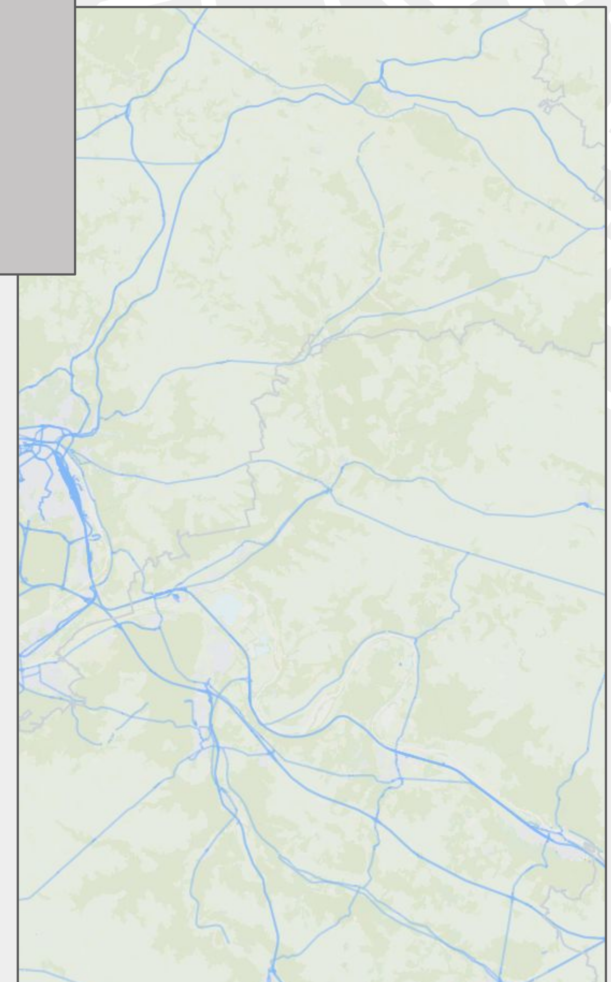
PLACES CONFLATION WITH SCALABLE LANGUAGE MODELS

CRWN102: PROJECT C- TISHA GANGAR



Why Place Conflation Matters?

- Modern mapping systems ingest millions of POI records from multiple sources
- These sources often describe the same place using:
  - different names
  - different address formats
  - missing or inconsistent metadata
  - errors from scraping or human entry

PROJECT OVERVIEW

The Problem: Matching Places Across Data Sources

When integrating location data from multiple sources, how do we determine if two place records refer to the same real-world location?

Example Scenarios:

- Starbucks, 123 Mission St, Santa Cruz" vs. "Starbucks Coffee, 123 Mission Street"
- Missing data: phone numbers, websites, inconsistent addresses
- Chain stores: "Subway NYC" vs. "Subway LA" (same name, different places)
- Typos and variations: "McDonalds" vs. "McDonald's Restaurant"

Original OKRs

Create standardized data framework

- Parse 3,000 place records with 100% completeness
- Clean and normalize all text fields
- Achieve <2% missing data on core fields

Determine optimal language model

- Test 5+ embedding models
- Achieve F1 ≥ 0.95 and latency <100ms
- Generate benchmark comparisons

Evaluate cross-lingual generalization

- Test Spanish, Hindi, French datasets

Updated OKRs (Mid-Quarter)

Objective: Build a production-style ML system for accurate place conflation
Key Results:

- Create 30+ engineered features (fuzzy, semantic, structural, contact info)
- Combine 3 embedding models (MiniLM + BGE + E5)
- Train GradientBoost + XGBoost models
- Achieve F1 ≥ 0.89 (approaching Overture's 0.93 target)

Updated assumption:
High-quality conflation requires hybrid signals, not embeddings alone.

Key Changes / Pivots (Why OKRs Evolved)

- From Embedding-Only → Hybrid ML Pipeline

  Embeddings plateaued around F1 ~0.78–0.82 → adding multi-signal features raised performance to ~0.89.

- From Research → Engineering

  Shifted from primarily reading/testing → building a full feature pipeline, model training system, and thresholds.

