National College of Ireland

Crime Category Classification in San Francisco using Machine Learning Techniques

MSc Research Project
MSc Data Analytics

Jerin Eldho
Student ID: x21196737

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jerin Eldho
Student ID:	x21196737
Programme:	MSc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	18/09/2023
Project Title:	Crime Category Classification in San Francisco using Machine Learning Techniques
Word Count:	9,544
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jerin Eldho
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Crime Category Classification in San Francisco using Machine Learning Techniques

Jerin Eldho
x21196737

Abstract

This research addresses the challenge of accurately categorizing crime incidents based on textual descriptions, aiming to enhance crime analysis and law enforcement strategies. The study investigates the effectiveness of various machine learning models, including LSTM, GRU, Logistic Regression, SVM, and Random Forest. Through comprehensive evaluation, the Logistic Regression, SVM, and Random Forest models exhibited remarkable accuracy, achieving an average of around 99.91% to 99.97% across categories. This work contributes by showcasing the potential of simpler algorithms in crime categorization tasks, highlighting their reliability and effectiveness. The results are consistent with the state of the art, stressing the importance of customized algorithm selection and feature representation. In reality, these models provide precise crime classification information for better public safety and law enforcement decisions. However, issues in differentiating complicated criminal categories remain, suggesting future study possibilities.

1 Introduction

The importance of utilizing machine learning algorithms to categorize crimes based on textual descriptions was addressed in this research. The study's goal was to assess how well these algorithms categorize crimes with accuracy, which would advance the fields of crime analysis and law enforcement. It is essential for law enforcement organizations to comprehend and classify crimes based on their textual descriptions. But traditional review of a high volume of crime reports might take a while and be subject to human error. The project's objective was to categorize the crimes according to the description of the crimes using various machine learning algorithms and methods for natural language processing that could swiftly and reliably detect crimes while also saving time.

1.1 Motivation and background

Crime Category Classification in San Francisco using Machine Learning Techniques was motivated by the pressing need to effectively analyze and categorize crimes in the United States. Crime classification plays a crucial role in law enforcement agencies' efforts to understand crime patterns, allocate resources efficiently, and devise targeted strategies to combat criminal activities. Still, the traditional manual categorization of crimes based on textual descriptions is time-consuming, subject to human error, and often lacks consistency. According to the problem's history, the large volume of crime data created across

the United States. Law enforcement entities, such as police departments and government agencies, generate and manage massive databases of textual crime reports. These descriptions, often rich in detail, provide valuable insights into the nature and characteristics of different criminal incidents. Though, the manual analysis of these descriptions poses significant challenges due to the sheer volume of data and the need for consistent and accurate categorization.

To address these challenges, the research prroject explored the potential of Natural Language Processing (NLP) techniques, along with other machine learning models, in automating the crime categorization process. The motivation behind employing NLP and machine learning models stemmed from their ability to capture the sequential and contextual information present in crime descriptions. By leveraging these techniques, the research aimed to develop a robust and accurate system that could automatically categorize crimes based on their textual descriptions, providing law enforcement agencies with a more efficient and reliable means of crime analysis. The completion of this research project contributes to the field of crime analysis in the United States by offering an advanced solution to the problem of crime categorization. By utilizing machine learning algorithms and NLP techniques, the research provides a method for analyzing large volumes of textual crime data in a time-efficient and consistent manner. The developed system not only enables law enforcement agencies to categorize crimes more accurately but also empowers them to gain deeper insights into crime patterns, identify trends, and allocate resources effectively.

The background and completion of this research underscore the importance of embracing technological advancements in the field of crime analysis. By leveraging NLP and other machine learning models, law enforcement agencies in the United States can enhance their decision-making processes and develop data-driven strategies to tackle crime more effectively. The research serves as a foundation for further advancements in automated crime categorization systems, offering potential avenues for improving public safety, resource allocation, and proactive crime prevention efforts across the nation.

1.2 Research Question

RQ: *How well can machine learning algorithms categorize crimes based on textual descriptions?*

It can inform the development and implementation of automated systems that streamline the crime analysis process. Moreover, they can provide valuable guidance to law enforcement agencies in leveraging machine learning algorithms to enhance their capabilities in categorizing crimes based on textual information, leading to improved resource allocation, more targeted strategies, and a deeper understanding of crime patterns.

1.3 Research Objectives and Contributions

The objectives mentioned below were derived to address the research question.

1. Critically evaluate the significance of text classification in crime analysis conducted by other researchers in the field of crime analysis.

2. The research aimed to collect a substantial dataset of crime reports containing textual descriptions.

3. To implement and evaluate multiple machine learning models to compare the performance of various models and assess their effectiveness in accurately categorizing crimes based on textual descriptions.

4. To apply and evaluate machine learning techniques, specifically LSTM and GRU, known for their effectiveness in handling sequential data, to enhance the analysis of crime descriptions. The objective was to leverage the sequential and contextual nature of textual data to improve the accuracy of crime categorization.

5. To perform comparison of the developed model mentioned in objectives 3 and 4.

Major Contribution: By employing NLP and machine learning methods, law enforcement agencies could easily categorize crimes with greater accuracy and efficiency compared to the traditional manual process. Thus the drawbacks of manual categorization, which is time-consuming, prone to human error, and lacks consistency can be avoided.

Minor Contribution: Minor contributions include the visualization of data after the exploratory data analysis.

The remaining sections of this technical report are structured as follows: Chapter 2 provides an overview of previous research on Crime Classification utilizing machine learning algorithms. Additionally, chapter 3 presents the modified CRISP-DM methodology approach. Chapter 4 covers the implementation, evaluation, and results of the machine learning algorithms. Chapter 5 presents the Discussion. Lastly, chapter 6 presents the final conclusion based on the obtained results and suggests directions for future research.

2 Literature Review of Crime Category Classification using Machine Learning Techniques

2.1 Introduction

Crime analysis is essential to law enforcement agencies' efforts to understand and effectively address crime patterns, allocate resources efficiently, and develop proactive crime prevention programmes Brown and Ballucci (2007). For a considerable amount of time, the United States has regularly monitored crime rates, which include common property crimes, violent crimes, and white-collar crimes. Despite crime rates, The United States has seen changes in crime rates throughout time, influenced by socioeconomic circumstances, demographics, law enforcement techniques, and sociological variables. These changes often reflect regional variations. Law enforcement organisations and decision-makers continue to place a high priority on addressing and preventing crime in the nation. This section looks into many aspects of categorising crimes and other text-based classifications made with the aid of several machine learning methods. Nasr et al. (2020) comprehensively covered traditional crime analysis approaches, machine learning techniques for text classification, natural language processing (NLP) methods applied to crime analysis, evaluation metrics for crime category categorization, and relevant research on automated criminal analysis systems in their study.

2.2 Significance of Text Classification in Crime Analysis

Text classification holds immense significance in crime analysis as it enables law enforcement agencies to effectively categorize and analyze crime incidents based on their textual descriptions. By accurately categorizing crimes, agencies can gain valuable insights into crime patterns, modus operandi, and emerging trends. This information allows them to allocate resources efficiently, deploy preventive measures, and develop targeted strategies to combat crime. Text classification also facilitates data-driven decision making, enabling agencies to identify correlations between different types of crimes and evaluate the effectiveness of existing interventions. Moreover, standardized crime categorization promotes collaboration and information sharing among law enforcement agencies, fostering a proactive and coordinated approach to crime management. By enabling the extraction of useful insights from unstructured textual data sources like incident reports, social media posts, news articles, and witness statements, text categorization plays a crucial role in crime analysis. In the article published by Bradley and Waller (2017), crime-related text data frequently contains important information regarding methods, motivations, suspects, and contextual factors that can help with correct categorization of crimes. Abdulrahman and Alkhader (2017) conducted a study on crime prediction in San Francisco using both the KNN and Na¨ıve Bayes classifiers. Their approach involved comparing the performance of these two classifiers. For the KNN classifier, they employed two different techniques: uniform and inverse. The Na¨ıve Bayes classifier, on the other hand, utilized Gaussian, Bernoulli, and Multinomial techniques. The findings revealed that the KNN classifier exhibited poor performance due to longer execution time during the classification and regression stage. The Na¨ıve Bayes Gaussian technique yielded unsatisfactory results, indicating that the data used was discrete rather than continuous. Conversely, the Na¨ıve Bayes Bernoulli and Multinomial techniques demonstrated better performance among the proposed approaches. It is worth noting that these techniques were directly applied to the training dataset without considering potential errors or outliers. As per Lyons (2016), Multimodal analysis can offer an all-encompassing perspective of a crime, improving the precision of category classification. Through social media research, text classification can assist law enforcement agencies in seeing new criminal patterns, tracking public opinion, and spotting potential threats. This proactive strategy helps with resource allocation and crime prevention. In order to speed up investigations and court cases, text categorization can help find pertinent evidence, hints, and witness testimony from enormous volumes of text data. By analyzing the language and content of crime-related text, insights into criminal behavior, motives, and patterns can be gained, contributing to a deeper understanding of crime dynamics. Text classification allows for the integration of structured and unstructured data, leading to a more comprehensive and accurate crime analysis. This fusion can improve the overall quality of crime category classification. John et al. (2021) discussed the issue of women's safety in major cities in India, which have been identified as some of the most unsafe places for women over the past decade and to provide a solution that helps women select states to travel to or relocate based on recent criminal activity. As per the author classifying news articles into categories is considered a multi-class classification problem, requiring a robust and powerful machine learning model. Another study by Shabat et al. (2014) combined Naïve Bayes, Support Vector Machine, and KNN classifiers using a weighted voting ensemble method. They created a new corpus of crime data and manually labeled it for training and testing purposes. The researchers tested this hybrid model on a dataset acquired from

