MKSSS's Cummins College of Engineering for Women , Pune

Named Entity Recognition (NER) Tool

Subject – AI/ML

Batch – B1

Presentated By :

UCE2023508 Arpita Barjibhe

UCE2023517 Shruti Desai

UCE2023543 Nishtha Shah

Table of Content

Problem Statement

The Challenge:

- Millions of unstructured text documents created daily
- Valuable information buried in text: names, companies, locations, dates
- Manual extraction is slow and error-prone
- Existing tools require technical expertise or expensive APIs

Our Goal:

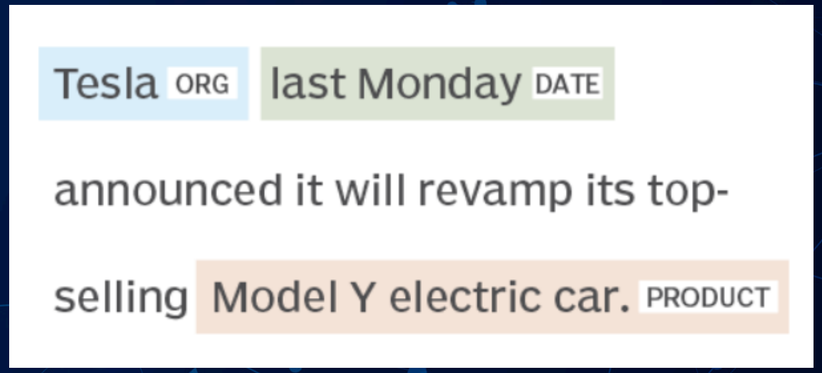
- Build an accessible web-based system that automatically identifies and extracts named entities from text with high accuracy and speed.

What is Named Entity Recognition?

Definition: Named Entity Recognition identifies and classifies named entities in text into predefined categories.

Entity Types We Detect:

- People: Elon Musk, Nishtha Sharma
- Organizations: Microsoft, Google, NASA
- Locations: California, India, Mumbai
- Dates: January 15 2024, yesterday
- Money: $2.5 billion, 500 rupees
- Miscellaneous: Products, events, languages



Technology Stack

Backend:

- Python 3.10
- FastAPI (REST API framework)
- spaCy / Transformers (NLP models)
- Uvicorn (ASGI server)

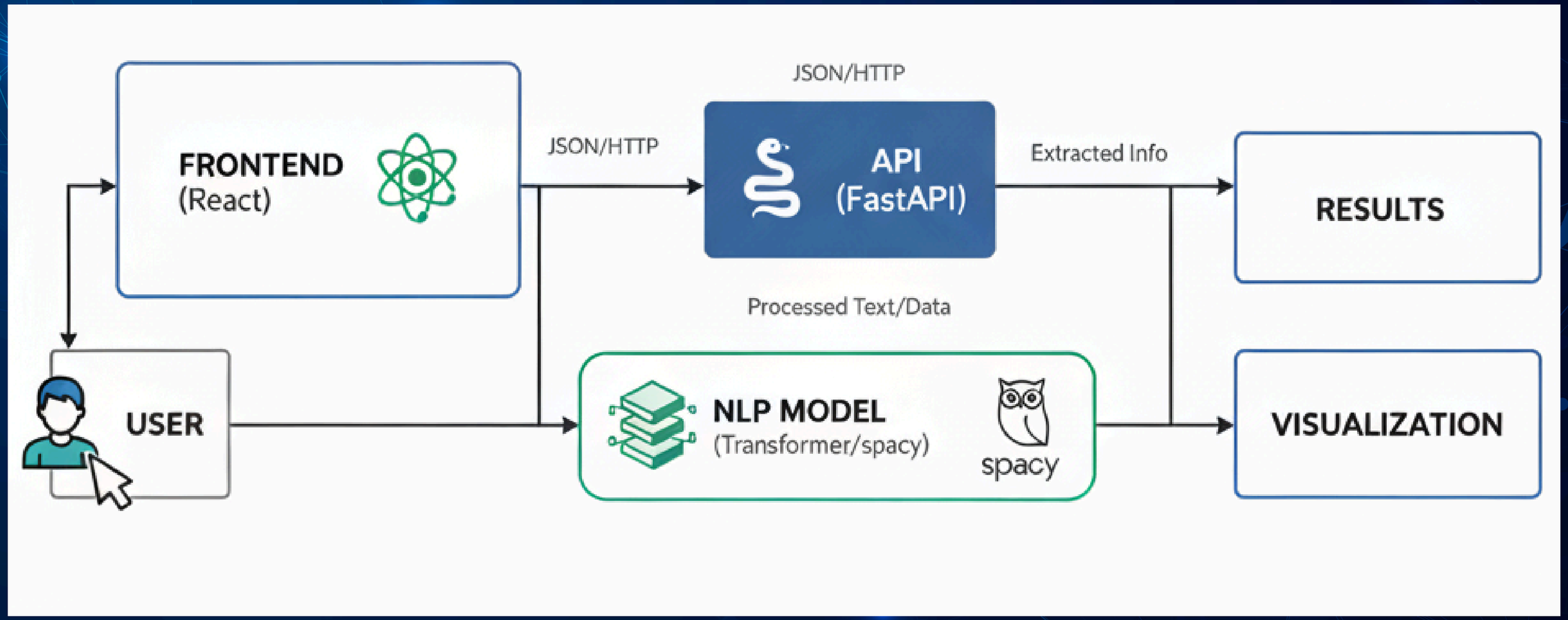
Frontend:

