Information Retrieval Course Project Proposal

Prateek Kumar Padhy
Computer Science
Virginia Tech
Falls Church, VA, USA
pprateekkumar@vt.edu

Jayanth Koripalli
Computer Science
Virginia Tech
Falls Church, VA, USA
jayanth28@vt.edu

Siji Chen
Computer Science
Virginia Tech
Falls Church, VA, USA
sijic@vt.edu

***Abstract*—Large Language Models (LLMs) have opened the door for revolutionary data analysis methods and insight production in information retrieval. The aim of this project is to investigate and optimize open-source LLMs, particularly those with parameter values under 2 billion, so as to meet the limitations of the available computing power—a 16GB GPU, in particular. This study attempts to find an effective yet potent model that can enable advanced crime data analysis by comparing models like LLaMA-3B, BERT, and RoBERTa. The project is structured around key milestones, including proposal submission, checkpoints, and final presentations, culminating in the application of the selected LLM to three critical areas: identifying crime hotspots, generating public safety reports, and developing a community engagement platform. Through the integration of LLMs with a targeted preprocessing and data handling backend and an intuitive front end, this study aims to provide new perspectives on crime patterns and risk factors, which will benefit the public safety and information science fields.**

I. Introduction

Crime prevention and prediction is a crucial problem in our society and hence attracted a lot of attention. In academia field, researchers have proposed numerous studies on crime analysis and predictions from varies aspects, such as crime hotspot discovery [1], incidence analysis and crime predictions [2]. Meanwhile, the advent of large language models (LLMs) has revolutionized natural language processing, enabling it with the capability to comprehend inference and reasoning tasks. Leveraging this advancement, we will employ the LLM to analyze thousands of crime-related news articles from Timesofindia, delving into the narratives concealed within the headlines.

Our application will furnish users with an intuitive interface to explore crime-related content through a free text queries. For example, users are able to find crime based on location, for which we will use the LLM to extract the location mentioned the article. In addition to the search interface for information retrieval, we will incorporate a dashboard function to visually display the analyses and patterns unearthed during data mining, such as crime hotspots within the city.

II. Related work

A. GraphBERT

GraphBERT [3] uses a integrated framework to assist the malicious behavior detection on social media platform. It's proposed to bridging the gap between graph and text; which can detect the behaviors with both semantic and user relational information. This paper doesn't have dataset/code published. They mined twenty million tweets from over one million twitter users; after cleanup, 91,500 tweets from 13,351 users are selected. Tasks includes malicious user detection; malicious tweet classification.

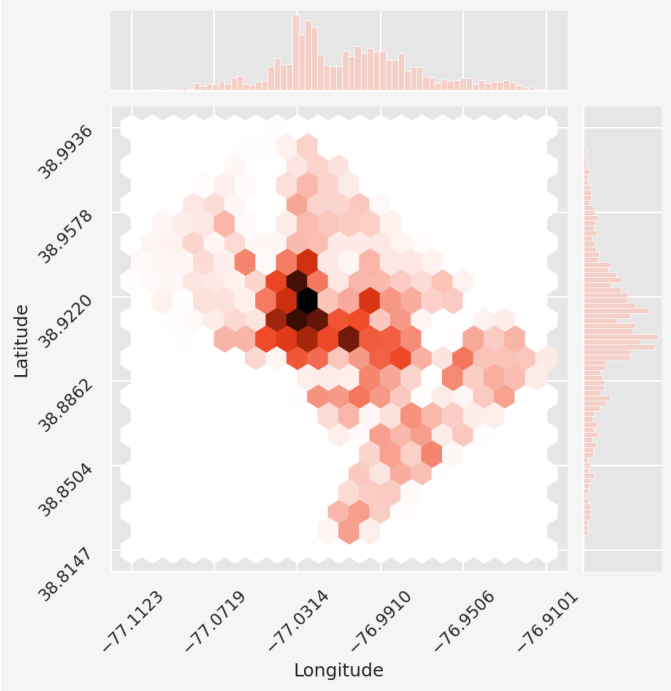

Fig. 1. D.C. 2023 crime overview. 26,091 crime incidents are recorded. Each hexagon represents a geographical area and the color intensity indicates the density of reported crimes in that area. Westeast area has the highest crime rate.

