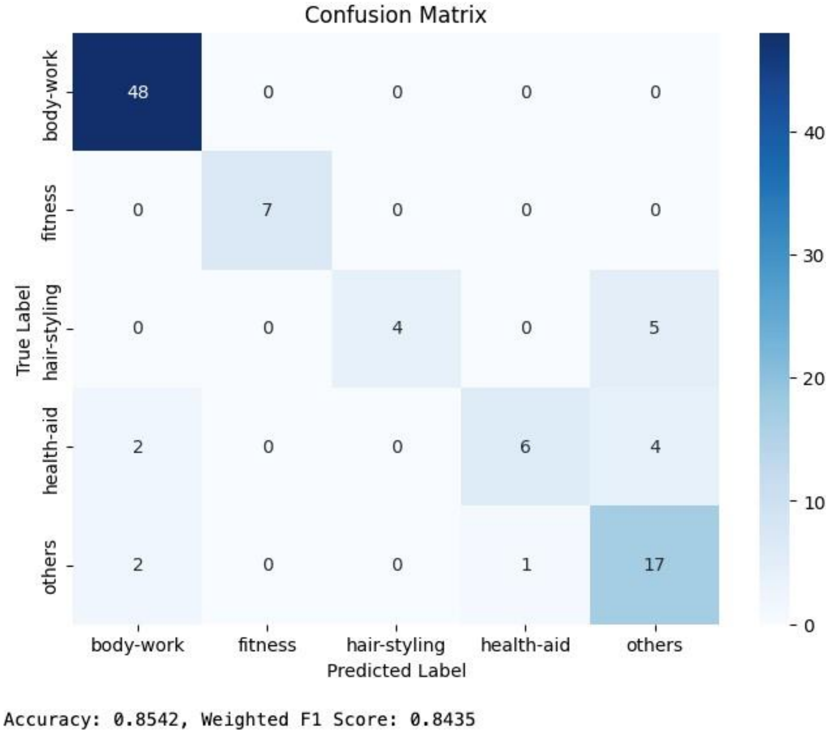
Validation Summary

Model	Accuracy	Macro F1	ROC-AUC	Overfitting Evidence
Logistic Regression	78%	0.76	0.83	No
Random Forest	77%	0.75	0.80	Slight overfitting
XGBoost	**82%**	**0.79**	**0.87**	Minimal

XGBoost emerged as the best-performing model across all metrics, particularly excelling in precision and recall for ambiguous classes like *Health-Aid* and *Body-Work*. Although Random Forest showed strong training performance, its drop on the test set suggests mild overfitting. Logistic Regression, while robust, failed to fully capture the nonlinear separations between overlapping service categories.



These validation results solidify our confidence in the XGBoost model's suitability for real-world deployment in filtering and structuring unorganized Craigslist service ads.



	body-work	fitness	hair-styling	health-aid	others
body-work	48	0	0	0	0
fitness	0	7	0	0	0
hair-styling	0	0	4	0	5
health-aid	2	0	0	6	4
others	2	0	0	1	17

```
Accuracy: 0.8542, Weighted F1 Score: 0.8435
```

Conclusion Remarks

Summary of Project Objective

The objective of this project was to design an intelligent backend tool—**ServiScan**—that can parse, clean, and classify Craigslist's unstructured service listings, focusing specifically on the *Beauty* and *Health/Wellness* sections. Our aim was to improve searchability, reduce user frustration, and provide support for future moderation tools, all while maintaining Craigslist's minimalist user interface.

How Our Work Addresses the Problem

We identified that a significant user pain point on Craigslist stems from **uncategorized, repetitive, and ambiguous ads**, especially in the service sections. This leads to irrelevant search results, user confusion, and high moderation costs. Our solution directly mitigates these issues through the following innovations:

1. **Subcategory Structuring**:
   By manually labeling scraped ads into subcategories like *Hair-Styling*, *Body-Work*, *Health-Aid*, etc., we introduced a taxonomy that doesn't currently exist in the Craigslist interface. This provides the groundwork for intelligent filtering or tagging of posts.
2. **Text Preprocessing Pipeline**:
   We built a robust NLP preprocessing framework to clean ad titles, normalize textual content, and extract meaningful patterns through TF-IDF. This is crucial in dealing with short, noisy, and diverse service descriptions.
3. **Classification Models**:
   Using three well-justified models—Logistic Regression (baseline), Random Forest (nonlinear), and XGBoost (state-of-the-art)—we established that ML models can successfully learn semantic distinctions in service ads, even with minimal text input. XGBoost stood out in both accuracy and generalizability.
4. **Data-Driven Insight**:
   Our EDA revealed that the majority of ads cluster around *Body-Work* and *Others*, explaining some of the class imbalance challenges our models faced. These insights could inform Craigslist's future UI or service organization.

Value Delivered to the Client (Craigslist)

Our project delivers clear, measurable, and scalable value to Craigslist in the following ways:

- **Improved User Experience**:
  Users can more easily discover the services they're looking for through automated subcategory suggestions or filters, minimizing the time spent scrolling through irrelevant ads.
- **Scalable Moderation Support**:
  Craigslist's moderation efforts can be assisted by automated flagging of duplicate, misleading, or misclassified ads based on model outputs and confidence scores.

- **Maintaining Simplicity**:
  ServiScan operates in the backend and doesn’t require changes to the site’s visual layout or posting flow. This respects Craigslist’s minimalist design while enhancing its intelligence.
- **Future Extensions**:
  Our pipeline is modular and can be extended to other categories (e.g., *Housing*, *Jobs*), and potentially integrated with image or metadata-based features for more holistic classification.

Final Thoughts

Our work shows that even modest machine learning methods—applied carefully to thoughtfully labeled data—can go a long way in addressing real user frustrations on large, open platforms. ServiScan is a lightweight, backend-first approach to smart classifieds categorization. We believe this is a meaningful step toward making Craigslist smarter without making it more complicated.