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.

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

[**?**] 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.

III. 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

IV. Dataset

7K Indian crime articles

V. 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;

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