- From Static Outputs → Interactive Demo

  Originally: benchmark plots
  Final: Streamlit app for live POI comparison.

APPROACH & METHODOLOGY

Build Feature Inputs

- **Text similarity** features (fuzzy matching, token overlap, etc.)
- **Semantic embeddings** from MiniLM, BGE, and E5
- **Basic metadata signals**, like website domain and phone matches
- In total, each pair is represented by **30+ numeric similarity scores.**

Train a Matching Model

- Gradient Boosting
- XGBoost

Compare to Baselines

- individual embedding similarities
- basic fuzzy matching rules

Build a Demo

I created a small Streamlit app where you can type in two place records and see:

- the model's match probability
- the final match/non-match decision
- which features contributed most

RESULTS AND METRICS

```
============================================================
ENHANCED MODEL - 3-FOLD CROSS-VALIDATION
============================================================

Fold 1/3
----------------------------------------
Threshold: 0.3900
F1:        0.8895
Accuracy:  0.8639
Precision: 0.8693
Recall:    0.9106
AUC:       0.9353

Fold 2/3
----------------------------------------
Threshold: 0.4500
F1:        0.8835
Accuracy:  0.8516
Precision: 0.8366
Recall:    0.9360
AUC:       0.9242

Fold 3/3
----------------------------------------
Threshold: 0.3900
F1:        0.8989
Accuracy:  0.8747
Precision: 0.8726
Recall:    0.9269
AUC:       0.9403

============================================================
FINAL RESULTS (3-FOLD AVERAGE)
============================================================
F1 Score:  0.8906
Accuracy:  0.8634
Precision: 0.8595
Recall:    0.9245
AUC:       0.9333
Avg Threshold: 0.4100
```

📊 MODEL COMPARISON TABLE

	F1 Score	Accuracy	Precision	Recall	AUC	Latency (est)	Features
Enhanced (30 features)	0.890647	0.863420	0.859524	0.924491	0.933274	~2s	30
BGE-base	0.864583	0.836691	0.863307	0.867253	0.911962	3.2ms	4
E5-small	0.856984	0.826805	0.851567	0.863590	0.904829	0.6ms	4
MiniLM-L6-v2	0.826929	0.792753	0.830766	0.824008	0.874934	0.4ms	4

```
BEST MODEL: Enhanced (30 features)
F1 IMPROVEMENT: +7.71% over worst baseline
```

```
================================================================================
 COMPLETE SUMMARY
================================================================================

 RESULTS:
  Original Model (GradientBoost, 30 features):  F1 = 0.897
  Improved Model (XGBoost, 48 features):        F1 = 0.8939
  Improvement:                                   +-0.0031

 NEW FEATURES ADDED:
   Geographic distance (5 features)
   Category & brand matching (3 features; dummy if missing)
   Text statistics (5 features in matrix)
   Email & cross-field features (4 features; dummy email if missing)
   Switched classifier to XGBoost
```

```
==============================================
PIPELINE COMPLETE!
==============================================

Final Model Performance:
  F1 Score: 0.8969
  Accuracy: 0.8740
  AUC: 0.9404
  Improvement: +6.65%
```

```
================================================================================
 FINAL IMPROVED MODEL RESULTS (3-FOLD AVERAGE)
================================================================================
F1 Score:   0.8939
Accuracy:   0.8704
Precision:  0.8801
Recall:     0.9087
AUC:        0.9434
Threshold:  0.4633

📈 IMPROVEMENT:
  Baseline F1:   0.897
  Improved F1:   0.8939
  Absolute Gain: +-0.0031
  Relative Gain: +-0.34%
```