the Malaysian National News Agency. The model's performance was evaluated using the manually annotated dataset, and the results of the experiments demonstrated promising outcomes. The findings demonstrated significantly improved performance, with an F-measure of 89.48% for identifying crime types and 93.36% for crime-related entities. Qi (2020) did a study on classification of theft crime data from a city from 2009 to 2019 using text classification technology. TF-IDF model was used to extract the features and for the data pre-processing.As per this researcher, the combination of TF-IDF and XGBoost is an effective text classification algorithm for theft crime data.

Support Vector Machine (SVM) is one classification algorithm that can be utilized to group news articles according to Rahmat et al. (2021).In this study, SVM is employed to classify news articles into five categories associated with ITE Law violations in Indonesia. The proposed classification system, employing the RBF kernel, achieves an accuracy rate of 86.27%. The implementation of Support Vector Machine serves as the chosen algorithm for effectively classifying news articles in this research. In the research conducted by Iqbal et al. (2013) used the classification, a supervised learning technique in data mining, to analyze a crime dataset and predicted the "Crime Category" for various states in the United States. The dataset used in the research was derived from socio-economic data, law enforcement data, and crime data collected from various sources. The study compared two classification algorithms, Naïve Bayesian and Decision Tree, and finds that the Decision Tree algorithm performs better, achieving an 83.95% accuracy in prediction of crime categories for different states in the USA.

Mantoro et al. (2022)have taken advantage of social media platform to analyze and understand various aspects of society, including competitive business strategies, decision-making processes, and predictive support systems. In their study, the focus was on content from Twitter and Facebook, where users often share information related to crimes that require police attention. The main objective of their research was to detect crime rates on social media and identify patterns and trends in the number of crime-related tweets. The researchers employed a text mining approach to classify the content of tweets and posts into ten different crime categories. The classification algorithms used for this task included Logistic Regression, Naive Bayes, Support Vector Machine (SVM), and Decision Tree. Upon analyzing the results, it was found that Logistic Regression achieved the highest accuracy, reaching 90%. This finding suggests that Logistic Regression is particularly effective in classifying crime-related content.

By Sundhara Kumar and Bhalaji (2016) classification methods of data mining is widely used in crime data analysis due to the global increase in crime rates. Their study focused on prediction of a crime whether its violent or non-violent using classification methods. The classification algorithms such as Gradient Boosting and Random Forest was used to determine the nature of a crime and evaluates their accuracy, precision, and recall values based on crime records. Experimental results indicated that those algorithms performed better than the k-Nearest Neighbor method, previously considered one of the best classification methods.
Thus the objective 1 has been accomplished.

2.3 Comparative Study of Methods Employed in Natural Language Processing

The article published byNadkarni et al. (2011), focuses on the NLP approach where due to its capacity to process and analyse unstructured textual data, such as that found in

event reports, social media posts, and news stories, Natural Language Processing (NLP) approaches have become increasingly popular in the classification of crime categories. In this article, we compare many NLP techniques used in crime analysis and classification. In continuous vector spaces. **?**,their research showcases the effectiveness of word embeddings in crime categorization using a dataset of 15,000 Italian news articles. Word embeddings like Word2Vec and GloVe are frequently used to represent words as dense vectors. By capturing the semantic relationships between words, these embeddings allow models to comprehend context and meaning. Word embeddings have been used to represent texts that deal with crimes, enabling machine learning models to learn contextual information and increase the accuracy of categorising crimes. NLP tasks have been transformed by pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) which has been mentioned in the article published by Gillioz et al. (2020). These models can be honed for particular tasks like the classification of crimes into different categories since they gain contextual knowledge from vast amounts of text data. They provide cutting-edge performance because they can capture intricate contextual interactions. Ensemble approaches increase classification performance by combining the predictions of various models. In the domain of NLP for crime investigation, integrating several NLP techniques has been investigated to increase accuracy and resilience. One example is mixing word embeddings with LSTM or CNN models. Many ways have been used to handle and analyse text data in the area of crime category classification utilising NLP techniques. Each approach has its own advantages and disadvantages, and the best one to use will frequently rely on the type of data being used and the precise specifications of the categorization task. To find the best method for applying NLP to categorise crimes, more investigation and testing are required. In Bai (2018) stated a text classification strategy that used attention mechanisms and Long Short-Term Memory (LSTM) networks in his study. The study started by outlining the importance of text categorization in Natural Language Processing (NLP) and noting the shortcomings of classic techniques such as Bag-of-Words and n-gram models. The author then introduced the LSTM network and its application in text categorization. LSTM networks, as recurrent neural networks (RNNs), possess the ability to capture long-term dependencies in sequences. Next, attention mechanisms were introduced as a means to focus on specific inputs during prediction generation. The study employed soft attention, assigning weights to each input token based on its relevance to the current prediction. The author described the experimental setup, including the utilization of the 20 Newsgroups dataset consisting of approximately 20,000 items distributed across 20 categories. The proposed model's performance was compared against several baseline models, including SVM, Naive Bayes, and a neural network using Bag-of-Words representation. The results indicated that the suggested model surpassed all baseline models in terms of accuracy, precision, and recall. Furthermore, the author conducted an ablation study to analyze the individual contributions of each component in the proposed model, demonstrating the significance of the attention mechanism and LSTM network in achieving high performance. The paper provided a comprehensive evaluation of the proposed model, showcasing its superiority compared to baseline models.Overall, the research findings suggested that employing LSTM networks and attention mechanisms in text categorization can achieve state-of-the-art performance on standard benchmarks. The suggested model exhibited promising results, outperforming conventional models with an accuracy rate of 97.02%, as reported by the author's comparison.

In a research conducted by Haider et al. (2022), the classification of crime using data analytics and machine learning techniques was explored. The researchers acknowledged the increasing prevalence of criminal activity and emphasized the importance of an efficient crime classification system. To address this, the authors proposed a technique that employed machine learning algorithms and data analytics to categorize crimes into distinct groups. The study focused on dividing crimes into three primary categories: property crime, violent crime, and drug-related crime. Data from sources such as police reports, crime statistics, and relevant papers were collected by the authors for analysis. To ensure data quality, noise, outliers, and missing values were removed through data preprocessing. Various feature extraction approaches were then employed to extract meaningful characteristics from the data. For crime categorization, the researchers utilized the Random Forest and Support Vector Machine (SVM) classification techniques. The performance of the proposed approach was evaluated using metrics such as accuracy, recall, and F1-score. The results indicated that the Random Forest method outperformed the SVM technique in terms of accuracy, precision, recall, and F1-score. Particularly, the accuracy of the system was determined to be 93.2% for the Random Forest method and 89.4% for the SVM algorithm. Overall, the study presented a potential approach for crime categorization by employing data analytics and machine learning algorithms. The use of Random Forest and SVM algorithms proved to be accurate and effective in the classification of crimes. Although, the researchers acknowledged certain limitations of the study, including the need for a larger dataset and further research to enhance the system's functionality. The primary objective of Haider et al. (2022)'s research was to develop a precise and effective criminal classification system for law enforcement organizations. The anticipated benefits included improved accuracy in crime classification and enhanced responsiveness to criminal situations.

Yang et al. (2019) proposed a convolutional gated-recurrent-unit (GRU) neural network to identify malicious URLs. According to their findings, the GRU neural network effectively learned the distinguishing features of malicious URLs and demonstrated high accuracy in classifying them. The experimental outcomes indicated that the GRU neural network is particularly well-suited for precision-oriented classification tasks. The text emphasizes the significance of utilizing deep learning for URL classification. Deep learning's ability to discern intricate patterns in data makes it particularly suitable for challenging tasks like identifying malicious URLs, where conventional methods may struggle. Their research represents a valuable addition to web security. The proposed GRU neural network shown great promise in accurately detecting malicious URLs.