- React 18 (UI components)
- Tailwind CSS (styling)
- Axios (HTTP client)
- Vite (build tool)

Why These Technologies?

- Fast performance
- Industry-standard tools
- Easy deployment
- Active community support

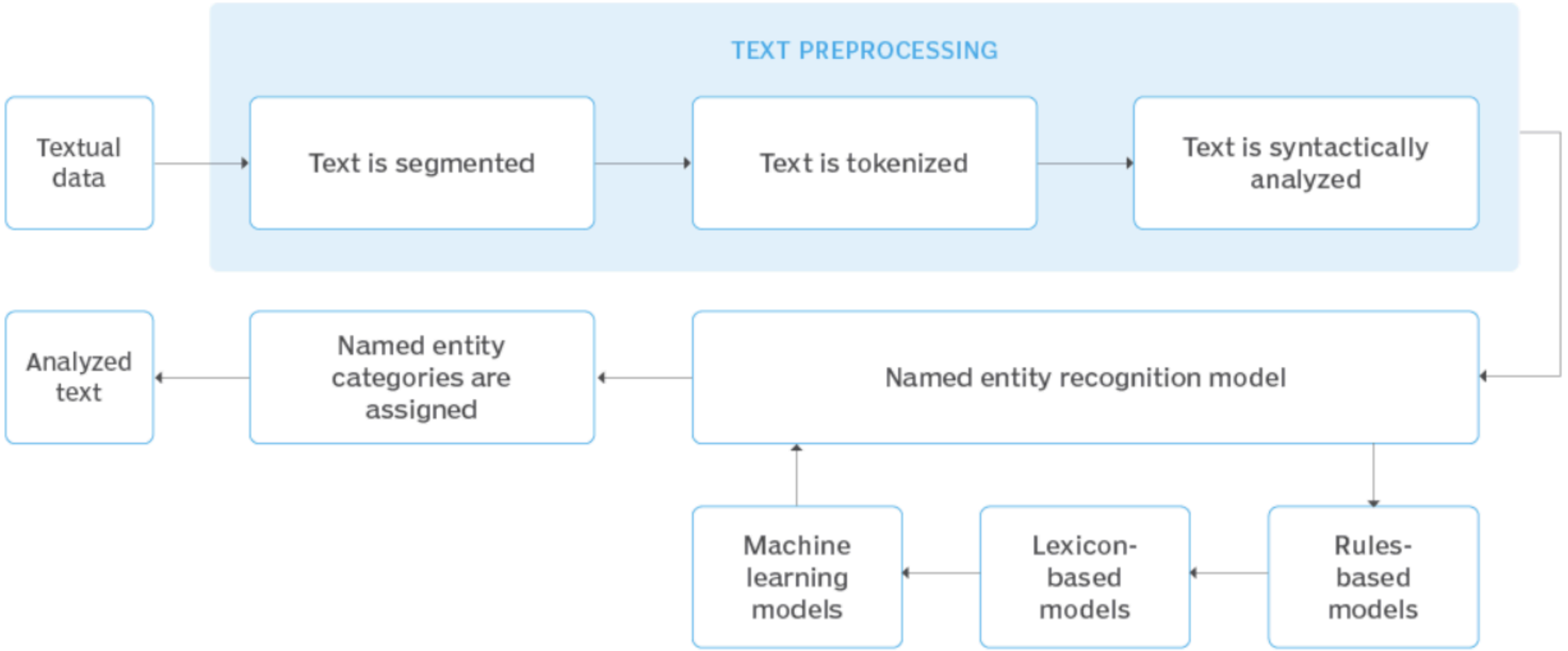
System Architecture



Features

- Real-Time Entity Recognition
- Text, PDF & Document Upload Support
- Color-Coded Highlighting
- Interactive Sidebar Entity List
- Sentiment Analysis Integration
- Contextual Categorization
- Analytical Dashboard

How preprocessing works



Model Training Process

1

Dataset Selection

- CoNLL-2003: 14,000+ news article examples
- WNUT-17: 3,400+ social media examples
- Combined for comprehensive coverage

2

Data Preparation

- Converted to spaCy training format
- Labeled entity boundaries and types
- Split into train/validation/test sets

3

Model Training

- Initialized blank spaCy model
- Trained for 30 iterations
- Validated performance continuously

