

STREAMLINING EMAIL CLASSIFICATION IN CRIME INVESTIGATION DEPARTMENT

**Capstone Project Report
MID SEMESTER EVALUATION**

Submitted by:

(102017070) Garima Chandna

(102017060) Simranjit Kaur

(102017065) Rashmeet Kaur

(102017059) Aakanksha Pandey

BE Third Year, CSE

CPG No: 162

Under the Mentorship of
Dr. Husanbir Singh Pannu
Assistant Professor



**Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala
July 2023**

ABSTRACT

The Crime Investigation Department receives a large number of complaints in the form of emails on a daily basis. The current process of classifying these emails is manual, that is CID Officials have to read every single complaint and then classify it to the correct category it belongs. This makes the whole process very tiring and time consuming. Moreover, the current process is error prone. Error may include misclassification of complaints, missing important information.

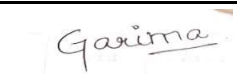
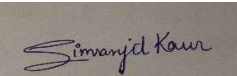
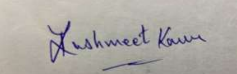
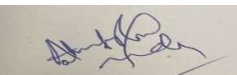
So, the objective of our project is to Streamline the Complaint Classification process in the Crime Investigation Department. The main goal of the project is to reduce the workload of investigators by automating the complaint classification process. Different categories such as accidents, missing person, molestation, murder, rape, and theft will be classified through the model, enabling investigators to focus on analyzing the information contained in the complaints, rather than spending time sorting and categorizing them manually.

To make this possible we will first collect the complaint data from CID and then we will move forward to the next step of building a ML model which will classify the complaints and the end product of our model is a website to be used by CID officials to classify the complaints and by complainants to file a complaint through our website.

DECLARATION

We hereby declare that the design principles and working prototype model of the project entitled **Streamlining Email Classification in CID** is an authentic record of our own work carried out in the Computer Science and Engineering Department, TIET, Patiala, under the guidance of Dr. Husanbir Singh Pannu during 6th semester (2023).

Date: 08-08-2023

Roll No.	Name	Signature
102017070	GARIMA CHANDNA	
102017060	SIMRANJIT KAUR	
102017065	RASHMEET KAUR	
102017059	AAKANKSHA PANDEY	

Counter Signed By:

Faculty Mentor:

Co-Mentor(if any):

Dr. Husanbir Singh Pannu

Dr. _____

Assistant Professor

Designation

CSED,

CSED,

TIET, Patiala

TIET, Patiala

ACKNOWLEDGEMENT

We would like to express our thanks to our mentor Dr. Husanbir Singh Pannu. He has been of great help in our venture, and an indispensable resource of technical knowledge. He is truly an amazing mentor to have.

We are also thankful to Dr. Shalini Batra, Head, Computer Science and Engineering Department, entire faculty and staff of Computer Science and Engineering Department, and also our friends who devoted their valuable time and helped us in all possible ways towards successful completion of this project. We thank all those who have contributed either directly or indirectly towards this project.

Lastly, we would also like to thank our families for their unyielding love and encouragement. They always wanted the best for us and we admire their determination and sacrifice.

Date: 04-08-2023

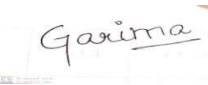
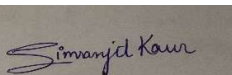

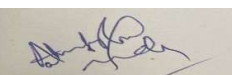
Roll No.	Name	Signature
102017070	GARIMA CHANDNA	
102017060	SIMRANJIT KAUR	
102017065	RASHMEET KAUR	
102017059	AAKANKSHA PANDEY	

TABLE OF CONTENTS

	Contents
ABSTRACT	2
DECLARATION	3
ACKNOWLEDGEMENT	4
TABLE OF CONTENTS	5
LIST OF TABLES.....	7
LIST OF FIGURES	8
LIST OF ABBREVIATIONS.....	9
1. INTRODUCTION	10
1.1 Project Overview	10
1.2 Need Analysis.....	12
1.3 Research Gaps	13
1.4 Problem Definition and Scope.....	14
1.5 Assumptions and constraints	15
1.6 Standards	16
1.7 Approved Objectives	16
1.8 Methodology.....	17
1.9 Project Outcomes and Deliverables.....	18
1.10 Novelty of Work.....	18
2. REQUIREMENT ANALYSIS.....	20
2.1 Literature Survey	20
2.1.1 Theory Associated with Problem Area.....	20
2.1.2 Existing System and Solutions	22
2.1.3 Research Findings for Existing Literature.....	23
2.1.4 Problem Identified	25
2.1.5 Survey of Tools and Technologies Used	25
2.2 Software Requirement Specification	25
2.2.1 Introduction.....	25
2.2.2 Overall Description.....	26
2.2.3 External Interface Requirements	28
2.2.4 Other Non-functional Requirements.....	28
2.3 Cost Analysis.....	30
2.4 Risk Analysis.....	30
3. Methodology Adopted.....	31
3.1 Investigative Techniques	31
3.2 Proposed Solution.....	33
3.3 Work Breakdown Structure.....	34