2.4 Identified Gaps in the Research

While there have been tremendous advancements in the study of applying NLP techniques to classify crimes according to their types, there are still many gaps and difficulties. Finding these gaps is essential for directing future study and enhancing the efficacy of crime analysis. The majority of research only use textual information to categorize crimes. However, adding additional media from crime scenes, such as photos, films, and audio recordings, may offer a more thorough picture of crimes. A more precise and reliable classification of crime categories may result from the integration of multimodal data. The thought by Al-Ghamdi et al. (2023), although pre-trained language models like BERT have achieved outstanding results in a variety of NLP tasks, their performance in crime analysis may be constrained by the absence of domain-specific fine-tuning. Creating

language models specifically for crimes could capture nuances and terminology unique to the topic and improve classification accuracy. Deep neural networks are one example of an advanced NLP method that lacks interpretability. Building trust in automated crime category classification systems requires the development of techniques for explaining the outcomes of complex models. While going through the article presented by Imamguluyev (2023), understanding the foundation of forecasts can also help law enforcement For NLP tasks, there are still few high-quality, varied, and sizable crime-related datasets available. Fair comparisons between alternative models and methodologies would be made easier by creating standardised benchmark datasets that encompass a variety of crime categories, contexts, and data sources .Class imbalance, or the considerable under representation of some crime categories, is a common problem in crime databases. Avoiding biased models that perform well for majority classes but poorly for minority classes requires the development of appropriate ways to handle imbalanced data. Regional differences exist in crime trends, linguistic preferences, and reporting conventions. Models developed using data from one area may not translate well to another. According Carter et al. (2020), the applicability of crime category classification models could be improved by incorporating strategies for domain adaptation and cross-region generalisation . There is a need for real-time crime category classification systems, even though some studies concentrate on retroactive crime analysis. Law enforcement authorities may be able to act quickly to prevent and respond to crimes by developing technologies that can process and classify text linked to crimes in real time The areas that need additional research and innovation are highlighted when these gaps in the classification of criminal categories are found utilising NLP approaches. Closing these shortcomings will result in crime analysis systems that are more reliable, resilient, and context-sensitive.

2.5 Conclusion

A study conducted by Shojaee et al. (2013), the use of NLP approaches in the classification of crime categories has demonstrated potential developments in automating crime analysis, upgrading law enforcement tactics, and boosting public safety. Through this analysis, some significant gaps and difficulties in the current research landscape have been discovered, paving the path for further studies. Incorporating many data sources, such as photographs and videos, into multimodal analysis is still an untapped field that could provide a more thorough knowledge of crimes and their settings. Creating domain-specific language models suited to criminal analysis could result in more precise and contextually aware classification of crime categories. There is still a critical demand for interpretable models in crime analysis. Building trust and promoting the use of complicated NLP models in law enforcement require an understanding of their judgements. Fair assessments and comparisons depend on the availability of high-quality, standardised benchmark datasets that reflect different crime categories and sources . To avoid biased models and guarantee correct classifications across all crime categories, effective solutions for handling imbalanced data are essential .Furthermore, overcoming the difficulty of cross-region generalisation is crucial for creating models that may be used in a variety of geographical contexts. Innovative technologies that analyse and classify crime-related material quickly are required to make the move from retrospective analysis to real-time crime categorization, enabling proactive law enforcement and crime prevention activities. The study conducted by Stalidis et al. (2021), proves the future of crime category classification using NLP techniques will be shaped by filling in these identified gaps as the

field develops, encouraging the creation of more precise, strong, and context-aware crime analysis systems.

3 Research Methodology and Design Specification

3.1 Introduction

This section explains the modified approach of CRISP-DM methodology of crime category classification and the two layer architectural design which is also implemented as part of this research.

3.2 Research Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely accepted and established approach in the fields of data mining and machine learning. For this research project, a modified version of the CRISP-DM approach, as shown in Figure 1, is utilized to create a more structured and flexible framework that effectively deals with the complexities of crime category classification. The primary aim is to make well-informed decisions, tackle challenges efficiently, and achieve significant outcomes that benefit both San Francisco and the stakeholders involved.

In comparison to the original approach, the modified methodology comprises 10 stages, which are as follows:

Problem Definition: The research problem revolves around the ability of machine learning algorithms to categorize crimes based on textual descriptions. Additionally, it involves identifying the specific crime categories to be predicted, such as theft, burglary, etc.

Data Collection and Understanding: Crime data from San Francisco is gathered from the official crime database of the San Francisco Police Department (SFPD)[1]. This dataset covers a wide range of crime categories and is thoroughly examined to understand its structure and quality. Initial data exploration reveals potential biases and missing values, which are addressed during the preprocessing stage. The data covered a diverse range of crime categories and was examined to understand its structure and quality. Initial data exploration revealed potential biases and missing values that were addressed during the preprocessing stage.

Data Preprocessing and Integration: The collected crime data undergoes extensive preprocessing. Missing values are imputed, outliers are handled appropriately, and any redundant or inconsistent information is corrected. As the data from the source is already comprehensive, data integration is not necessary for this research.

Data Analysis and Visualization: In-depth data analysis is conducted to gain insights into the relationships between variables and identify essential features for crime category classification. Visualization techniques are employed to support decision-making during the model selection process.

Model Selection: Several machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machines, and LSTM and GRU, which are recurrent neural network architectures commonly employed in natural language processing tasks,

[1]SFPD : San Francisco Police Department

were used in this research for crime category classification. The objective is to choose the best-performing model that achieves high prediction accuracy.

Model Training and Tuning: The selected machine learning models are trained on the preprocessed data using appropriate hyperparameters. Extensive hyperparameter tuning is performed to optimize each model's performance, ensuring they are adequately generalized.

Model Evaluation and Model Interpretation:: Rigorous evaluation of the trained models is carried out using various metrics, such as accuracy, precision, recall and F1 score This evaluation allows for the identification of the best-performing model, which demonstrates superior performance in classifying crime categories. The selected models are interpreted to gain insights into their decision-making process. Feature importance analysis is conducted to understand which variables contribute most to the crime category predictions, aligning the results with domain knowledge.

Model Comparison: In this stage, the models identified in the objectives are compared to assess their relative strengths and weaknesses.

Model Deployment: The project's outputs are visually represented in a tabular format to ensure that the project goals are achieved. A general assessment of the project is also conducted. The outperformed model is then deployed for categorizing crimes.

Documentation and Final Report: The entire research process, including data preprocessing steps, model selection, evaluation results, and model interpretation, is thoroughly documented. This comprehensive documentation ensures that the research findings can be reproduced and shared with other researchers in the form of a final report.

By following this modified CRISP-DM approach, the research project aims to provide valuable insights and results that can contribute to better crime category classification and enhance the overall understanding and response to crime-related issues in San Francisco.

