Gun Violence Prediction Using Machine Learning Models (LLMS)

Sarthak Sethi

***Abstract*—This paper studies various machine learning methods to forecast the pattern or trend of gun violence in the United States. Some models used in this study are Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest; the latter is the most robust. Further, an attempt was made to utilize transformer-based models like BERT, though the limitation of GPU was not allowed to go further. Preprocessing data included extensive cleaning and integration from multiple Kaggle datasets to ensure reliability. Preliminary findings hint that predictive modeling might present some promise in pinpointing high-risk factors and informing proactive interventions despite challenges in computational resources and visualization techniques.**

Keywords: Gun violence, machine learning, predictive analytics, Logistic Regression, Decision Tree, SVM, Random Forest, BERT, data preprocessing, feature selection, hyperparameter tuning, model evaluation, data integration, visualization challenges.

I. Introduction

GUN violence in the United States is a critical public health and safety concern. At the end of September 2024, there had been over 385 mass shootings-a little more than one each day. The predictive modeling performed in this space will yield useful insights into possible risk factors and also proactively inform public policy measures toward community safety. This project uses a dataset of U.S. mass shootings through prior years to find patterns and make predictions by training machine learning models on the data; it is hoped that the output might provide a better understanding of what factors-both core and peripheral-lay at the roots of gun violence.

II. Proposed Methodology

This research adopts a multi-phase approach to predict gun violence trends using the following steps:

- **Data Collection:** The dataset, compiled from multiple Kaggle sources and enriched with external information, includes detailed records of U.S. mass shootings up to 2024. It features key attributes such as:
  - **Incident details:** Incident IDs, dates, states, cities, total fatalities, injured counts, total victim counts, and descriptive summaries.
  - **State-specific gun laws:**
    - Minimum age to purchase a handgun and long gun.
    - Whether open carry is permitted.
  - **Demographic and geographic data:**
    - Population of the incident location.
    - Latitude and longitude for spatial analysis.
  - **Temporal features:**
    - Years, months, and days were extracted to assess trends over time (e.g., seasonal effects).

  This dataset offers a comprehensive view of incidents with structured and enriched information, making it suitable for analyzing patterns, evaluating regional and temporal trends, and exploring potential predictors related to mass shootings in the U.S.
- **Data Preprocessing:** Preprocessing involved extensive cleaning for missing values and standardizing inconsistent formats in many datasets. Normalization of column names was performed, along with the removal of irrelevant features. New information was added by enriching columns, splitting them when sourced from external data, including state gun laws, the minimum age to purchase handguns and long guns, and geographical latitude and longitude for regional analysis. Temporal features, like years, months, and days, were extracted from date fields to check for seasonal and yearly trends that could be influential. This allowed for a deeper look into the data. The data was split into training and testing sets to provide a realistic model evaluation. Techniques such as GridSearchCV were applied for the tuning of model hyperparameters to optimal performance.
- **Model Selection and Training:** Various machine learning algorithms were chosen for their suitability in classification tasks:
  - **Logistic Regression:** Useful for binary classification, offering insights into factors influencing gun violence.
  - **Decision Tree:** Provides interpretable decision-making paths for identifying critical features.
  - **Support Vector Machine (SVM):** Useful for complex data relationships, adding robustness to predictions.
  - **Random Forest:** Demonstrated robust performance with high accuracy and precision. Hyperparameter tuning further improved its performance.
  - **BERT:** Bidirectional Encoder Representations from Transformers, an advanced transformer-based model was attempted to analyze textual features, though GPU limitations hindered full implementation.

  Note: K-Nearest Neighbors (KNN) was initially considered but dropped due to inefficiencies with large datasets.
- **Evaluation Metrics:** Models were evaluated using accuracy, precision, recall, and F1 score to assess prediction performance.