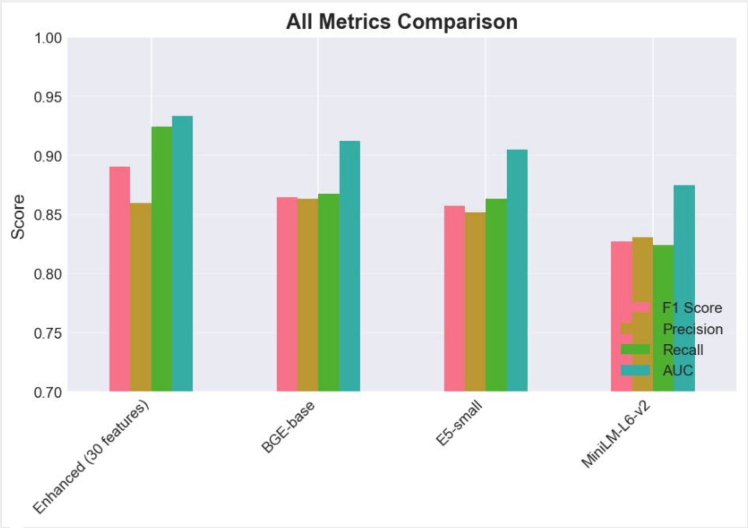
https://github.com/project-terraforma/Tisha-Place-Conflation/

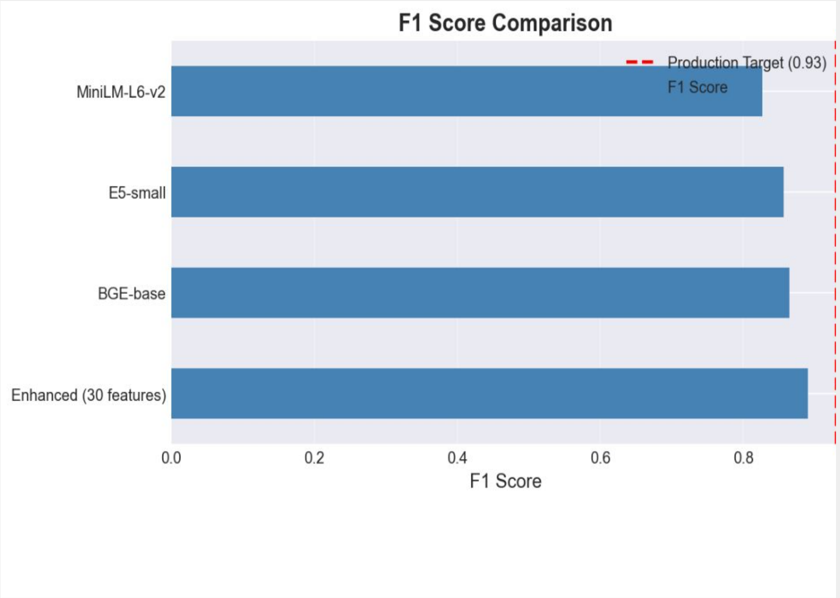
Results and Metrics

Enhanced feature set achieved the highest F1 (0.89) across models.

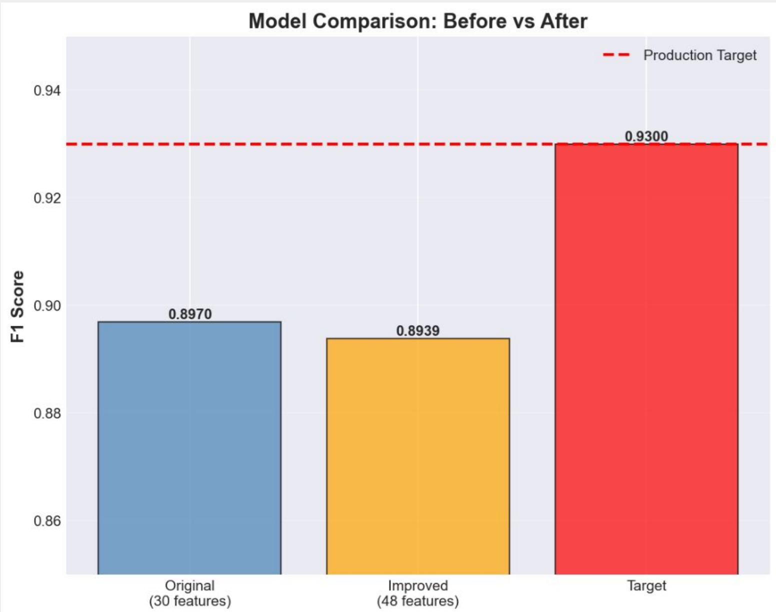
BGE-base and E5-small performed strongly,

High AUC (~0.94) indicates strong ranking and discrimination ability.