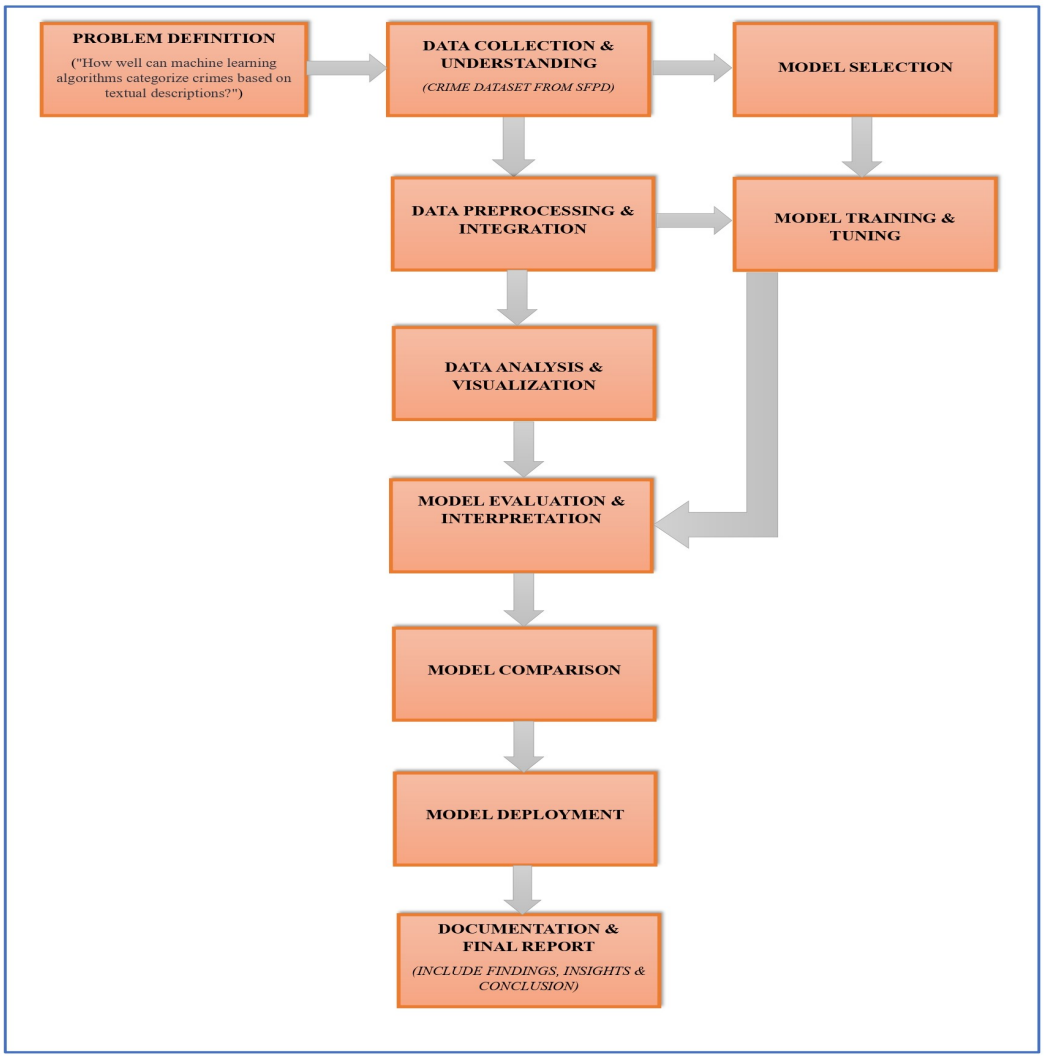

Figure 1: Stages involved in adapted Modified CRISP-DM approach

3.3 Architectural Project Design Process

Figure 2 illustrates the process of the architectural design for the project. The chosen approach was a two-tier architecture, which involved the creation of two distinct layers known as the Presentation layer and the Business logic layer. In essence, the first tier, also referred to as the Client side or front end, is responsible for the visualization and display of data and results from the back end. Python visualization techniques were employed to accomplish this task seamlessly. On the other hand, the second tier, the Business logic layer, represents the back end of the architecture. In this section, the application of several essential Machine Learning algorithms, which greatly contributed to the research, is presented. Notably, the utilized algorithms included LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), SVM (Support Vector Machine), Random Forest, and Logistic Regression. To sum up, the project's architectural design followed a two-tier approach, with the front end responsible for user interaction and data visualization through Python visualization, while the back end handled the application of essential Machine Learning algorithms like LSTM, GRU, SVM, Random Forest, and Logistic Regression.

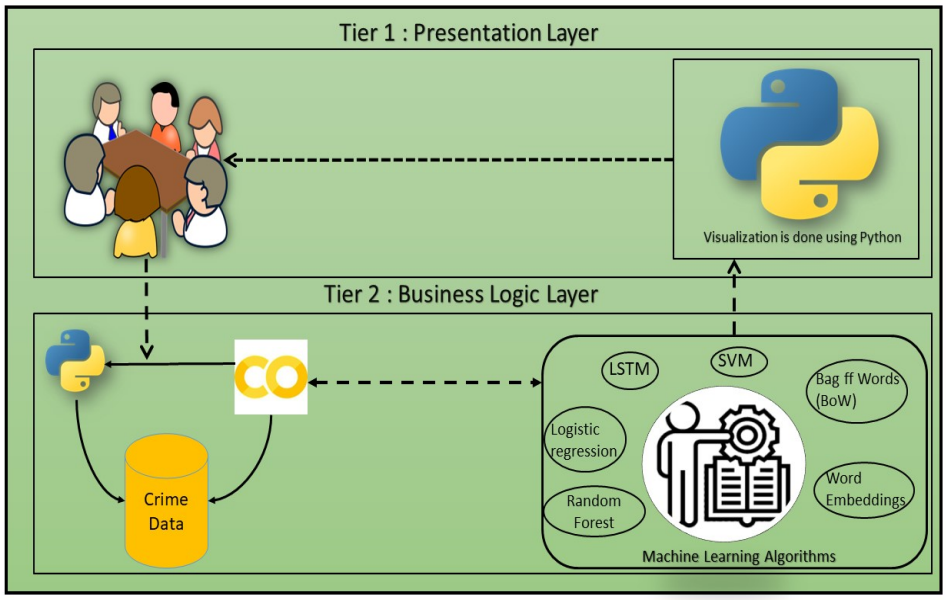

Figure 2: Project Architectural Design for Crime Category Classification

4 Crime Category Classification: Model Implementation, Evaluation, and Results

4.1 Introduction

The following section provides a full overview of the attempt involving crime category classification, going into every aspect of its implementation, evaluation, and results. The primary goal of this part is to provide a complete grasp of the complex process from establishing objectives to revealing conclusive results.

The implementation phase, which acts as the research's foundation, became absorbed in the complex process of collecting key features from the precisely curated dataset. This

dataset, carefully selected for its relevance, was analyzed in minute detail to confirm its suitability for the research aims. In the preceding part, the succeeding phases in the feature extraction process were methodically explained, establishing the framework for the subsequent stages of the classification process.The use of Python programming along with Google Colab—a dynamic and versatile platform praised for its ability to handle large datasets—was critical to the project's success. Each model was rigorously sculpted and trained using Python and Google Colab, assuring precision in model creation and permitting the application of complex approaches to the dataset. The essential essence of this part, however, shows itself in the following step—the evaluation of the implemented models. To objectively measure the performance of the constructed models, a set of generally accepted assessment criteria was deployed. **Precision, Accuracy, F1 score, and Recall** stand out as pillars, providing a comprehensive view of the models' efficacy.

Precision, as elucidated in the literature, stands as a hallmark of the model's ability to correctly identify instances within a specific crime category.

The **Accuracy** metric, a quintessential benchmark, gauges the overall correctness of predictions made by the model.

F1 score, a harmonic blend of precision and recall, encapsulates the trade-off between these two critical elements of model performance.

Recall, the fourth pillar, delineates the model's capability to successfully identify all relevant instances within a specific crime category.

This meticulous adoption of these evaluation metrics not only permits an in-depth comprehension of the models' capabilities but also establishes a solid foundation for objective comparisons between different models. By considering these multifaceted facets of performance, a holistic appraisal of the implemented crime category classification models emerges—a reflection of the intricate processes undertaken, the astute coding efforts, and the culmination of strategic choices made throughout the development journey. In the following sections, the performance of each model is meticulously discussed, analyzing the numerical findings within the context of the above mentioned evaluation measures. This method not only captures the final results but also allows for critical reflection, resulting in insights that might influence future advancements and research activities.

4.2 Analyzing Crime Data: Strategies for Pre-processing and Exploration in Category Classification

In this section, a comprehensive analysis of the crime dataset was undertaken, focusing on the strategies employed for data pre-processing and exploratory data analysis to pave the way for effective category classification. The dataset, sourced from the SPFD website[2]., spanned from 2018 to 2023 and was presented in a CSV format. With a considerable size of 260 MB, the dataset was a substantial reservoir of information, necessitating strategic steps to manage and process its content.

To facilitate efficient handling of the sizable dataset, it was first downloaded and then subsequently mounted onto Google Drive due to its substantial size. Leveraging the Python Pandas library, the dataset was seamlessly loaded into the Google Colab environment, setting the stage for a comprehensive analysis. The dataset itself comprised 35 columns and a substantial 746,052 records, reflecting the richness of the data at hand. Recognizing the sheer volume of data, an initial step was taken to curate the dataset for

[2]https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/data

relevance. Focusing on the time frame from 2022 to 2023, corresponding records were extracted using filters based on start and end years which was more than enough for the modelling. This allowed for a more manageable dataset, tailored to the research objectives.A thorough strategy was used to improve the dataset in preparation for later analysis and model creation. Unneeded columns, such as row id, incident numbers, and report type codes, were discovered and eliminated because they were deemed unnecessary for the model development. A function in Python was carefully built to replace spaces within column names with underscores to maintain uniformity and ease of analysis. This basic yet effective change improved the dataset's usability during analysis. In order to ensure data integrity, the isnull().sum() function was used to do a thorough check for null values. These null values were treated attentively to avoid any potential disruptions during subsequent analyses. Null entries in both integer and string columns were seamlessly replaced with appropriate average values, adding to the dataset's completeness.One of the pivotal components of the dataset, the "Incidentdatetime" column, was converted into a datetime format. This transformation not only facilitated time-based analysis but also ensured accurate temporal referencing throughout the study.