III. Challenges and Solutions

- **Data Imbalance:** Can decrease the overall precision of the models; it typically means fewer instances of incidents with higher severities in most situations. Some possible techniques for overcoming this, such as resampling and the generation of synthetic data, are being explored.
- **Data Quality Issues:** Partial records needed preprocessing to improve data quality for better input to the model. Inconsistent formats and missing values in `.csv` files led to significant preprocessing challenges, including manual edits using Modern CSV on MacOS and conversions of data types (e.g., `int` to `float`).
- **Performance Trade-offs:** High accuracy often required significant computational resources, presenting a challenge. Hyperparameter tuning and cross-validation were used to optimize performance.
- **Computational Limitations:** GPU constraints hindered the training of BERT and required purchasing Colab GPU credits.
- **Visualization Problems:** Heatmaps failed to render properly, even after troubleshooting with language models (LLMs).

IV. Related Work

Various works have developed predictive models for different aspects of public safety, including gun violence. For example, Swedo et al. (2023) apply machine learning techniques to estimate U.S. firearm homicides in near real-time using a two-phase pipeline that combines optimal models into a stacked ensemble for accurate weekly estimates. A predictive policing framework based on the Random Forest algorithm for urban crime by Jain and Jain 2022 will go a long way in developing safe smart cities. Further, Wheeler and Steenbeek (2020) demonstrated that Random Forests can yield accurate, long-term crime forecasts for micro-places and significantly improve interpretability thanks to its advanced model summaries. This has been built on a body of research, targeting gun violence especially, by using a wide range of algorithms including Random Forest, attempts to utilize transformer models like BERT. This, in addition to the extensive dataset we use, will enhance the prediction and give a fuller knowledge of what contributes to gun violence.

V. Results and Analysis

The performance of different machine learning models was analyzed using various metrics, including accuracy, precision, recall, and F1 scores. Additionally, BERT's performance on test and unseen datasets was explored using predictive distributions and training loss trends.

A. Model Comparison

- **Random Forest**: Achieved **90.60% accuracy**, outperforming other models in most metrics. Weighted average precision, recall, and F1 scores were all **91%**, showcasing its reliability for classification tasks.
- **Decision Tree**: Competitive performance with an accuracy of **90.60%**, comparable to Random Forest. High interpretability makes it suitable for actionable insights in high-risk cases.
- **Logistic Regression**: Lower accuracy at **40.13%**, but with **high recall** for certain critical classes, beneficial for identifying high-risk incidents.
- **Support Vector Machine (SVM)**: Performed below expectations, achieving **39.81% accuracy**. Despite showing high recall for some scenarios, overall performance was inconsistent.
- **Dropped Model - KNN**: K-Nearest Neighbors was excluded due to inefficiency with large datasets and poor performance during preliminary testing.

B. Insights from BERT Predictions

BERT was applied to both test and unseen datasets, demonstrating robust prediction capabilities despite computational constraints. Key findings are illustrated in Figure 1:

- **Prediction Distribution**: Predictions for the unseen dataset showed high concentration around 1.0, which signifies very high confidence in most samples. In contrast, the test set showed a more spread-out distribution with wider variations in predictions.
- **Prediction Ranges**: The prediction range for the unseen dataset had relatively small boxplot ranges compared to the test set. It contained fewer outliers compared to the test set.
- **Sample-Specific Predictions**: Sample-Specific Predictions: The scatterplots show consistent prediction trends for the unseen dataset across the indices, with minimal deviation. The test set had a bit more variability, which was expected given its wider distribution of predictions.
- **Training Loss**: The training loss decreased significantly over three epochs, indicating effective convergence and improved performance with each iteration.

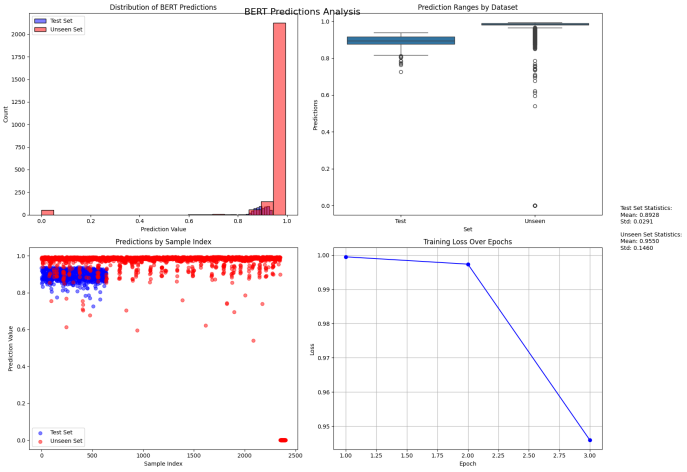

Fig. 1. BERT Predictions Analysis: Distribution, Prediction Ranges, and Training Loss.