B. Discrimination/ hate crimes across U.S. Cities

[4] City-level analysis. They combined hate crime data from FBI; twitter data collected from 2011 to 2016; census data. Tasks include classifications on discrimination; spatial crime distribution; hate crimes and social media relationship. They found the top 20 features that can predict 25% of the discrimination. Phoenix, AZ had the highest number of race, ethnicity or national-origin based hate crimes over the 6 years.

C. Crime Analysis and Prediction using Data Mining

[5] Applying data mining techniques to the interface of computer science and criminal justice represents a revolutionary approach to crime research and prevention. This methodology allows the visualization of crime-prone areas and the prediction of regions with a high probability of crime occurrence by focusing on the systematic identification and analysis of crime patterns and trends rather than traditional factors like political enmity or the criminal background of the offender. The introduction of electronic systems has greatly improved the skills of crime data analyzers, enabling law enforcement officials to solve crimes more quickly. Using unstructured data to extract previously undiscovered and valuable information, this method highlights a move toward utilizing technological improvements for more effective and efficient techniques.

D. Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions

[6] The field of crime prediction has witnessed a significant transformation with the advent of machine learning and deep learning techniques. A comprehensive review which explored over 150 articles, emphasizing the diverse range of algorithms employed in crime prediction. This body of work underlines the critical role of predictive analytics in identifying crime patterns and trends, offering invaluable insights for law enforcement strategies. Key findings from the review highlight the efficacy of various machine and deep learning models in not only forecasting crime occurrences but also in unveiling underlying factors contributing to criminal activities. The review further sheds light on the datasets commonly used in this research area, providing a foundational resource for future investigations. Acknowledging the potential gaps identified in current methodologies the paper propose future directions aimed at enhancing prediction accuracy, such as integrating socio-economic indicators and adopting more sophisticated neural network architectures. This systematic review serves as a cornerstone for our research, guiding our exploration of novel approaches in crime prediction and contributing to the ongoing dialogue within the academic community on improving public safety through advanced analytics.

E. Bias and ethical concerns

Language models like LLM (Large Language Models) can reflect biases present in the data they're trained on [7], [8]. They learn patterns and associations from vast amounts of text data, which can include biases present in society, such as gender, racial, or cultural biases. These biases can manifest in various ways, including in the language generated by the model or in the responses it provides.

Efforts are made to mitigate these biases during training and in post-training evaluation. Techniques like data preprocessing, fine-tuning, and bias mitigation strategies are employed to address these issues. However, complete elimination of bias is challenging and often an ongoing process as models evolve and new biases are identified and addressed. It's important to continuously evaluate and improve models to reduce bias and ensure fairness and inclusivity in their outputs.

III. Proposed Approaches

A. Selection and Optimization of LLMs

Our aim is to identify the most efficient and effective Large Language Models (LLMs) that are suitable for analyzing crime data within the computational constraints of a 16GB GPU. We plan to evaluate the performance of various LLMs, including LLaMA-3B, BERT, and RoBERTa, on benchmark datasets related to crime. I will implement parameter tuning and model pruning techniques to optimize these models for enhanced performance while ensuring they remain computationally efficient.

B. Data Preprocessing and Integration

We intend to develop a robust preprocessing pipeline that can clean, standardize, and enrich crime datasets for more effective analysis. We will design and implement a preprocessing pipeline that incorporates data cleaning, normalization, and the integration of additional data sources for a richer analysis context. We aim to utilize geospatial data processing techniques to aid in the identification and categorization of crime hotspots.

C. Crime Hotspot Identification

Our goal is to use optimized LLMs to analyze crime data effectively and identify potential crime hotspots. We plan to apply natural language processing (NLP) techniques to extract and analyze relevant information from crime reports and related texts. We will integrate geospatial analysis with NLP findings to identify and visualize areas with high crime rates.

