Paper 1: Terrorism Analytics: Learning to Predict the Perpetrator

Objective and Scope

- Analysed terrorist attacks in India using machine learning to predict the group responsible for a given attack.
- Dataset: Indian subset of the Global Terrorism Database (GTD), covering data from 1970–2015.

Key Methods and Algorithms

1. Support Vector Machine (SVM)

- A supervised learning algorithm for classification and regression tasks.
- Used for predicting the perpetrator based on attributes like attack type, target type, and weapon type.
- Dataset attributes were categorical but effectively modeled using a linear kernel.

2. Decision Tree (C4.5 Algorithm)

- Builds a tree structure by recursively splitting data based on the attribute that provides the highest information gain (entropy reduction).
- Used to classify and predict the terrorist group responsible for attacks.
- Attributes like target type, attack type, and weapon type were key decision points.

3. Random Forest

- **Uses bagging (bootstrap aggregation) to train each tree on a random subset of the data.**
- **Applied to classify and predict perpetrators using the GTD dataset.**

4. Factor Analysis of Mixed Data (FAMD)

- Reduces high-dimensional data into principal components while handling mixed (numerical and categorical) attributes.
- Identifies attributes with the highest contribution to variance.
- Used for dimensionality reduction and feature selection in GTD data.

- Key attributes extracted include year, attack type, target type, weapon type, and location (latitude/longitude).

Key Results

Effective attributes for prediction include attack type, target type, weapon type, and incident location.

Applications

- Potential to aid investigative agencies in narrowing down suspects based on attack patterns.
- Suggests exploring ensemble classifiers and deep learning for further accuracy improvements.

Paper 2: ConfliBERT: A Language Model for Political Conflict

Objective:

Developed a domain-specific language model (ConfliBERT) for analyzing political conflicts and violence.

Key Methods and Innovations

- **Fine-Tuned BERT Architecture:** Trained on a 33.7GB curated corpus of conflict and political violence data.
- **Tasks Supported:**
    1. Binary Classification: Identifies texts related to political violence.
    2. Multi-Class Classification: Categorizes conflict events (e.g., bombings, armed assaults).
    3. Named Entity Recognition (NER): Extracts entities like actors, victims, locations, and dates.

Key Results

- **Performance Metrics:**
    - Outperforms generalist LLMs in precision, recall, and computational efficiency.
- **NER Examples:**
    - Extracts detailed event attributes (e.g., actors, targets) from texts like news reports and datasets.
- **Binary Classification:**
    - Accurately filters political violence-related texts, achieving confidence levels over 99% in test cases.

Applications

- Enables efficient and scalable event coding for political science research.
- Suitable for real-time analysis of emerging conflicts, significantly reducing human annotation costs.
- Potential for extensions into multilingual support and downstream tasks.