3.4	Tools and Technologies Used.....	35
4.	Design Specifications	36
4.1	System Architecture.....	36
4.2	Design Level Diagrams	37
4.2.1.	State Chart Diagram	37
4.2.2	Activity Diagram	39
4.2.3	Component Diagram.....	40
4.2.4	Entity Relationship Diagram	41
4.2.5	Class Diagram.....	42
4.2.6	Data Flow Diagram.....	43
4.3	User Interface Diagrams	46
4.3.1	Use-Case Diagram	46
4.3.2	Use-Case Template.....	47
4.4	Snapshots of Working Prototype	51
5.	Conclusions and Future Scope.....	57
5.1	Work Accomplished	57
5.2	Conclusions	57
5.3	Environmental (/ Economic/ Social) Benefits	58
5.4	Future Work Plan.....	58
	APPENDIX A: References.....	59

LIST OF TABLES

Table No.	Caption	Page No.
1.1	Assumptions of StreamlineCID	15
1.2	Constraints of StreamlineCID	15
1.3	Standards of StreamlineCID	16
2.1	Research Findings for existing literature of StreamlineCID	23
3.1	Investigative Techniques of StreamlineCID	31
4.1	Use Case Template of Login	47
4.2	Use Case Template of Filing a Complaint	47
4.3	Use Case Template of Viewing Segregated Complaint	48
4.4	Use Case Template of Downloading Segregated Complaint Data	49

LIST OF FIGURES

Fig. No.	Caption	Page No.
2.1	Text Classification Process	21
3.1	Work Breakdown Structure	34
4.1	System Architecture	36
4.2	State Chart Diagram for Complainant	37
4.3	State Chart Diagram for Admin	37
4.4	State Chart Diagram for CID Official	38
4.5	Activity Diagram for Complainant	39
4.6	Activity Diagram for CID Official	39
4.7	Component Diagram	40
4.8	Entity Relationship Diagram	41
4.9	Class Diagram	42
4.10	Data Flow Diagram – Level 0	43
4.11	Data Flow Diagram – Level 1	44
4.12	Data Flow Diagram – Level 2	45
4.13	Use Case Diagram	46
4.14	Home Page of StreamlineCID	51
4.15	CID Official Registration Page of StreamlineCID	52
4.16	Complainant Registration Page of StreamlineCID	52
4.17	Complaint Filing Page of StreamlineCID	53
4.18	Segregated Complaints Page of StreamlineCID	54
4.19	Category Page of StreamlineCID	55
4.20	About Page of StreamlineCID	55
4.21	Contact Page of StreamlineCID	56
4.22	Policy Page of StreamlineCID	56

LIST OF ABBREVIATIONS

S.NO.	Abbreviation	Full Form
1.	SVM	Support Vector Machine
2.	CID	Crime Investigation Department
3.	NLP	Natural Language Processing
4.	EDA	Exploratory data Analysis
5.	GPU	Graphical processing unit
6.	TPU	Tensor processing unit
7.	UI	User Interface
8.	ER	Entity Relationship
9.	UML	Unified modeling Language
10.	TF-IDF	Term Frequency-inverse document frequency
11.	IDE	Integrated Development Environment
12.	KNN	K nearest neighbor
13.	NLTK	Natural Language Toolkit
14.	CSS	Cascading Style sheets
15.	OS	Operating system
16.	QA	Quality Assurance

1. INTRODUCTION

1.1 Project Overview

Introduction:

Crime reporting and classification are critical aspects of law enforcement and public safety. Manual categorization of crime reports can be time-consuming and error-prone. This project aims to develop a text classification model to automatically categorize crime reports into five categories: "rape," "murder," "theft," "kidnap," and "accident." The model will streamline crime data analysis and enhance the efficiency of crime investigation processes.