C. Visualizations

- **Heatmaps**: Illustrated comparative performance of precision, recall, F1 score, and accuracy across all models (see Figure 2).
- **Bar Plots**: Highlighted accuracy comparisons, with Decision Tree and Random Forest achieving the highest scores. Detailed class-specific performance metrics (precision, recall, and F1 scores) further emphasized the strength of Random Forest.
- **BERT-Specific Visualizations**: Distribution and range of predictions were depicted through histograms and boxplots, contrasting the test and unseen datasets (Figure 1).
- **Precision-Recall Analysis**: Showcased performance for each label across models, reinforcing the strengths of Random Forest and Decision Tree in balancing these metrics.

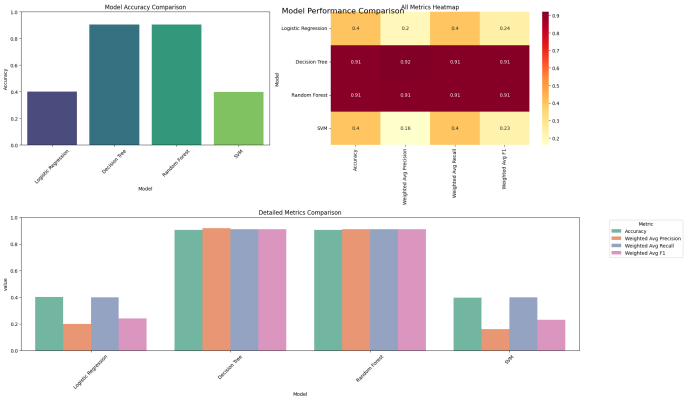

Fig. 2. Heatmap of Precision Scores Across Models and Classes.

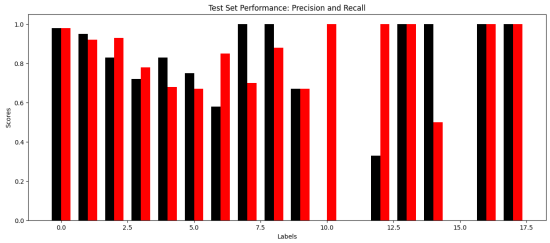

Fig. 3. Bar Plots Comparing Model Accuracies.

VI. Conclusion

This paper does a comparative study of various machine learning models for the prediction of gun violence in the United States. The results here show how high performance and interpretability can be achieved with more advanced models such as Random Forest and Decision Tree, thus being reliable for actionable insights. While models such as Logistic Regression and SVM proved useful in certain scenarios, they are less consistent in their overall performance.

After correlating the textual and numerical data by integration of BERT, further promise was extracted from that too, enhancing the case of hybrid approaches. Not very reassuringly, though, other challenges were noticed in their method regarding variability of test set predictions and with computational resource availability.

In the future, further efforts will be directed at enhancing preprocessing techniques, considering ensemble methods to increase model robustness, and augmenting the dataset to increase generalizability. Addressing these areas will help predictive modeling be an even stronger tool for proactive interventions in mitigating gun violence.

VII. Sources

1) Swedo, E. A., Alic, A., Law, R. K., Sumner, S. A., Chen, M. S., Zwald, M. L., Van Dyke, M. E., Bowen, D. A., & Mercy, J. A. (2023). Development of a machine learning model to estimate US firearm homicides in near real time. *JAMA Network Open, 6*(3), e233413. https://jamanetwork.com/journals/jamanetworkopen/
2) Jain, R., & Jain, S. (2022). Predictive policing in urban environments using random forest framework for safer smart cities. *IEEE Xplore*. https://ieeexplore.ieee.org/document/10723873
3) Wheeler, A. P., & Steenbeek, W. (2020). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology, 36*(3), 445–469. https://link.springer.com/article/10.1007/s10940-020-09457-7