Impact of Social Factors on Criminal Activities: A Decade-long Pakistan-based Analysis

Project Proposal

Session: 2022 – 2026

Submitted by:

Rehan	2021-SE-47
Sadia Sher	2022-SE-07
Muntaha Kashif	2022-SE-08
Maheen Zahid	2022-SE-12
Rimsha Aamir	2022-SE-13

Supervised by:

Mr. Nadeem Iqbal

Department of Computer Science, New Campus

University of Engineering and Technology

Lahore, Pakistan

Table of Contents

1 Proposal Synopsis

1.1 Abstract

This project aims to analyze and predict crime patterns in Pakistan using criminal records from the past decade. We will collect data from newspapers, TV channels, and socio-environmental sources and annotate it to train an LLM-based system. This system will classify and predict crime trends while exploring correlations with weather, population demographics, and socioeconomic conditions. Geo-spatial and temporal analysis will support the identification of crime hotspots and trends over time.

1.2 Introduction

Crime continues to pose a significant challenge in Pakistan, undermining public safety, economic stability, and citizen trust in law enforcement. As cities grow and socio-political dynamics shift, understanding the evolving nature of criminal activity becomes more important than ever. Despite the abundance of historical crime data maintained by police departments and government institutions, much of it remains underutilized, especially in predictive and analytical domains. Meanwhile, the rapid growth of digital journalism and media reporting has resulted in the availability of vast unstructured crime-related content from newspapers, television broadcasts, and social media platforms. These sources often contain rich details about criminal incidents, public sentiment, and contextual factors—data that can be crucial in identifying hidden trends and forecasting crime patterns.

However, existing crime analysis tools in Pakistan primarily rely on manual, structured data entry systems that lack scalability and depth. Most predictive tools are rule-based, regionally limited, or imported from Western contexts, making them ill-suited for addressing the complex social, cultural, and economic fabric of Pakistani society. In addition, these tools rarely incorporate real-time external data such as weather conditions, population density, or socio-economic indicators, all of which can significantly influence crime rates.

With advancements in natural language processing (NLP) and the emergence of large language models (LLMs), it is now possible to analyze and extract insights from diverse sources of unstructured data. Our proposed project aims to bridge this technological gap by developing a localized, LLM-based crime prediction and classification system tailored specifically for Pakistan. By combining criminal records with media reports and socio-environmental data, we intend to build a system that not only identifies current crime hotspots but also predicts future trends with contextual awareness. The project will also integrate spatial and temporal analysis to provide a more comprehensive understanding of crime evolution over time and across different regions.

Ultimately, this research has the potential to transform how law enforcement agencies, policymakers, and researchers understand and respond to crime. It seeks to shift crime response from reactive to proactive—helping prevent incidents before they occur through timely insights and data-driven recommendations.

1.3 Problem Statement

Despite increasing crime rates and the availability of diverse crime-related data in Pakistan, there exists no comprehensive, data-driven system that leverages large language models (LLMs) to classify, analyze, and predict criminal activity. Current tools are either rule-based, geographically irrelevant, or incapable of processing unstructured media data. This lack of intelligent, context-aware predictive systems leaves policymakers, security agencies, and local authorities without actionable insights to detect emerging crime trends or deploy early interventions. There is a critical need for a localized, LLM-powered solution that integrates multiple data sources—including media reports, weather, demographics, and historical crime records—to provide real-time, scalable, and explainable crime analysis for Pakistan.

1.4 Objectives

- Collect crime-related data from newspapers, TV news, and external sources.
- Annotate crime data with relevant categories and influencing factors.
- Develop and fine-tune a large language model (LLM) to classify and predict crime types.
- Conduct exploratory analysis for factors such as weather, population density, and income levels.
- Perform spatial and temporal visualizations to identify crime hotspots and evolving patterns.

1.5 Related Work

Related Work	Weakness	Proposed Project Improvement
Basic Text Classifiers	Poor performance on Urdu/English mix	Fine-tuned local LLM for language support
CrimeStat Tools	Focus on GIS, no AI	LLM for both prediction & insight extraction
U.S.-based Predictive Tools	Not tailored to Pakistan	Region-specific training and analysis
Rule-based NLP for News	Doesn’t scale for deep semantic understanding	Context-aware deep language models
Existing Police Records	Structured but lacks news & external data	Richer, multi-source dataset

1.6 Proposed Methodology/System

1. Historical Data Collection
   - Pakistani English newspaper archives
   - TV crime news transcripts
   - External datasets: weather, census, socioeconomic reports