Objectives:

The primary objective of this project is to build an accurate and robust text classification model capable of handling diverse and complex crime reports. The specific goals are to:

- Create a comprehensive labeled dataset of crime reports representing the five categories.
- Preprocess the text data to prepare it for model training.
- Explore and select an appropriate machine learning architecture for text classification.
- Train the model on the labeled dataset to achieve high accuracy in categorizing crime reports.
- Deploy the model in a user-friendly interface for real-time crime report classification.

Approach:

The project will follow these main steps:

- **Data Collection:** Gather a diverse and labeled dataset of crime reports from law enforcement agencies, news articles, and public databases.
- **Data Preprocessing:** Clean, tokenize, and normalize the text data to prepare it for model training.
- **Model Selection:** Explore and select an appropriate machine learning architecture, considering naïve bayes, SVM, logistic regression, for text classification.
- **Model Training:** Fine-tune the selected model on the preprocessed crime reports dataset.
- **Model Evaluation:** Assess the model's performance using classification metrics on a separate test dataset.

- **Model Deployment:** Implement the trained model in a user-friendly interface for real-time crime report classification.

Key Components:

The key components of the project include:

- A diverse and labeled dataset of crime reports.
- Text preprocessing techniques (tokenization, stop word removal, stemming, lemmatization).
- A selected machine learning model for text classification.
- Model training and evaluation processes.
- A user-friendly interface for real-time crime report classification.

Timeline:

The project is expected to be completed by the end of December, divided into the following phases:

- Data Collection and Preprocessing
- Feature Extraction
- Model Selection and Training
- Model Evaluation
- Model Deployment and Interface Development

Expected Impact/Benefits:

The successful implementation of the text classification model will provide law enforcement agencies with an efficient tool for crime report analysis and classification. Automating the classification process will save time and resources, enabling investigators to focus on more critical tasks. The model will enhance data-driven decision-making and contribute to faster crime solving and public safety.

Risks and Mitigation:

Potential risks include data quality issues, model overfitting, and interface usability challenges. We will mitigate these risks through thorough data cleaning, implementing regularization techniques, and conducting user testing during the interface development phase.

Budget:

There is no cost involved as our project does not have any hardware component.

Conclusion:

The text classification model for crime reporting aims to revolutionize crime investigation processes by automating and streamlining the categorization of crime reports. With the successful implementation of this project, law enforcement agencies will have a powerful

tool to enhance their crime analysis capabilities, ultimately leading to safer communities and improved public safety.

1.2 Need Analysis

Introduction:

A crucial role in maintaining law and order in society is played by the Crime Investigation Department (CID), which receives a large volume of complaint emails related to various crimes, such as theft, murder, accidents, and more. The complaint classification process is currently manual and time-consuming, leading to delays and errors. Therefore, the need arises to streamline the email classification process in the CID by implementing an automated system that can accurately categorize incoming complaint emails into various categories.

Current Situation:

Total of 60,96,310 cognizable crimes comprising 36,63,360 Indian Penal Code (IPC) crimes and 24,32,950 Special & Local Laws (SLL) crimes were registered in 2021. A total of 29,272 cases of murder were registered during 2021, showing a marginal increase of 0.3% over 2020 (29,193 cases) [1].

A total of 4.28 lakh cases of crime against women were registered in 2021, showing an increase of 15.3% over 2020. The rate of crimes against women has increased from 56.5 per lakh in 2020 to 64.5 per lakh in 2021. In 2021, there were 1.73 lakh deaths that were caused due to Traffic Accidents, accounting for around 44% of the total accidental deaths for the year [2]. These statistics depict the volumes of crime reports CID receives and has to manage.

Currently, the complaint email classification process in the CID is manual and dependent on the human eye's efficiency. Each complaint email is read, analyzed for its content, and manually classified into a specific category. This process is time-consuming, leading to delays in response time and action on the complaints received. Moreover, human error is a common occurrence, leading to misclassification of emails, which can result in serious consequences.

Proposed Solution:

The proposed solution is to implement an automated email classification system that uses natural language processing to analyze and classify incoming complaint emails into specific categories accurately. The emails will be classified into appropriate categories such as accidents, kidnaps, murder, rape/molestation, and theft using natural language processing

(NLP) techniques. The system will be capable of handling the high volume of emails received daily, reducing response time and errors.

Benefits:

Several benefits will be brought to the CID through the implementation of an automated complaint email classification system. Response time will improve, errors will be reduced, and efficiency will be increased. The system will enable the department to take prompt action on complaints received, leading to better law and order maintenance. Furthermore, the system will reduce the workload of the human workforce, allowing them to focus on other essential tasks.