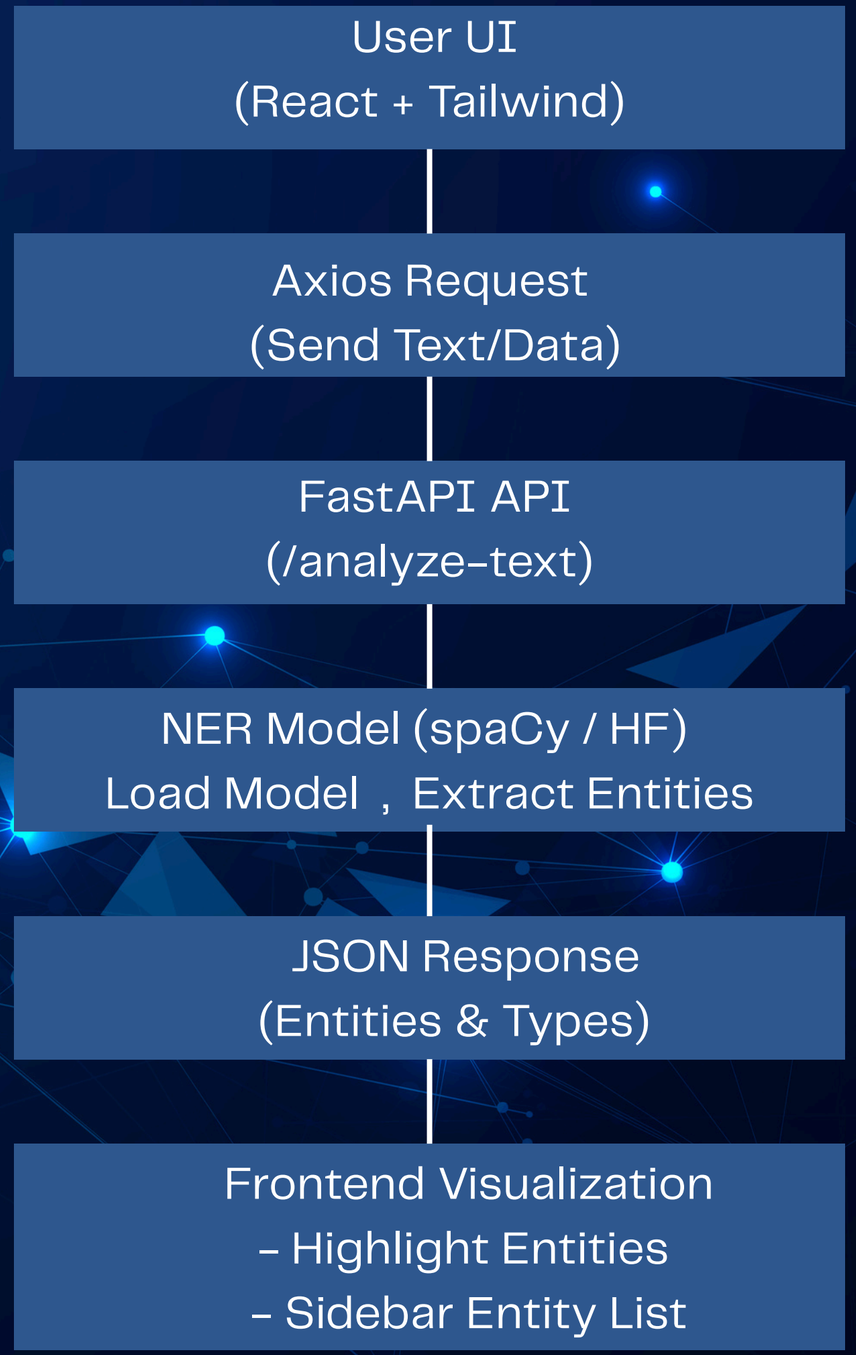
4

Evaluation

- Tested on held-out data
- Calculated precision, recall, F1 scores
- Result: Custom-trained model ready for deployment

NER Model / Pipeline

- Text Processing: Tokenization
- NER Prediction: Model predicts entity spans + labels
- Post-processing: Preparing JSON output (entity text, label, start/end)
- Model Variants: Use spaCy's small model (en_core_web_sm) or transformer model (en_core_web_trf) based on accuracy/speed tradeoff



System Flow

Frontend Flow & UI

- User enters or pastes text / uploads document
- Click “Analyze” → Axios call to API
- Response: JSON of entities
- UI: Highlight entities in text + show list in sidebar
- Interactive: clicking entities in sidebar highlights them in text

Backend Flow

- Text sent via API
- FastAPI endpoint receives request
- Model loaded in backend (spaCy or transformer)
- Entities extracted and packaged in JSON
- Response sent back to frontend

Performance Results

Accuracy :

Entity Type	Precision	Recall	F1 Score
PERSON	0.94	0.93	**0.94**
ORGANIZATION	0.92	0.89	**0.91**
LOCATION	0.91	0.90	**0.91**
MISCELLANEOUS	0.84	0.82	**0.83**
AVERAGE	**0.90**	**0.89**	**0.90**

Speed :

Document Length	Processing Time	Entities Found
100 words	180 ms	8-12
500 words	320 ms	35-45
1000 words	480 ms	70-90
2000 words	550 ms	140-180

Real-World Applications

1

Business Intelligence

- Extract companies and executives from reports
- Track competitor mentions
- Identify market trends
- Result: 85% time reduction in analysis