D. Public Safety Report Generation

We aim to automatically generate comprehensive public safety reports based on the analyzed crime data. We will develop a template for public safety reports that includes essential metrics, trends, and actionable recommendations. Using the optimized LLMs, we plan to populate this template with data-driven insights and analyses.

E. Community Engagement Platform Development

We intend to design and implement a web-based platform that is integrated with the backend LLM analysis system. The platform will feature interactive maps, report generation capabilities, and mechanisms for community feedback.

IV. Experiment

A. Large Language Model (LLM)

We are looking for open-sourced LLM that could be integrated into our project. Based on each type of tasks, we maybe find out the need to use inference only or fine-tune the open source LLM. Hence the GPU memory is a consideration. The GPU Quadro we have is 16GB so need a compact and accurate model which has less than 2 billion parameters. To run the 7B

model in full precision, one need $7*4 = 28GB$ of GPU RAM. One should add $torch_dtype = torch.float16$ to use half the memory and fit the model on a T4 which is 16GB. To train, 56GB is needed to store gradients for each parameters. Table I summarized the popular LLM size.

Model	Parameter	Memory Full precision
LLaMA-3B	3B	12 GB
LlaMA -7B	7B	28 GB
Vicuna	13B	52 GB
BERT	345M	1.4 GB
RoBERTa	355M	1.5 GB
MPT-7B	7B	28 GB
Flan-T5	11B	44 GB

TABLE I
CURRENT OPEN SOURCE LLM PARAMETERS

Based on Table I, we plan to start with the following open source model: 1) BERT based IR+QA Paper 2)Pytorch roBERTa 3)Keras/TensorFlow roBERTa

B. Streamlit

Our project will make use of the open-source app framework Streamlit to provide a user-friendly API that links the front end and our backend Large Language Models (LLMs). We will be able to quickly prototype and implement a dynamic interface for our crime data analysis platform thanks to Streamlit's ease of use and effectiveness in developing interactive web apps. With the use of this interface, customers will be able to communicate with our LLMs in real-time and get instant access to functions like community involvement, public safety report generating, and crime hotspot identification. We want to make sure that our platform is not just strong in its analytical skills but also easily accessible and interesting to a diverse user base, which is why we have planned on integrating Streamlit.

C. Huggingface

Hugging Face is a company that develops and maintains an open-source platform for natural language processing (NLP) called the "Transformers" library. The library provides easy-to-use interfaces to access and fine-tune pre-trained language models, including models like BERT, GPT, and others.

The Transformers library, along with the associated ecosystem of tools and resources provided by Hugging Face, has become immensely popular among researchers, developers, and practitioners in the field of NLP. It simplifies the process of working with state-of-the-art language models, enabling tasks such as text classification, question answering, text generation, and more.

With its "pipelines" feature, the Transformers library from Hugging Face makes it easier to use cutting-edge machine learning models for computer vision and natural language processing (NLP) applications. These pipelines provide a high-level, user-friendly interface for carrying out common tasks like text classification, token classification (e.g., named entity recognition), question answering, translation, summarization, and more. They abstract away the complexity involved in the preprocessing of data, prediction, and postprocessing of results.

Starting with advanced NLP and machine learning models is quite simple because each pipeline is tailored for a specific goal and comes preset with default models and tokenizers that have been tuned for that particular activity. In addition, users have the ability to modify pipelines by selecting various models and tokenizers, which allows them to adapt the pipeline to their own requirements and datasets.

To put it briefly, Hugging Face's pipelines offer developers and researchers a smooth and effective means of utilizing sophisticated machine learning models with little setup, freeing them up to concentrate more on their particular issues rather than the nuances of model implementation and data preprocessing.

In addition to the Transformers library, Hugging Face also offers a model hub where users can discover, share, and download pre-trained models and model checkpoints for various NLP tasks. They also provide tools for model training, evaluation, and deployment, making NLP accessible to a wider audience of developers and researchers.

