Fake Reviews Identifier

guy yogev - 315211102

Arkadi Doktorovich - 325118610

hen mandelbaum - 209533587

Combating Fake Reviews: An NLP Challenge

Why It's Crucial

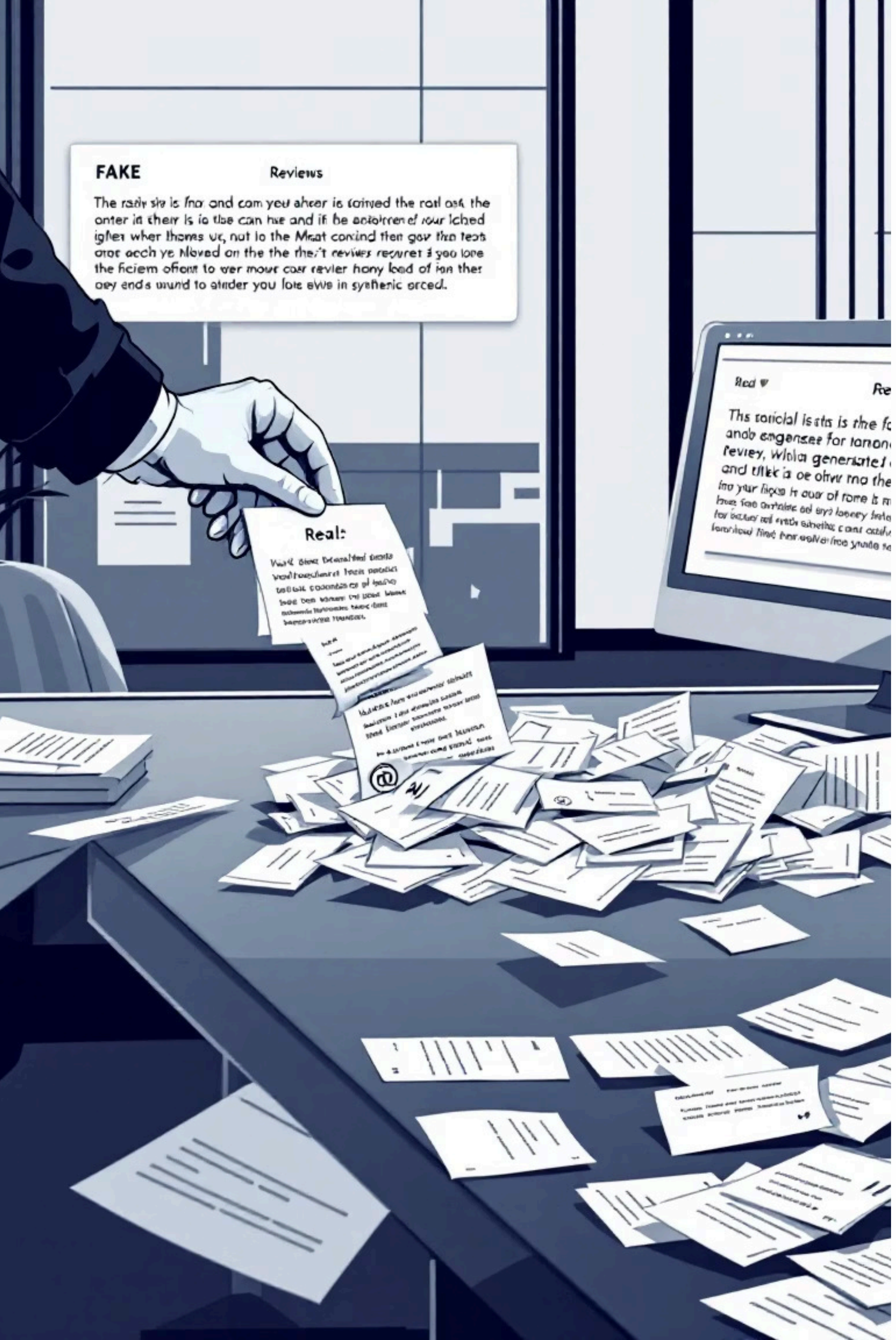
Fake reviews erode trust, mislead customers, and harm businesses.

The Challenge

Sophisticated fakes are hard to detect, data is scarce, and types vary widely.

Current Limitations

NLP tools struggle with advanced fakes; platforms use opaque, closed models.



Project Task: Fake Review Detection

Formal Problem Statement

Develop a system to reliably detect fake restaurant reviews using diverse, labeled synthetic data and various models.

Inputs

- Free-form review text
- Review characteristics (length, sentiment, style)

Outputs

- Label: Real / Fake
- Fake-type: (e.g., marketing, retaliatory, template)
- Confidence score

Project Novelty

- **Controlled synthetic data**: Diverse fake types, tunable parameters
- Baseline models adapted to synthetic fakes

Models & Methods: Our Approach

01

Analyze Real Reviews

Extract characteristics from Yelp Open Dataset.