2

News Analysis

- Monitor person/organization mentions
- Track geographic coverage
- Identify trending topics
- Result: Process 1000+ articles/hour

3

Academic Research

- Extract researchers and institutions
- Identify methodologies and datasets
- Build collaboration networks
- Result: 95% automation of literature review

4

Legal Documents

- Extract parties and dates
- Identify jurisdictions
- Track regulatory references
- Result: 95% extraction accuracy

Future Work

Short-Term (Next 3 months):

- Entity linking to Wikipedia for context
- Relation extraction (who works where, who founded what)
- Batch processing for multiple documents
- Export in JSON and XML formats

Long-Term Vision:

- Multilingual support (10+ languages)
- Mobile application (iOS/Android)
- Confidence threshold controls
- Custom entity type definitions
- Active learning from user corrections

Research Directions:

- Few-shot learning for new entity types
- Cross-lingual entity recognition
- Zero-shot entity extraction

Conclusion

What We Built

- Production-ready NER web app
- State-of-the-art NLP + simple UI

Impact

- 99% faster entity extraction
- Useful for non-technical users
- Privacy-friendly deployment
- Extensible for domain-specific NER

Key Achievements

- ~90% F1 score
- <500ms processing
- Easy-to-use interface
- Supports multiple models
- Trained on 17k+ examples

Reference

- SpaCy Documentation Link
- CoNLL-2003 Dataset Paper
- WNUT-17 Dataset Paper
- FastAPI Documentation
- HuggingFace Transformers


Thank You !