To summarize, the process of data pre-processing and exploratory analysis began after a lengthy process of refinement and intentional filtering. Each step, from data extraction, filtering, and transformation to null value handling and category mapping, was critical in building the groundwork for subsequent classification attempts, ensuring that the data was not only large but also relevant and suitable to fruitful analysis. The Objective 2 is thus achieved.

4.3 Data Modeling, Feature Engineering and Visualization

In this section, a thorough and systematic methodology was used to shape, refine, and graphically reveal insights from crime data. This phase provided a critical transition from raw data to a refined and enhanced foundation for future analysis and modeling efforts.

The first exploratory investigation of the target variable, "unifiedcategory," revealed a multi class categorization problem centered on the city of San Francisco's 26 unique crimes. This first analysis also found discrepancies in the frequency of crime types, leading the prudent grouping of some crime categories. This strategic mapping not only streamlined the classification process, but it also reduced the total number of crime types, making subsequent analysis and modeling more manageable. In the area of feature engineering, an organized transformation was planned to improve the models' capacity for prediction. The "IncidentDatetime" column's ability to extract complex temporal information, including the year, month, day, hour, and minute, allows for a detailed analysis of the temporal patterns present in the crime data. Further enriching the dataset and highlighting hidden trends within it were the derivation of "hour type" (morning, afternoon, evening, and night), "season" (winter, summer, fall, and spring), and the insertion of a binary "weekend" feature (1 for weekends, 0 for weekdays).

The following focus of this phase was visualization, a powerful tool for uncovering patterns and trends hidden within data. By analyzing the distribution of crime occurrences among police districts, a startling discovery emerged: the Central Police District stood out as a hotspot for high-reported offenses(Fig.3). Turning to the time component, an intuitive graphical representation identified March as the month with the highest number of recorded crimes(Fig. 4). The depiction of crime incidents across different hour types revealed "Afternoon" as the temporal window indicated by a notable surge in criminal

activity, providing further temporal detail. The distribution of crimes over the seasons was strongly correlated with a springtime uptick, suggesting that some crimes were seasonally variable (Fig. 5). Additionally, a graphical representation showed a difference between weekends and weekdays, with weekdays having higher crime rates (Fig 6). The graphic highlighted the prevalence of the "Property Crime" category, which stood tall with a noticeably high count, by showing how crimes are distributed among other crime categories(Fig. 7). This complete blend of exploratory analysis, strategic feature engineering, and perceptive visualization techniques highlights a significant transformation of raw data into a refined foundation. This base, supported by a contextual understanding of temporal, spatial, and grouping dynamics, provides as a solid basis for the subsequent phases of modeling and analysis, smoothly matching with the overall goals of this endeavor. Following the feature extraction procedure, which resulted in the addition of a suite of valuable temporal and categorical variables to the dataset, a critical step was performed to assure the preservation of this enriched dataset. The data frames, which now included the original properties as well as the precisely constructed features, were strategically exported to a new CSV file called "Processeddata."