V. DATASET

Crimes in India India is a country of 1.3 billion people, along with it's high population there's tens of thousands of crimes recorded all over the country. This dataset contains links to the articles from times of India website which describes the M.O. of the crimes. This dataset will help us to learn crime trends in different regions of India

Crimes in Washington DC The Washington DC government provided this carefully selected information, which is a crucial representation of specific locations and characteristics of criminal incidences that were reported in 2023. It helps scholars, decision-makers, and the general public comprehend the crime landscape in the nation's capital by providing a window into the dynamics of the city's criminal activity. This dataset comprises of 26,091 crime incidents along with 28 features. Key attributes include neighborhood clusters, offense types, geographical details, timestamps, and associated crime classifications. In this dataset, nine categories of offence

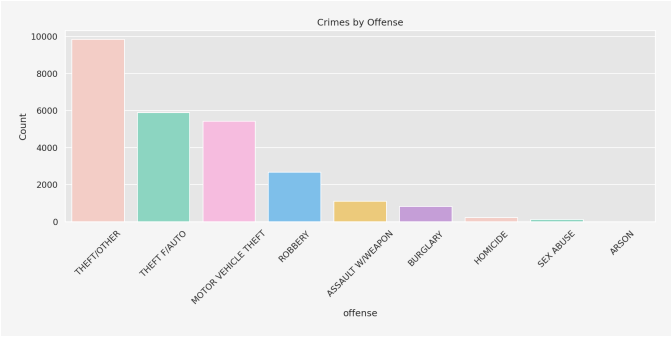

Fig. 2. D.C. 2023 crime with offense type. The three offense types of theft together account for 81% of the total crime. High correlations are observed for banks and ATMs. Museums has the lowest correlation.

types are reported: theft/other, theft f/auto, motor vehicle theft, robbery, assault w/dangerous weapon, burglary, homicide, sex

