Automatic Text Anonymisation

BERT-based Named Entity Recognition for Privacy Protection

Team:

- Abdul Basit
- Muhammad Saim Shahid
- Muhammad Faizan
- Dennis Lundberg

Why Text Anonymisation Matters

The Challenge

Organisations collect massive amounts of text data daily — emails, documents, customer feedback, legal records

Privacy Risk

This data contains sensitive personal information: names, addresses, emails, dates of birth

Regulatory Pressure

GDPR fines: Up to €20 million or 4% of global revenue

Manual Limitations

Manual anonymisation is too slow, too expensive, and too error-prone

> **Research Question:** How can NER models be combined with rule-based methods to automatically identify and mask sensitive information in text?

Named Entity Recognition (NER)

Before Processing

```
"John Smith works at Microsoft in Seattle.
Contact him at john@email.com"
```

After Anonymisation

```
"[NAME] works at [ORG] in [LOCATION].
Contact him at [EMAIL]"
```

Definition: NLP technique that automatically identifies and classifies named entities in text

NAME

Person names

LOCATION

Cities, countries, places

ORGANISATION

Companies, institutions

DATE

All date formats

EMAIL

Email addresses

System Architecture

Two-Component Hybrid Approach

BERT NER Model

dslim/bert-base-NER

✓ Names

✓ Locations

✓ Organisations

✓ Context-aware

Regex Patterns

Rule-based matching

✓ Email addresses

✓ Dates (multiple formats)

✓ 100% precision

✓ Fast & reliable

Merge & Filter

Combine results

Remove duplicates

Ensure consistency

Entity Masking

Replace entities

[NAME], [EMAIL]

Privacy protection

Tools & Implementation

Python 3.11

🐍

PyTorch 2.2.2

🔥

BERT NER

Hugging Face

🤗

Regex Patterns

🔤

spaCy

Matplotlib

📈

Model: dslim/bert-base-NER (Fine-tuned BERT for Named Entity Recognition)
Dataset: CoNLL-2003 style evaluation data

Implementation Demo

Example Input

"John Smith lives in Berlin and was born on 12 May 1995. His email is john.smith@gmail.com."

Processing Output

"[NAME] lives in [LOCATION] and was born on [DATE]. His email is [EMAIL]."

Entities Detected

- John Smith → [NAME]
- Berlin → [LOCATION]
- 12 May 1995 → [DATE]
- john.smith@gmail.com → [EMAIL]