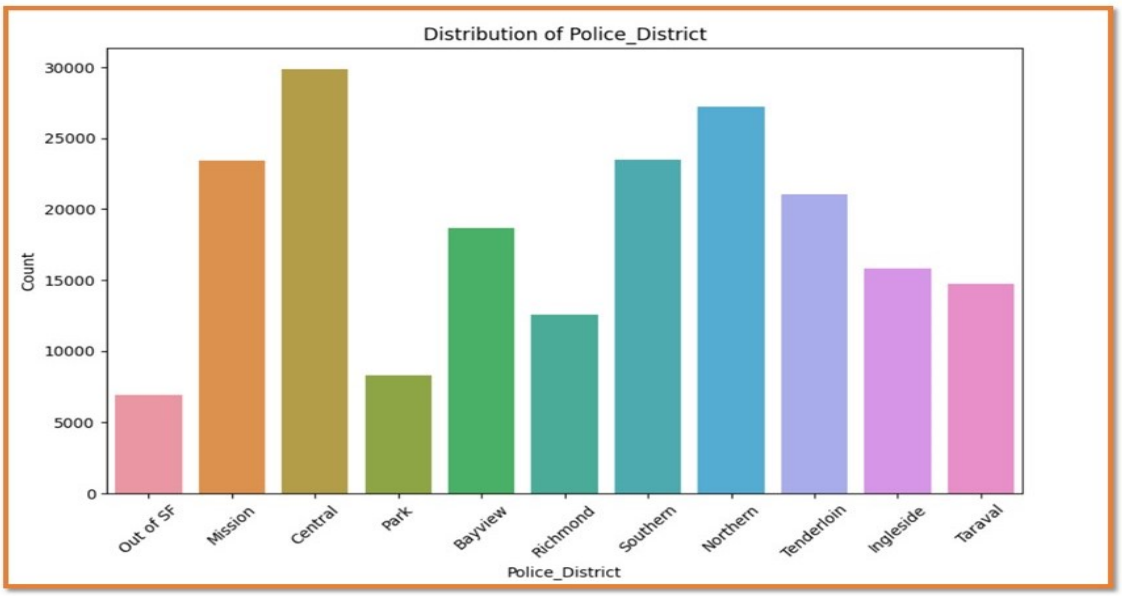

Figure 3: Distribution of Crime Occurrences Among Police Districts

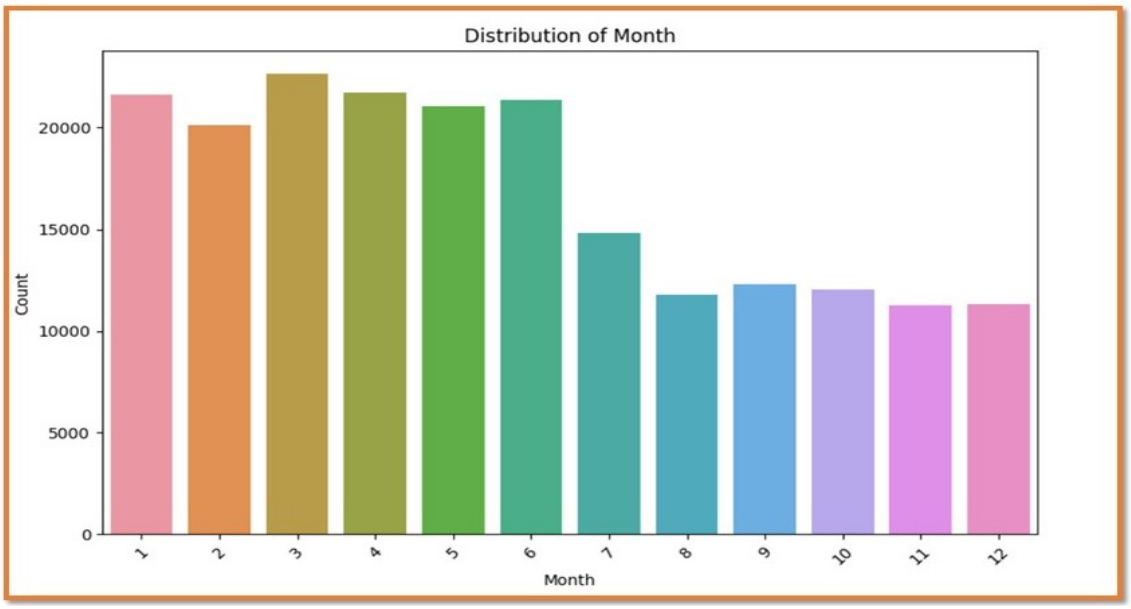

Figure 4: Crime Incidences by Month

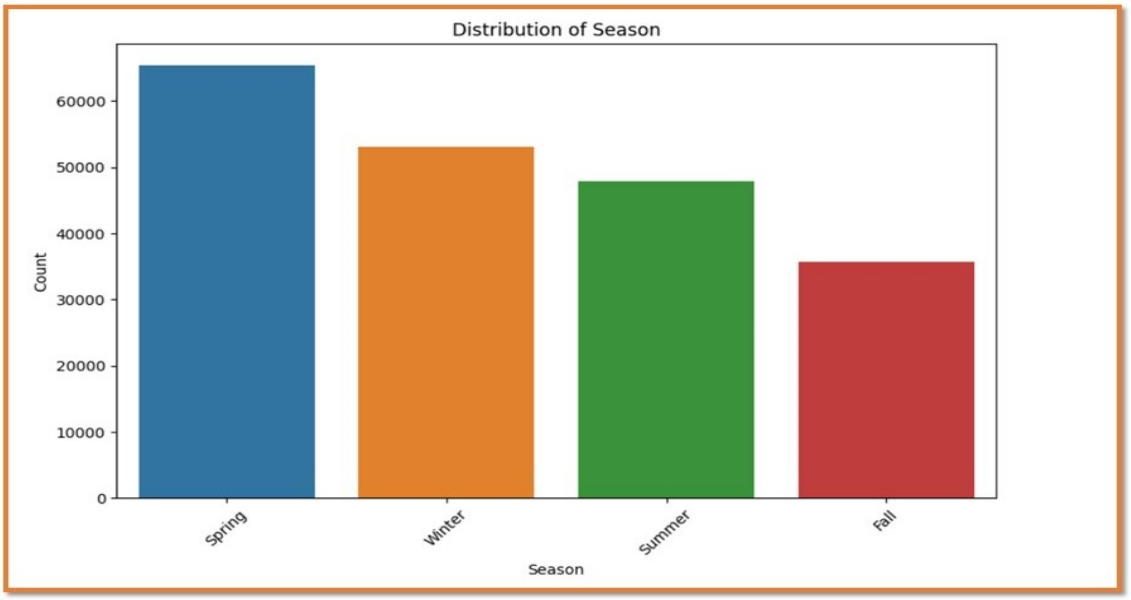

Figure 5: The distribution of crimes over the seasons

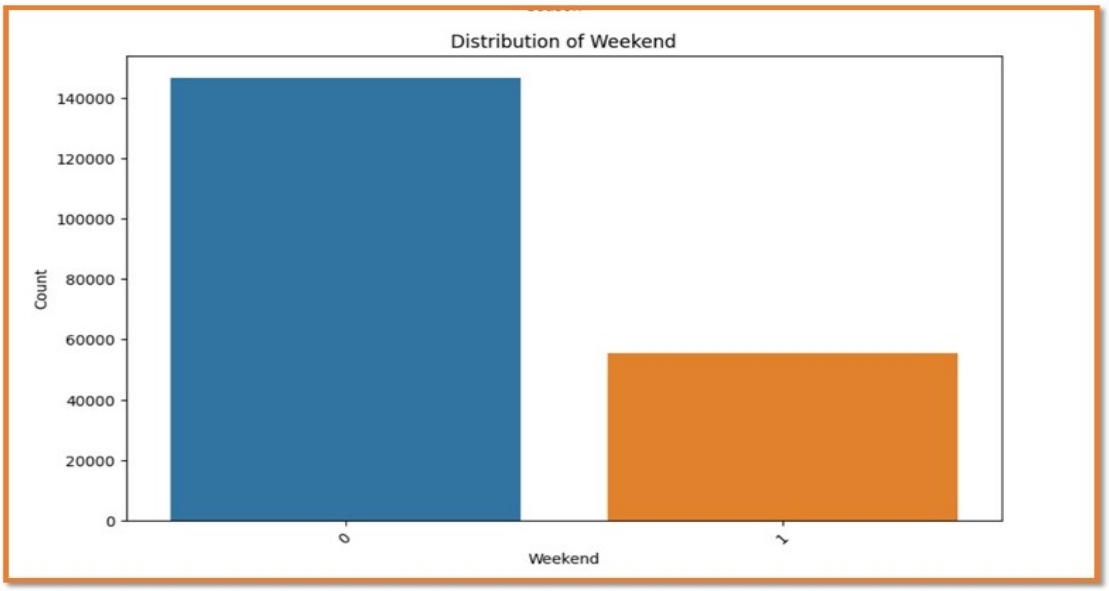

Figure 6: Comparative Analysis of Weekdays and Weekends

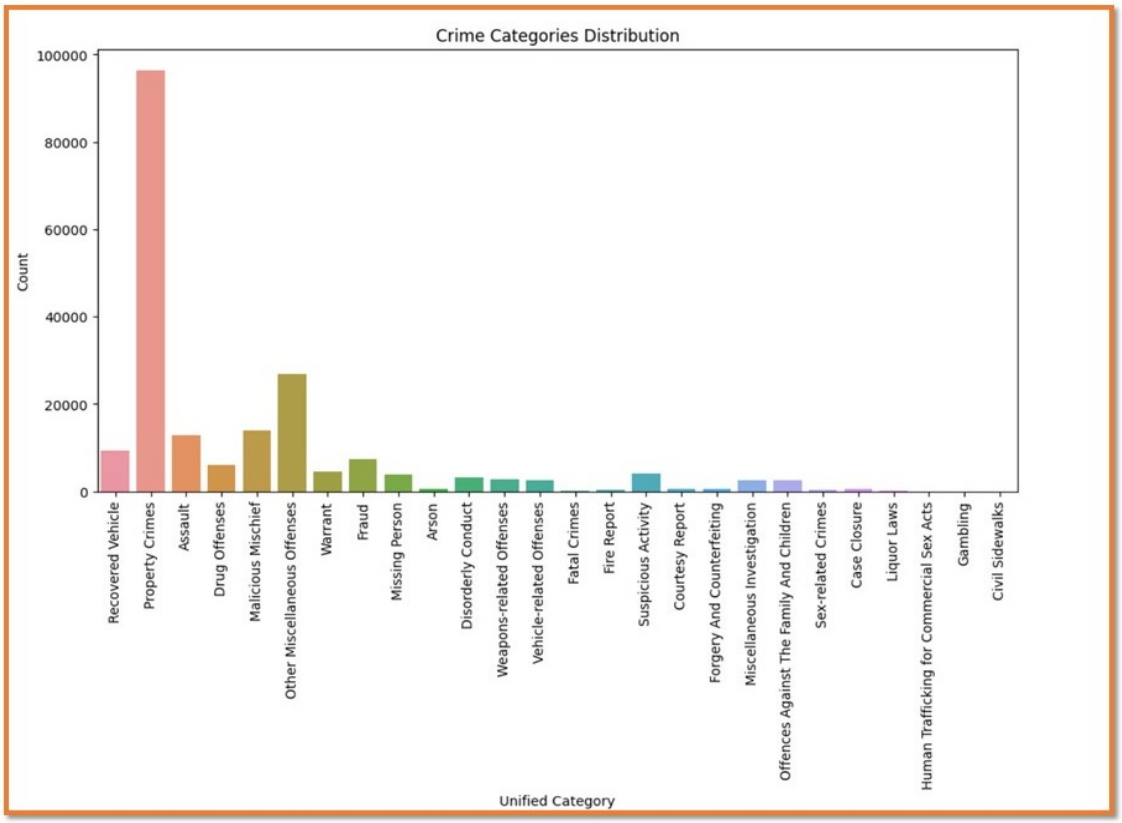

Figure 7: Distribution of Crime Categories

4.4 Implementation, Evaluation and Results of LSTM Model

The objective was to create a robust model for the precise categorization of crime incidents based on their textual descriptions. Accurate crime classification is of paramount importance as it has the potential to significantly improve crime analysis, resource allocation, and policy-making, ultimately bolstering public safety and law enforcement strategies. This approach harnesses the power of Natural Language Processing (NLP) and Sequential Modeling to address the classification challenge. In particular, Long Short-Term Memory (LSTM) networks were utilized due to their proficiency in handling sequential data, allowing for the effective processing and categorization of crime descriptions.

4.4.1 Implementation

The implementation began with the preprocessing of the dataset, entailing the conversion of crime descriptions and their corresponding categories into numerical representations. The crime descriptions were tokenized and transformed into sequences, while the categories were encoded into numerical labels. Subsequently, the Keras library[3] was utilized to craft an LSTM-based neural network. The model architecture comprised an embedding layer that learned contextual representations, succeeded by two LSTM layers that captured temporal dependencies within the sequences. The final dense layer, featuring a softmax activation function, predicted the crime category.

4.4.2 Evaluation and Result

The model's performance was assessed by dividing the dataset into training and testing subsets. The model was trained on training data before being evaluated on testing data using accuracy and loss criteria. A confusion matrix was also created to provide insight into the model's performance across several crime categories. The matrix indicated the categories that were correctly categorised as well as those that needed more attention. The model exhibited promising outcomes in the classification of crime categories based on textual descriptions. After training and evaluation, the model achieved a test accuracy of approximately 47.48% and a corresponding test loss of approximately 1.98. The confusion matrix depicted the model's performance for each crime category, illustrating areas of accuracy and potential areas for refinement.

4.5 Implementation, Evaluation and Results of GRU Model

The implementation of the Gated Recurrent Unit (GRU) into this research effort was critical due to its effectiveness in capturing long-term dependencies and sequential patterns within textual data. Using the GRU architecture, the effort attempted to improve the model's comprehension of the sequential nature of crime descriptions, resulting in more accurate and informed crime category predictions.