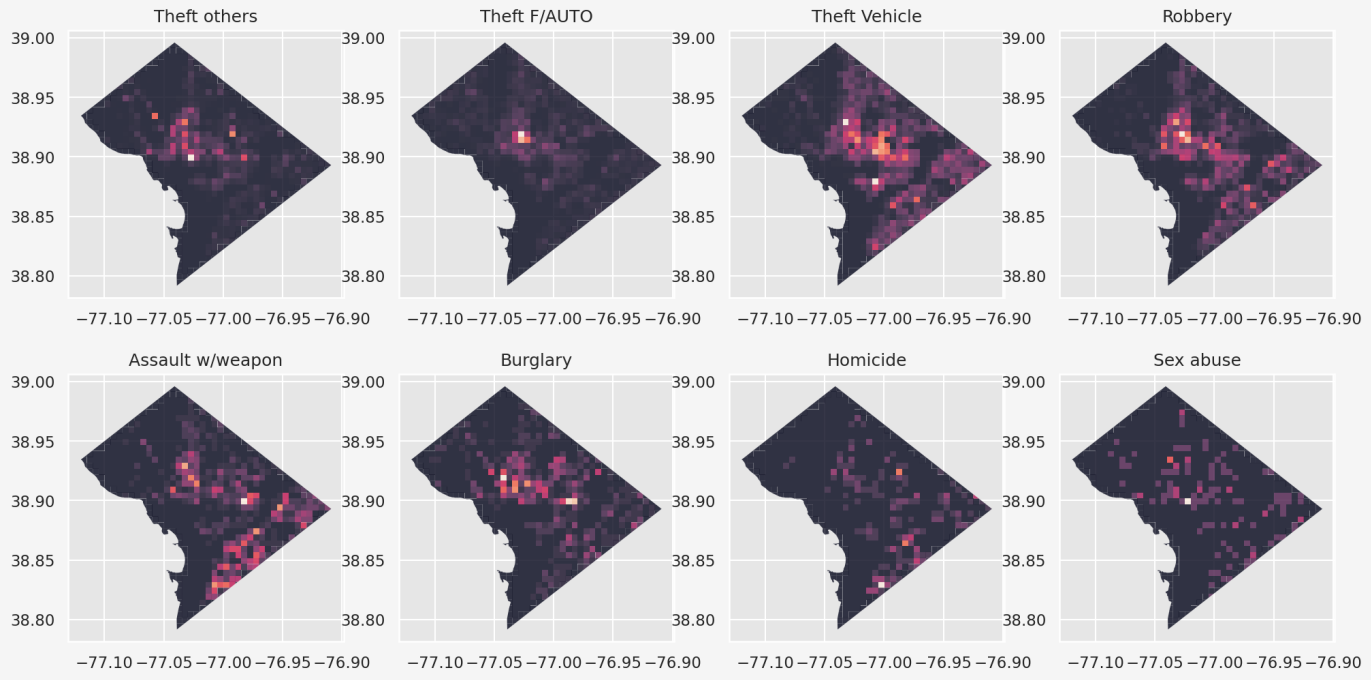

Fig. 3. D.C. 2023 crime breakdown by offence type. Different patterns are observed for different offense types.

abuse, and arson as shown in Figure. 2. In Figure. 1, the aggregated crimes are plotted and west-east area has the highest crime rate, which is counter-intuitive. We will analyze it more in-depth to hopefully find an explanation for it. In Figure. 3, crime counts are breakdown by the offense type.

Open Data DC It has city wide data for Washington D.C. In Figure. 4, the D.C. area is partitioned into 800 areas, other information is filled into each area such as the number of ATMs, banks, libraries, gas stations, parks, etc. Banks and ATMs have the highest correlation to total crime incidents. Followed by metro stations and gas stations. Museums have the lowest correlation to crimes, even lower than police stations.

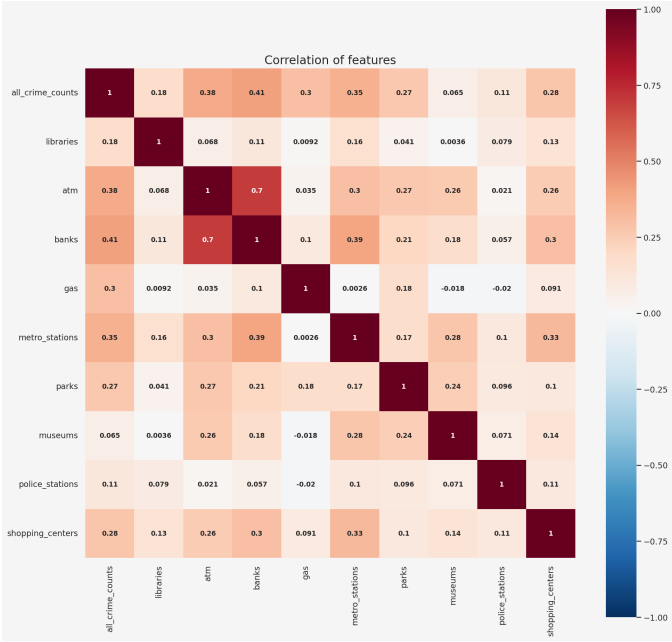

Fig. 4. D.C. 2023 crime correlation with other spatial features.

VI. System Architecture

The system architecture depicted in the diagram provides a seamless interface for the user to input data and receive refined responses. Upon submission of a query through the user interface (UI), it undergoes a tokenizing process that converts the input into a form suitable for machine understanding. Simultaneously, a representation of the query is matched against a vector database to find contextually relevant embeddings. This ensures that the language model has the most pertinent information available. Data filtering mechanisms safeguard the process by preventing the model from processing any unauthorized data.