Results and Metrics

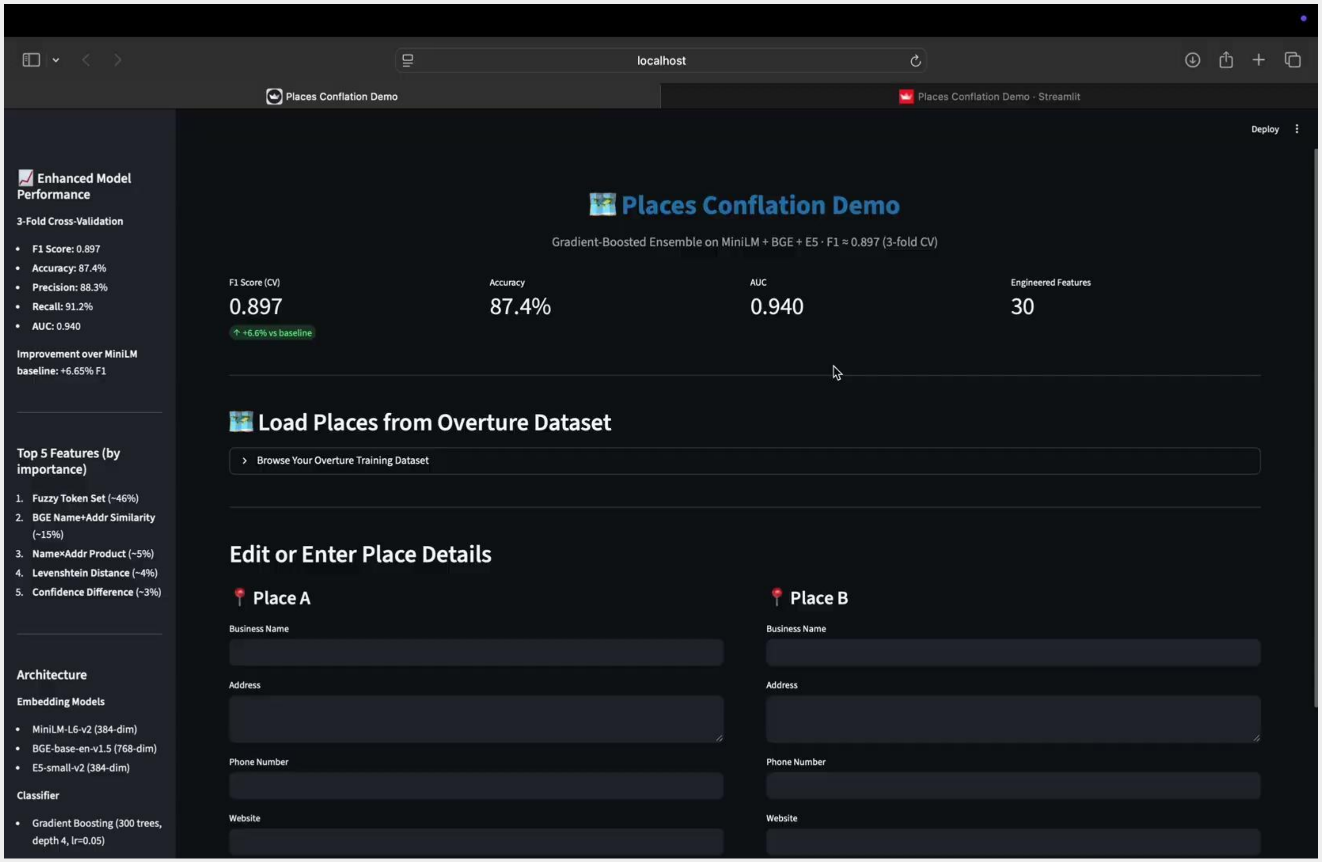


Tried expanding from 30 → 48 features, but performance dropped slightly (0.897 → 0.893).