4.5.1 Implementation

After the preparation of the dataset, the crime descriptions and their corresponding categories were extracted from the dataset and converted into lists. Text tokenization was carried out using the Keras Tokenizer, which also involved limiting the vocabulary size

[3] An open-source deep learning framework that facilitates the creation and training of neural networks.

to 100,000 words. These tokenized sequences were then padded to ensure equal lengths, allowing for effective model training. The category labels were encoded into numerical format using the LabelEncoder class and further converted into one-hot encoded vectors. Subsequently, a neural network model was constructed using the Keras library. The architecture included an embedding layer for contextual understanding, followed by two Gated Recurrent Unit (GRU) layers. A dropout layer was introduced after the embedding layer to mitigate overfitting. The model was compiled using categorical cross-entropy loss[4] and the Adam optimizer.[5]

4.5.2 Evaluation and Results

The dataset was split into training and testing subsets for model evaluation. The model was trained on the training data for 5 epochs with a batch size of 64. The evaluation included measuring the accuracy and loss metrics on the testing data. Additionally, a confusion matrix was generated to visually represent the model's performance across different crime categories.

After training and evaluation, the GRU-based model displayed mixed results in classifying crime categories based on their textual descriptions. The model achieved a test accuracy of approximately 47.48]%, accompanied by a test loss of around 1.98. These results were consistent with those obtained in previous model implementations using LSTM layers. The confusion matrix illustrated the model's performance across different crime categories. Most categories displayed low accuracy, which was likely due to the complexity and diversity of the crime data. Despite utilizing the GRU architecture, the model struggled to distinguish between crime categories effectively.

4.6 Implementation, Evaluation and Results of Logistic Regression

Logistic Regression model was implemented for categorizing incident descriptions into unified categories using the TfidfVectorizer for feature extraction. The project sought to provide insights into the feasibility of utilizing this approach for accurate incident classification, which is crucial for enhancing public safety and law enforcement efforts.

4.6.1 Implementation

The project began by preparing the training and testing datasets. Incident descriptions and their corresponding unified categories were extracted from the datasets. The categorical labels were then encoded into numerical format using the LabelEncoder. Text data was vectorized using the TfidfVectorizer, which transformed the incident descriptions into numerical features while considering the importance of words in the dataset.

4.6.2 Evaluation and Results

The Logistic Regression model was created and trained using the vectorized text data. This model was chosen due to its simplicity and effectiveness in text classification tasks.

[4]A standard loss function for classification tasks, quantifying the difference between predicted and actual distributions.

[5]An effective optimization algorithm that adapts learning rates for parameters, enhancing training convergence and performance.

The model's performance was evaluated using accuracy as the primary metric, indicating the proportion of correctly predicted categories among all predictions. Additionally, a detailed classification report was generated, providing insights into precision, recall, and F1-score for each category.

The model achieved remarkable accuracy, with an average accuracy of approximately 99.91% across all categories. The classification report presented high precision and recall values, confirming the model's ability to accurately predict various unified categories. While some labels had a limited number of predicted samples, overall performance was exceptional, highlighting the effectiveness of the TfidfVectorizer [6] and the Logistic Regression algorithm in accurately categorizing incident descriptions.

4.7 Implementation, Evaluation and Results of SVM

The TF-IDF vectorization technique was utilized in Support Vector Machine (SVM) model. SVM is a widely used machine learning algorithm for classification tasks, and TF-IDF allows us to convert text data into a numerical format suitable for modeling.

4.7.1 Implementation

Initially the dataset were splitted into training and testing purpose. The different crime categories were encoded into numerical format using the LabelEncoder. Next, the text data was vectorized using the TfidfVectorizer, which transformed the incident descriptions into numerical feature vectors while considering the importance of terms in the corpus. A linear SVM model was selected for its interpretability and effectiveness in handling high-dimensional data. The SVM model was trained using the vectorized training data.

4.7.2 Evaluation and Results

The trained SVM model was evaluated using the test dataset. The accuracy of the model was calculated, indicating the proportion of correct predictions. Additionally, a more detailed evaluation was performed using the classification report, which includes precision, recall, and F1-score for each category. These metrics provide insights into the model's performance across different categories, measuring its ability to correctly classify instances.

The SVM model achieved an impressive accuracy of approximately 99.95% on the test dataset, showcasing its effectiveness in categorizing incident descriptions accurately. The classification report further validates the model's strong performance, with high precision, recall, and F1-scores for most categories. Notably, the macro and weighted averages for precision, recall, and F1-score were close to 1.00, indicating a well-balanced performance across categories.

4.8 Implementation, Evaluation and Results of Random Forest

Random Forest classifier, known for its ensemble-based learning and robustness, emerges as an appealing choice. The reason behind its selection rely upon its ability to handle high-dimensional data, capture intricate relationships between features, and provide insights into the feature importance contributing to classification decisions..

[6] Calculates the Term Frequency-Inverse Document Frequency (TF-IDF) values of words, capturing their significance in text data. This helps in effective feature representation for machine learning models.

4.8.1 Implementation

The implementation involved several key steps. Firstly, the categorical labels were transformed into numerical format using a LabelEncoder. The text data was then preprocessed and vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer, which represented the descriptions as numerical feature vectors. A Random Forest classifier was chosen for its ensemble-based nature and the ability to handle complex relationships within the data. The classifier was trained using the training data's TF-IDF vectors and the encoded labels.

4.8.2 Evaluation and Results

Using the TF-IDF vectors of the test data, we generated predictions. These predictions were then transformed from numerical values back into their respective categorical labels. This transformation was achieved through the inverse transformation process of the LabelEncoder.The accuracy score was calculated by comparing the predicted labels to the actual labels in the test data. Additionally, a more detailed evaluation was conducted using the classification report, which provided precision, recall, and F1-score for each class.

The model achieved an accuracy of approximately 99.97%, indicating a high level of accuracy in its predictions. The classification report further validated the model's excellence, with high precision, recall, and F1-scores across all categories. The results signify that the combination of TF-IDF vectorization and the Random Forest algorithm effectively captured the intricate patterns present in the incident descriptions, enabling accurate and reliable classification. TF-IDF vectorization was used to convert text data into numerical vectors, which were then fed into a Random Forest classifier for training and prediction. The LabelEncoder was utilized to handle the conversion between numerical and categorical labels, ensuring that the final predictions were in the form of category labels. This combination allowed for the accurate classification of incidents based on their textual descriptions.

The objectives 3 and 4 outlined in research objectives section have been successfully accomplished.

4.9 Comparison of Developed Models

Fig. 8 depicts the comparison table of the developed models. Precision, Recall, and F1-Score values are described as "High" for models where the classification report indicated strong performance across most categories. LSTM and GRU models did not have detailed precision, recall, and F1-Score.

LSTM Model: The LSTM model exhibited promising outcomes in classifying crime categories based on textual descriptions. The evaluation process involved training the model on the training data and assessing its performance on the testing data using accuracy and loss criteria. The test accuracy was approximately 47.48%, accompanied by a test loss of around 1.98. The confusion matrix highlighted areas of accuracy and potential refinement, showcasing the model's performance for each crime category.

GRU Model: Similar to the LSTM model, the GRU-based model was evaluated after training on the training data for 5 epochs with a batch size of 64. The results were consistent, with a test accuracy of approximately 47.48% and a corresponding test loss of around 1.98. Despite utilizing the GRU architecture, the model struggled to

Model	Test Accuracy	Test Loss	Precision	Recall	F1-Score
LSTM	47.48%	1.98	N/A	N/A	N/A
GRU	47.48%	1.98	N/A	N/A	N/A
Logistic Regression	99.91%	N/A	High	High	High
SVM	99.95%	N/A	High	High	High
Random Forest	99.97%	N/A	High	High	High

Figure 8: Comparison of Developed Models

effectively distinguish between crime categories, as indicated by the confusion matrix. The complexity and diversity of the crime data posed challenges for accurate categorization.

Logistic Regression: In contrast to the LSTM and GRU models, the Logistic Regression model achieved remarkable accuracy, with an average accuracy of approximately 99.91% across all categories. The model's performance was confirmed by high precision and recall values, reflecting its ability to predict various unified categories accurately. The TfidfVectorizer and Logistic Regression algorithm's effectiveness played a pivotal role in the model's exceptional performance.

SVM Model: The SVM model demonstrated impressive accuracy, achieving approximately 99.95% on the test dataset. The detailed classification report reinforced the model's strong performance, with high precision, recall, and F1-scores for most categories. The macro and weighted averages also indicated a well-balanced performance across categories. The SVM model's accuracy and comprehensive evaluation metrics highlighted its efficacy in accurately categorizing incident descriptions.