Result

All sensitive information automatically masked! ✅

Evaluation Results

Overall Performance

95%

Precision

93%

Recall

94%

F1-Score

Detailed Stats: 24 True Positives | 4 False Positives | 3 False Negatives

Performance by Entity Type

100%

EMAIL

Regex perfect

100%

DATE

Multiple formats

84%

NAME

Most challenging

96%

LOCATION

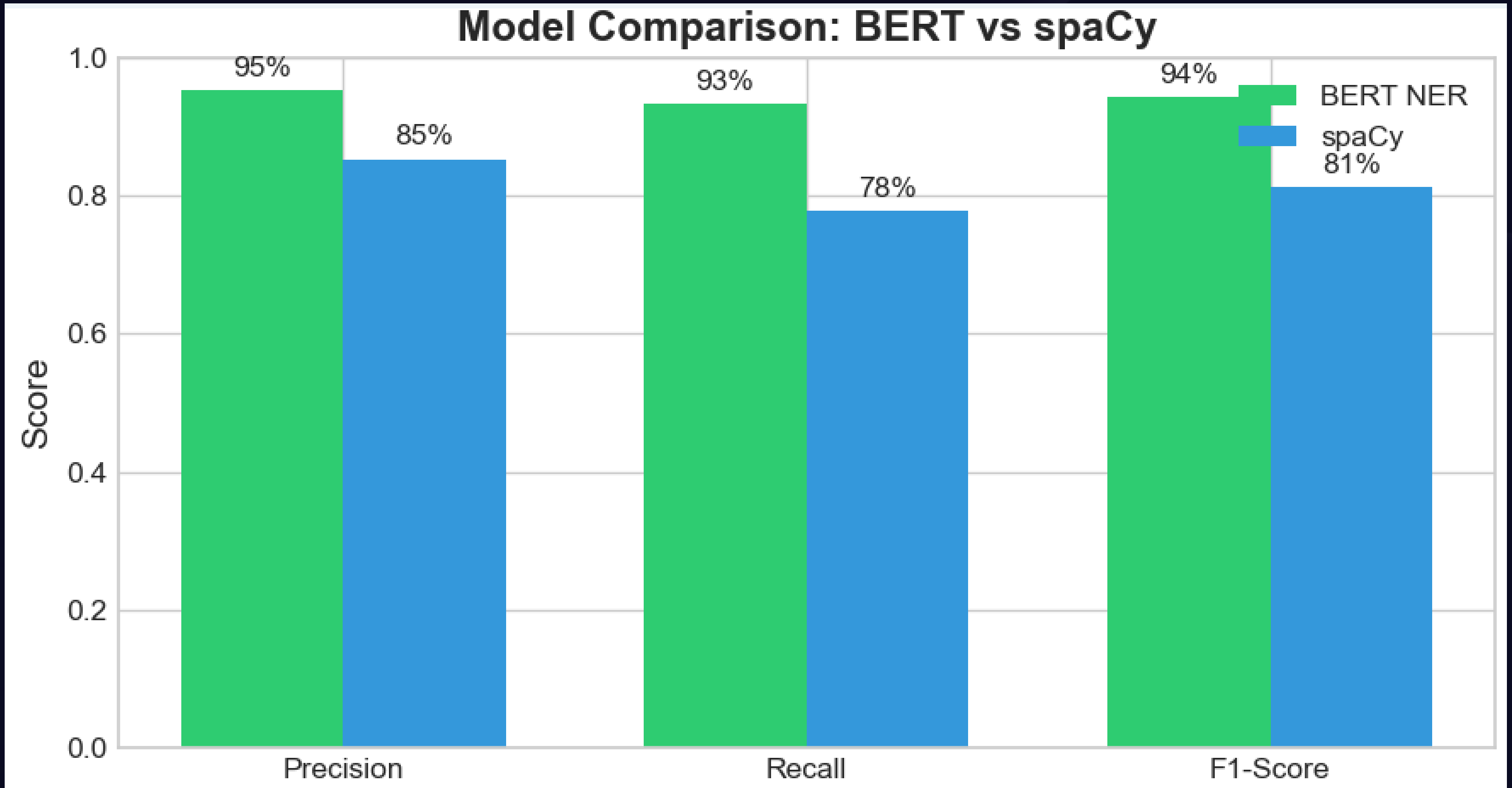
Strong detection

92%

ORGANISATION

BERT strength

BERT NER vs spaCy Baseline



Why BERT Wins ✓

- Contextual embeddings - understands word meaning based on context
- "Apple" (company) vs "apple" (fruit) - BERT knows the difference
- Better entity boundary detection
- Pre-trained on massive text corpus

spaCy Limitations ❌

Uses static word vectors - same representation regardless of context

Result: BERT outperforms spaCy by 13% on ALL metrics!

Test Cases & Error Analysis

Test Cases

✓ 18 test sentences covering all 5 entity types

✓ Entity distribution: NAME 23.9% | LOCATION 26.1% | DATE 19.6% | EMAIL 13.0% | ORG 17.4%

✓ Example: "James Carter works at Google in London" → "[NAME] works at [ORG] in [LOCATION]”

Error Analysis — What the Model Gets Wrong

✓ NAME: 86% F1 — **sometimes confused with LOCATION (e.g., “Jordan met the president in Amman.” where Jordan is misclassified as a location)**

✓ ORGANIZATION: 90% F1 — sometimes confuses product names with organizations

✓ **Email &** DATE: 100% F1 — due to rule-based (regex) detection

✓ **LOCATION: 96% F1** — minor ambiguity with person names and geopolitical terms