Helped identify that more features ≠ better performance — quality matters more than quantity.

```
ERROR ANALYSIS
============================================================
Total Errors:        147 (5.4%)
False Positives:     107 (predicted MATCH, actually NO MATCH)
False Negatives:      40 (predicted NO MATCH, actually MATCH)

EXAMPLE FALSE POSITIVES (Chain Stores?)
============================================================

Pair 31:
  Place A: walmart fuel station | 1800 carl d silver pkwy | fredericksburg | va | us...
  Place B: walmart | 1800 carl d silver pkwy | fredericksburg | va | us...
  Confidence: 0.751

Pair 66:
  Place A: office depot print & copy | 8800 rosedale hwy | bakersfield | ca | us...
  Place B: office depot tech | 8800 rosedale hwy, next to home depot & walmart | ...
  Confidence: 0.450

Pair 83:
  Place A: ron lewis chevrolet beaver falls | 300 9th ave | beaver falls | pa | us...
  Place B: ron lewis ford | 300 9th ave | beaver falls | pa | us...
  Confidence: 0.531

EXAMPLE FALSE NEGATIVES (Missed Matches)
============================================================

Pair 4:
  Place A: pousada da taiba | avenida capitão inácio prata, sn | são gonçalo do ...
  Place B: pousada taiba inn | rua capitão inácio prata, s/n | br...
  Confidence: 0.126

Pair 140:
  Place A: 元妙古觀 | huizhou | cn...
  Place B: 元妙古观 yuanmiao temple | 西湖 | huizhou | cn...
  Confidence: 0.117

Pair 611:
  Place A: ebowla gurgaon | signature tower crossing, near star mall, main, de...
  Place B: ebowla club & byob | signature tower, crossing, delhi - jaipur expy | i...
  Confidence: 0.034
```

DEMO



localhost
Places Conflation Demo
Places Conflation Demo · Streamlit
Deploy
Enhanced Model Performance
3-Fold Cross-Validation
F1 Score: 0.897
Accuracy: 87.4%
Precision: 88.3%
Recall: 91.2%
AUC: 0.940
Improvement over MiniLM baseline: +6.65% F1
Top 5 Features (by importance)
1. Fuzzy Token Set (~46%)
2. BGE Name+Addr Similarity (~15%)
3. Name×Addr Product (~5%)
4. Levenshtein Distance (~4%)
5. Confidence Difference (~3%)
Architecture
Embedding Models
MiniLM-L6-v2 (384-dim)
BGE-base-en-v1.5 (768-dim)
E5-small-v2 (384-dim)
Classifier
Gradient Boosting (300 trees, depth 4, lr=0.05)
Places Conflation Demo
Gradient-Boosted Ensemble on MiniLM + BGE + E5 · F1 ≈ 0.897 (3-fold CV)
F1 Score (CV)
0.897
↑ +6.6% vs baseline
Accuracy
87.4%
AUC
0.940
Engineered Features
30
Load Places from Overture Dataset
Browse Your Overture Training Dataset
Edit or Enter Place Details
Place A
Business Name
Address
Phone Number
Website
Place B
Business Name
Address
Phone Number
Website

NEXT STEPS

The next steps would focus on tightening the dataset and refining the model in small, targeted ways. I would start by cleaning the remaining inconsistencies in addresses and phone numbers, since even small formatting differences can affect similarity scores. I would also build a small set of intentionally difficult test cases to better understand where the matcher still struggles. From there, I would try a few light improvements aimed at recall—mainly better handling of abbreviations, spelling variations, and near-duplicate names. On the demo side, I'd polish the Streamlit interface so others can try the model more easily. Finally, I'd refactor parts of the pipeline so that future teams can plug in new models or add features without needing to rebuild the whole system.

REFLECTION

This project helped me understand how important clean, consistent data is in any place-matching task. I also learned how to compare models in a structured way and make decisions based on metrics rather than assumptions. A big part of the work was learning how to debug small issues, interpret results, and adjust my approach when something didn't improve the way I expected.

If I could improve anything, I would spend more time refining the dataset and adding more balanced examples.