2. Data Annotation
   - Tagging for crime class, influencing factors, and locations

3. LLM-Based Classification
   - Development and fine-tuning of LLM models for classification

4. Enhanced LLM for Crime Prediction
   - Few-shot prompting for forecasting crime types and trends

5. Exploratory Insights
   - Weather-crime correlation
   - Demographic and crime linkage
   - Socioeconomic and crime correlation
   - Temporal crime pattern analysis
   - Geo-spatial crime hotspot identification

1.7 Tools and Technologies

- Languages & Libraries: Python, PyTorch, HuggingFace, Pandas
- Models: GPT-3.5/4, BERT, UrduBERT
- Annotation: Prodigy, Label Studio
- GIS & Visualization: QGIS, Seaborn, Plotly
- Data Handling: Excel, JSON, CSV
- Dev Tools: Jupyter, Google Colab

1.8 Work Division

Team Member	Tasks
Member 1	News & TV crime data collection + cleaning
Member 2	Annotation of text, crime types, external factor integration
Member 3	LLM fine-tuning and classification model

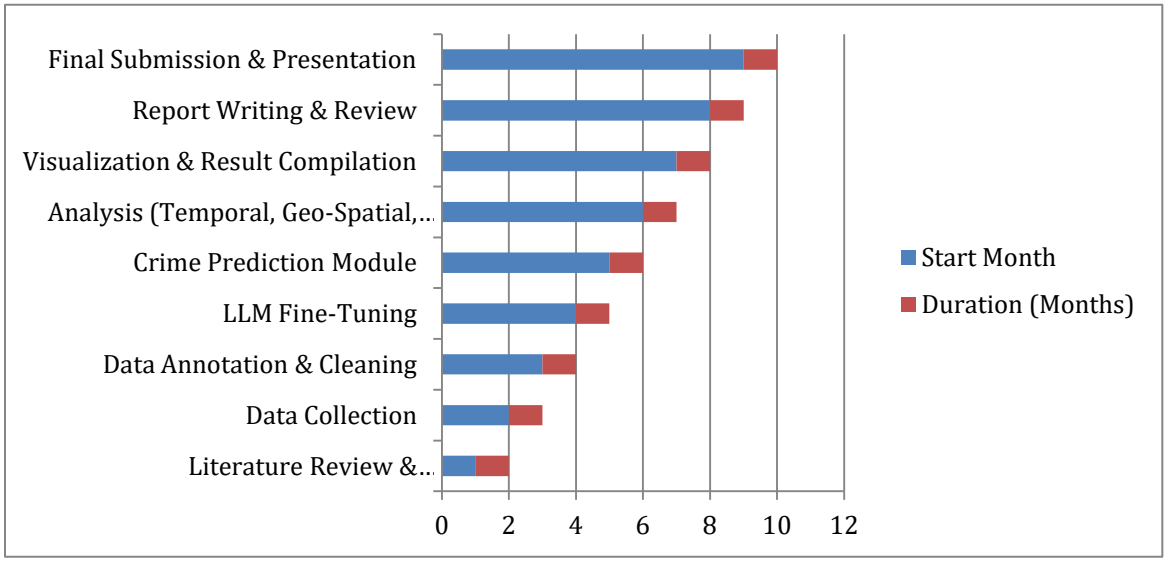
	development
Member 4	Prediction logic, prompting, and testing
Member 5	Exploratory analysis (temporal, spatial, demographic), visualization

1.9 Data Gathering Approach

We will extract data from digital archives of Pakistani newspapers (e.g., Dawn, Jang, The News), TV broadcast summaries, and external datasets like Pakistan Meteorological Department reports, Bureau of Statistics surveys, and World Bank data. Manual tagging and automated annotation tools will be used to prepare the dataset.

1.10 Timeline/Gantt Chart

Month	Task
Month 1	Literature Review & Requirement Gathering
Month 2	Data Collection
Month 3	Data Annotation & Cleaning
Month 4	LLM Fine-Tuning
Month 5	Crime Prediction Module
Month 6	Analysis (Temporal, Geo-Spatial, Factor-Based)
Month 7	Visualization & Result Compilation
Month 8	Report Writing & Review
Month 9	Final Submission & Presentation



References

- Ahmed, M. (2021). Trends in Urban Crime in Pakistan. Journal of Social Issues.
- Dawn, Jang, Geo News archives

- Pakistan Meteorological Department
- Pakistan Bureau of Statistics