The next phase involves the incorporation of context snippets into the prompt, which is then cached or retrieved from the Large Language Model (LLM) cache to optimize response times. The LLM API, powered by DistilledGPT2, processes the prompt using a transformer architecture, which may further incorporate insights from specialized databases such as a crime database. Finally, the output generated by the LLM is subjected to content filtering to ensure it aligns with prescribed content guidelines before being presented back to the end user through the UI. This architecture ensures that the user receives a prompt, accurate, and contextually relevant response, bolstered by an efficient and secure backend

workflow.

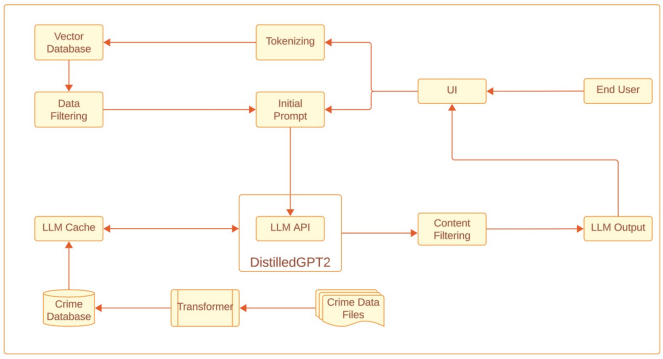

Fig. 5. System Architecture

VII. Appendix

A. Timelines

1) 2/20 Proposal
2) 3/19 check point
3) 4/9 presentation
4) 4/23, 4/23 final presentation
5) 5/7 project due

B. Tasks

1) User interface: webapp/Tableau/Streamlit.
2) Backend: model training; model validation and testing; API to UI(maybe not need)
3) database: maybe we don't need inmemory, postgresql and others.
4) Testing
5) Document writing;

C. Progress Report

1) Siji Chen:
   - (In progress) Exploratory data analysis (EDA) for the D.C. crime dataset.
   - (In progress) Check more dataset such as metro stations, schools, and gas stations to merge in our current crime dataset.
   - Define the interface to the front end and LLM model.
2) Jayanth Koripalli:
   - Finalizing the LLM Model we shall be using (distill gpt 2)
   - (In progress) Testing more non-gpu intensive models which have more parameters.
3) Prateek Kumar Padhy:
   - Designing the System Architecture.
   - (In Progress)Embedding the data and training the model.
   - Fine-tuning the transformer.

References

[1] J. E. Eck, S. Chainey, J. G. Cameron, M. Leitner, and R. E. Wilson, "Mapping Crime: Understanding Hot Spots."

[2] H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime Rate Inference with Big Data," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 635–644. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939736

[3] J. Wu, C. Zhang, Z. Liu, E. Zhang, S. Wilson, and C. Zhang, "GraphBERT: Bridging Graph and Text for Malicious Behavior Detection on Social Media," in *2022 IEEE International Conference on Data Mining (ICDM)*. Orlando, FL, USA: IEEE, Nov. 2022, pp. 548–557. [Online]. Available: https://ieeexplore.ieee.org/document/10027673/

[4] K. Relia, Z. Li, S. H. Cook, and R. Chunara, "Race, Ethnicity and National Origin-Based Discrimination in Social Media and Hate Crimes across 100 U.S. Cities," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 417–427, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3354

[5] S. Sathyadevan, M. Devan, and S. S. Gangadharan, "Crime analysis and prediction using data mining," in *2014 First international conference on networks & soft computing (ICNSC2014)*. IEEE, 2014, pp. 406–412.

[6] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime prediction using machine learning and deep learning: A systematic review and future directions," *IEEE Access*, 2023.

[7] P. N. Venkit, S. Gautam, R. Panchanadikar, T.-H. K. Huang, and S. Wilson, "Nationality Bias in Text Generation," Feb. 2023, arXiv:2302.02463 [cs]. [Online]. Available: http://arxiv.org/abs/2302.02463

[8] O. Shaikh, H. Zhang, W. Held, M. Bernstein, and D. Yang, "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning," Jun. 2023, arXiv:2212.08061 [cs]. [Online]. Available: http://arxiv.org/abs/2212.08061