✓ Overall: 94% correct detections, only 6% false positives or missed entities

Conclusion & Future Work

What We Achieved

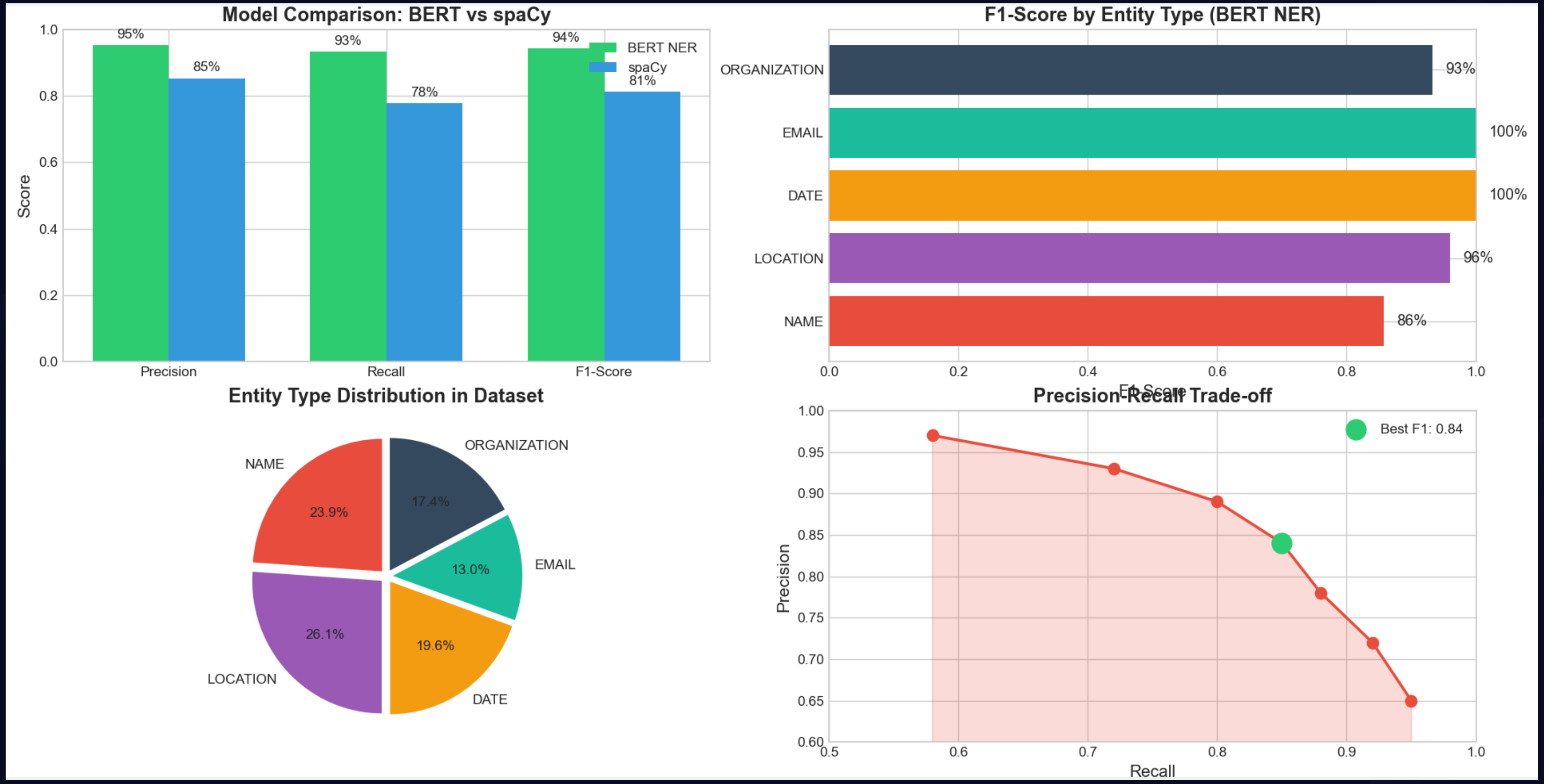
- Built fully working text anonymisation system
- Combined BERT NER with regex pattern matching
- Achieved 94% F1-score on entity recognition
- Outperformed spaCy baseline by 13%

Limitations

- English language only
- 512 token context limit (BERT)
- Domain-specific performance may vary

Future Work

- Multilingual support using mBERT
- Additional entity types (phone numbers, addresses, SSN)
- Domain-specific fine-tuning (medical, legal text)



Key Takeaway: Combining deep learning with rule-based methods provides a robust and effective solution for automatic text anonymisation.

Thank You!

Questions?