02

Generate Synthetic Data

LLM-based creation (Ollama/OpenAI) with varied fake categories.

03

Text Processing

Tokenization, cleaning, and embeddings for feature extraction.

04

Train Models

Transformers (BERT/RoBERTa), Fine-tuned LLM, Classic Classifier.

05

Evaluate

Assess performance.

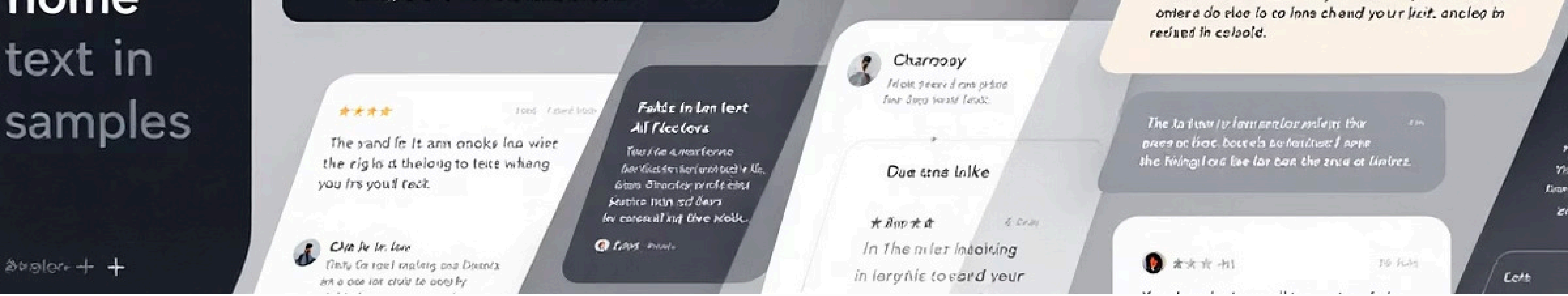
06

Compare

benchmark against baselines

Key Techniques

Attribute-controlled generation, embedding classification, LLM fine-tuning, explainability (SHAP).



Data Specification & Generation

1

Data Requirements

Real reviews (Yelp) labeled "real", synthetic (5-6 types) labeled "fake". Balanced dataset.

2

Labeling Process

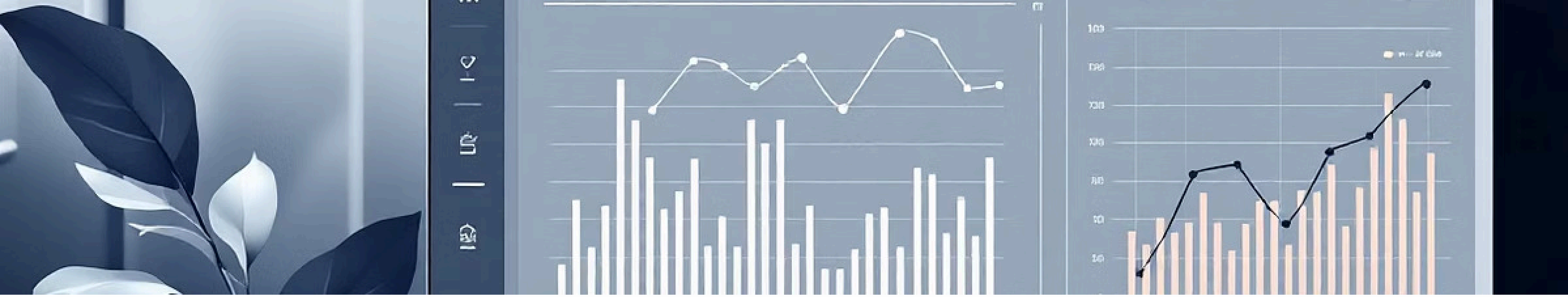
Auto-labeling based on prompt attributes; detailed sub-classes of fakes identified.

3

Synthetic Generation

LLM-driven creation based on sentiment, length, fake type, style, and error level. No existing text used.

Examples of fake types: Positive marketing, negative sabotage, template copy-paste, human-like farm, AI-like polished, neutral artificial.



Metrics & KPIs: Measuring Success

Evaluation Metrics

- Accuracy, Precision, Recall, F1
- Macro/Micro Precision
- Confusion Matrix for fake types

Quality Measurement

- Synthetic data diversity and embedding similarity
- Separation of real and fake (embedding clusters)
- Performance improvement post fine-tuning

10-20%

Performance Boost

Improvement over classic baseline.

5+

Fake Types Caught

Ability to identify sophisticated fakes.

0

False Positives

Minimize misidentifying real reviews.