Random Forest Model The Random Forest model excelled in its predictions, achieving an accuracy of approximately 99.97%. This high accuracy indicated the successful capture of intricate patterns present in incident descriptions. The combination of TF-IDF vectorization and the Random Forest algorithm played a pivotal role in the model's excellence. The classification report further validated the model's reliability, with high precision, recall, and F1-scores across all categories.

In summary, while the LSTM and GRU models struggled to distinguish between complex crime categories, the Logistic Regression, SVM, and Random Forest models showcased exceptional performance. The latter three models demonstrated high accuracy, precision, recall, and F1-scores, highlighting their effectiveness in accurately categorizing incident descriptions. The choice of algorithm and feature representation played a significant role in achieving reliable and insightful results. The research objective 5 set forth were met with success.

5 Discussion

The primary research question addressed in this project was: "How well can machine learning algorithms categorize crimes based on textual descriptions?" To answer this question, various machine learning models were implemented, trained, evaluated, and their results were analyzed. The research aimed to determine the effectiveness of different algorithms in accurately classifying crimes using textual descriptions.

The results of the research project provided valuable insights into the performance of the implemented models. The LSTM and GRU models, which were specialized recurrent neural networks, exhibited promising outcomes in classifying crime categories based on textual descriptions. However, their performance was limited, with both models achieving a test accuracy of approximately 47.48% and corresponding test losses of around 1.98. These results suggested that while the models captured certain patterns in the data, they struggled to effectively distinguish between crime categories, likely due to the complexity and diversity of crime-related language. In contrast, the Logistic Regression model demonstrated exceptional performance in accurately categorizing incident descriptions. Leveraging the TfidfVectorizer for text vectorization and the simplicity of Logistic Regression, this model achieved an average accuracy of approximately 99.91% across all categories. The detailed classification report further confirmed the model's high precision and recall values, underscoring its ability to accurately predict various unified categories. The SVM model also showcased strong capabilities, achieving an impressive accuracy of approximately 99.95%. The classification report highlighted high precision, recall, and F1-scores for most categories. These results indicated that the SVM model effectively categorized incident descriptions with remarkable accuracy and consistency. Furthermore, the Random Forest model, utilizing TF-IDF vectorization, excelled in accurately classifying crime categories. With an accuracy of approximately 99.97%, this model's performance was exceptional, as supported by high precision, recall, and F1-scores across all categories.

In conclusion, the research project's findings indicated that while specialized neural network models like LSTM and GRU demonstrated potential, simpler algorithms such as Logistic Regression, SVM, and Random Forest could achieve superior performance in categorizing crimes based on textual descriptions. The substantial differences in accuracy and precision among the models underscored the importance of selecting appropriate algorithms and techniques tailored to the nature of the data and the research objectives.

6 Conclusion and Future Work

In this research project, various machine learning models were implemented and evaluated for the task of categorizing crime incidents based on textual descriptions. The models included LSTM, GRU, Logistic Regression, SVM, and Random Forest. The objective was to enhance crime analysis, resource allocation, and policy-making through accurate crime classification. The models were trained, evaluated, and compared based on accuracy, precision, recall, and F1-scores.

The LSTM and GRU models, although showing promise, struggled to effectively distinguish between complex crime categories. They achieved modest test accuracy and encountered challenges in capturing diverse crime-related language. In contrast, the Logistic Regression, SVM, and Random Forest models demonstrated exceptional performance. Logistic Regression achieved high accuracy and precision, leveraging TfidfVectorizer

for feature extraction. SVM effectively utilized the TF-IDF vectorization technique to categorize incidents with impressive accuracy, precision, and recall. Random Forest's ensemble-based learning excelled in capturing intricate patterns within descriptions, resulting in high accuracy and comprehensive performance metrics.

Future Work: The following are some future research directions that may be explored. First, examining how ensemble strategies for model fusion might improve accuracy and adaptability. Second, exploring sophisticated textual elements like word embeddings or transformer-based models like BERT has the potential to help us comprehend crime descriptions better. Third, strategies for data augmentation can be considered to support models, particularly LSTM and GRU. Finally, crucial stages to transfer research results into useful applications include tuning hyperparameters, assuring the interpretability of complex models, and applying the best-performing model in real-time crime analysis scenarios.

7 Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr.Catherine Mulwa, for her invaluable guidance, unwavering support, and insightful feedback throughout this research endeavor. Her expertise has been instrumental in shaping this work. I am also deeply thankful to my parents for their constant encouragement and belief in my abilities. I extend my appreciation to my friends Karthik and Dona for their friendship and thoughtful discussions. Furthermore, I am immensely grateful to my girlfriend for her unwavering presence and encouragement, which provided me with the strength to persevere. Their combined support has been crucial in every step of this journey.

References

Abdulrahman, N. and Alkhader, W. (2017). Knn classifier and naive bayse classifier for crime prediction in san francisco context, *International Journal of Database Management Systems* **9**: 1–9.

Al-Ghamdi, S., Al-Khalifa, H. and Al-Salman, A. (2023). Fine-tuning bert-based pretrained models for arabic dependency parsing, *Applied Sciences* **13**: 4225.

Bai, X. (2018). Text classification based on lstm and attention, pp. 29–32.

Bradley, J. and Waller, I. (2017). The importance of data in crime prevention: Diagnosis and evaluation.

Brown, E. and Ballucci, D. (2007). Specialized knowledge: Understanding crime analyst's roles and responsibilities and the impact of their work, *Criminology & Criminal Justice* **0**(0): 17488958221095980.
URL: *https://doi.org/10.1177/17488958221095980*

Carter, E., Ward, T. and Strauss-Hughes, A. (2020). The classification of crime and its related problems: A pluralistic approach, *Aggression and Violent Behavior* **59**: 101440.

Gillioz, A., Casas, J., Mugellini, E. and Khaled, O. A. (2020). Overview of the transformer-based models for nlp tasks, *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 179–183.

Haider, M. N., Mumtaz, R. and Zaidi, S. M. H. (2022). Crime classification using machine learning and data analytic, *2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, pp. 093–098.

Imamguluyev, R. (2023). The rise of gpt-3: Implications for natural language processing and beyond, *International Journal of Research Publication and Reviews* **4**: 4893–4903.

Iqbal, R., Murad, M., Mustapha, A., Hassany Shariat Panahy, P. and Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction, *Indian Journal of Science and Technology* **6**: 4219–4225.

John, J., Varkey, M. S. and M, S. (2021). Multi-class text classification and publication of crime data from online news sources, *2021 8th International Conference on Smart Computing and Communications (ICSCC)*, pp. 64–63.

Lyons, A. (2016). *Multimodality*, pp. 268–280.

Mantoro, T., Permana, M. and Ayu, M. (2022). Crime index based on text mining on social media using multi classifier neural-net algorithm, *TELKOMNIKA (Telecommunication Computing Electronics and Control)* **20**: 570.

Nadkarni, P., Ohno-Machado, L. and Chapman, W. (2011). Natural language processing: An introduction, *Journal of the American Medical Informatics Association : JAMIA* **18**: 544–51.

Nasr, M., Karam, A., Atef, M., Boles, K., Samir, K. and Raouf, M. (2020). Natural language processing: Text categorization and classifications, **12**: 4542–4548.

Qi, Z. (2020). The text classification of theft crime based on tf-idf and xgboost model, pp. 1241–1246.

Rahmat, R., Faza, S., Adnan, S., Situmorang, D., Gunawan, D. and Zata Lini, T. (2021). News articles classification for electronic information and transaction law in indonesia using support vector machine, pp. 106–110.

Shabat, H., Omar, N. and Rahem, K. (2014). Named entity recognition in crime using machine learning approach, pp. 280–288.

Shojaee, S., Mustapha, A., Sidi, F. and A. Jabar, M. (2013). A study on classification learning algorithms to predict crime status, *International Journal of Digital Content Technology and its Applications* **7**: 361–369.

Stalidis, P., Semertzidis, T. and Daras, P. (2021). Examining deep learning architectures for crime classification and prediction, *Forecasting* **3**: 741–762.

Sundhara Kumar, K. B. and Bhalaji, N. (2016). A study on classification algorithms for crime records, *in* A. Unal, M. Nayak, D. K. Mishra, D. Singh and A. Joshi (eds), *Smart Trends in Information Technology and Computer Communications*, Springer Singapore, Singapore, pp. 873–880.

Yang, W., Zuo, W. and Cui, B. (2019). Detecting malicious urls via a keyword-based convolutional gated-recurrent-unit neural network, *IEEE Access* **7**: 29891–29900.