CEES: Crime Escalation Early-warning System

Using Natural Language Processing (NLP) to Predict Escalation Risk in Crime Narratives

Author: Ewelina Gradwicka

Year: 2025

Abstract

This project presents **CEES – Crime Escalation Early-warning System**, a prototype Natural Language Processing (NLP) model designed to identify early indicators of behavioural escalation in short crime narratives. Using a dataset of 29 police-style incident descriptions across three escalation levels (low, medium, high), CEES applies TF-IDF vectorisation and multinomial logistic regression to classify risk. Despite the limited dataset, the model correctly captured meaningful linguistic patterns aligned with criminological theory, including repeated monitoring, stalking precursors, coercive control dynamics, threats, and breaches of restraining orders.

An extended version, CEES 3.0, applied a transformer-based zero-shot classifier (BART-large MNLI), showing the differences between pattern-based and deep semantic approaches. The findings demonstrate the feasibility of narrative-based early-warning systems in policing and public safety and provide a foundation for future development.

1. Introduction

Crime narratives—short, textual descriptions of incidents—often contain behavioural cues that indicate risk escalation. Traditionally, the identification of these indicators relies on human interpretation, which can be subjective, inconsistent, and influenced by cognitive biases.

This project explores whether **Natural Language Processing (NLP)** can support early detection of escalation risk. CEES classifies incidents into three categories:

- **Low risk:** minor disturbances, isolated incidents
- **Medium risk:** interpersonal conflict, neighbour disputes, early warning signs
- **High risk:** threats, coercive control, stalking, repeated monitoring, breaches of legal orders

The goal is to evaluate whether machine learning models can recognise these behavioural signals and assist in early intervention strategies.

2. Dataset

A pilot dataset of **29 synthetic crime-style narratives** was created to emulate real incident patterns observed in policing environments:

- **10 low-risk** incidents
- **10 medium-risk** incidents
- **9 high-risk** incidents

The narratives cover a range of behaviours, including domestic abuse, harassment, stalking, antisocial behaviour, neighbour disputes, and general conflict escalation.

Each narrative was manually labelled with an escalation level, guided by criminological theory and common police risk assessment frameworks.

3. Methodology

3.1 Preprocessing

- Lowercasing
- English stopword removal
- Tokenisation

3.2 Feature Engineering

Narratives were encoded using **TF-IDF vectorisation**, transforming text into numerical features representing word importance across the dataset.

3.3 Classification Model

A **multinomial Logistic Regression classifier** was trained to predict escalation level.

Data split:

- **70% training**,
- **30% testing**,
- **Stratified** to preserve class balance.

Evaluation metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion matrix

3.4 Transformer Zero-shot Classification

To compare traditional and modern NLP methods, the model was extended using:

- **facebook/bart-large-mnli**
- With labels: *low escalation risk*, *medium escalation risk*, *high escalation risk*

This model performs classification **without task-specific training**, relying on semantic understanding.

4. Results – TF-IDF + Logistic Regression (CEES 2.0)

The model achieved an accuracy of **0.56**, which is expected for such a small dataset.

Despite limitations, meaningful criminological patterns were identified.

4.1 High-risk indicators captured by the model:

- partner
- ex
- victim
- followed
- consecutive
- threats
- restraining
- weapons
- monitoring

These reflect **stalking patterns, coercive control and repeat victimisation**, aligning with criminological research.

4.2 Medium-risk indicators:

- neighbours
- arguing
- returning
- night
- injuries
- customer
- incident

These represent **ongoing disputes and conflicts**.

4.3 Low-risk indicators:

- parking
- construction
- noise
- drivers
- teenager
- flat

These relate to **minor public disturbances**.

4.4 Confusion Matrix Summary

- **High risk** was often correctly identified
- **Low risk** achieved high recall
- **Medium risk** was the most ambiguous (common in real policing)

The model struggled with knife-related threats when contextual detail was lacking — a known limitation of TF-IDF models.

5. Results – Transformer Zero-shot Model (CEES 3.0)

The zero-shot model achieved an accuracy of **0.44**, but demonstrated different strengths:

Strengths

- High precision for **high-risk** cases
- Strong recall for **medium-risk** cases
- Better understanding of narrative semantics

Weaknesses

- Frequent confusion between low and medium risk
- Tendency to choose “medium” when uncertain
- Lack of domain fine-tuning reduces calibration

This highlights the difference between shallow and semantic NLP models.

6. Discussion

Both models showed that **text alone carries strong behavioural signals**, especially when describing:

- repeated contact

- monitoring and following
- coercive control dynamics
- threats
- breaches of legal boundaries

These patterns align with criminological theories such as:

- Repeat victimisation
- Cycle of violence
- Coercive control framework
- Threat escalation models

The results suggest that **machine learning can meaningfully support early-warning systems**, especially when combined with criminological expertise.

7. Limitations

- Small dataset
- Synthetic narratives
- No contextual metadata (history, relationship status, timestamps)
- No fine-tuning of transformer model
- Results should not be interpreted as operational-level performance

8. Future Work

Future development of CEES will include:

Data Expansion

- Larger labelled dataset
- Multi-jurisdictional narratives
- Real police report structures

Model Improvements

- Fine-tuning transformer models (BERT, RoBERTa)

- Adding sentence embeddings
- Sequence models with temporal data

Explainability

- SHAP/LIME for feature attribution
- Highlighting key narrative segments
- Model uncertainty estimates

Deployment Potential

- Integration with safeguarding units
- Policing risk assessment tools
- Real-time flagging of escalation indicators

9. Conclusion

CEES demonstrates that even simple NLP methods can detect meaningful escalation patterns in crime narratives. Although preliminary, the project shows strong potential for integrating criminology and AI. With expanded datasets and transformer fine-tuning, CEES could evolve into a powerful early-warning system for domestic abuse units, threat assessment teams and policing services.