Conclusion:

The current email classification process in the Crime Investigation Department is manual and time-consuming, leading to delays and errors. The proposed solution to implement an automated email classification system that uses machine learning algorithms and NLP techniques will streamline the email classification process, improve efficiency, and reduce errors. The implementation of such a system will bring significant benefits to the department and enable them to provide better services to society.

1.3 Research Gaps

Existing text classification techniques suffer from the following research gaps:

- A majority of the recent research has been studied on the English language, and only a few studies have been conducted on the Hindi language.
- There isn't a single system that both enables complainant grievance submission and automatic categorization of cases for the Crime Investigation Department.
- The current manual process can be influenced by the officer's mood, potentially resulting in biased outcomes.
- The current techniques focus on data preprocessing and modeling tasks but lacks adequate emphasis on data visualization during the exploratory analysis phase.
- Semi-supervised learning is limited due to lack of rule-based structural grouping [3].
- Research gaps exist in investigating ensemble techniques that combine predictions from multiple text classification models trained with different

algorithms. Ensemble methods can enhance prediction accuracy and provide robustness against individual model biases [4].

- No transparency and openness in model development, including sharing datasets, code, and model architectures, to facilitate reproducibility and encourage collaboration among researchers and practitioners.

1.4 Problem Definition and Scope

Problem Definition:

The problem addressed in this project is "Streamlining Email Classification in the Crime Investigation Department." The goal is to develop an automated system that can accurately and efficiently classify emails received by the Crime Investigation Department (CID) into categories: "murder", "rape", "kidnap", "theft", and "accident". The CID receives a large volume of emails daily, reporting various incidents, and manually processing and categorizing them is time-consuming and error-prone. The project aims to leverage natural language processing (NLP) techniques and machine learning to build a robust and scalable email classification system that enhances the department's investigative workflow and response time.

Scope:

The project's scope encompasses the development of a text classification system focused on email categorization for the specific categories: "murder", "rape", "kidnap", "theft", and "accident". It aims to streamline the crime investigation department's email handling process, augmenting the investigative efforts and providing timely responses to reported incidents. While the project focuses on these specific categories, the methodologies employed can be extended to address other crime-related classifications in the future. Additionally, the project emphasizes the ethical use of AI, ensuring privacy, transparency, and fairness in email classification to maintain public trust and adherence to legal and regulatory frameworks.

1.5 Assumptions and constraints

Table 1.1 Assumptions of StreamlineCID

S.NO.	Assumptions
1.	Sufficient and Labeled Data: The project assumes the availability of a substantial amount of labeled data containing textual complaints along with their corresponding categories. This data will be used for model training, validation, and evaluation.
2.	Homogeneous Data Distribution: The project assumes that the data used for training and testing the model is representative of the real-world distribution of complaints. It is assumed that the distribution of complaint categories in the training data is similar to that of unseen, real-world complaints.
3.	Predefined Categories: The project assumes that the complaint categories (e.g., murder, theft, accident, kidnap) are predefined and well-defined. Each complaint falls into one and only one category.
4.	Sufficient Computational Resources: It is assumed that the project will have access to sufficient computational resources to train and evaluate deep learning models. Training deep learning models can be computationally intensive, so access to GPUs or TPUs is beneficial.
5.	Language: The project assumes that all textual complaints are in Hindi Language. Handling multilingual complaints is beyond the scope of this specific project.

Table 2.2 Constraints of StreamlineCID

S.NO.	Constraints
1.	Data Quality: The quality of the complaint data can significantly impact the system's performance. Noise, missing labels, or inaccurately labeled complaints may affect the model's ability to generalize well.
2.	Scalability: While the project aims to create an efficient classification system, scalability may be a concern when dealing with a large volume of real-time complaints. Ensuring real-time responsiveness could pose challenges.
3.	Domain-specific Jargon: Complaints may contain domain-specific jargon or language, which could lead to difficulties in understanding or categorizing them accurately.

4.	Bias in Data: The complaint data might be biased towards certain categories or specific demographics, leading to potential bias in the model's predictions. Care should be taken to minimize and address any bias in the dataset and model.
5.	Ethical and Legal Considerations: The project must adhere to ethical guidelines and legal regulations concerning handling sensitive data and ensuring privacy and security.

1.6 Standards

Table 3.3 Standards of StreamlineCID

S.No	Standards	Details
1	Web 2.0	Web 2.0 refers to websites that emphasize user-generated content, ease of use, participatory culture, and interoperability for end users.
2	ISO/IEC 90003	Software engineering -- Guidelines for the application of ISO 9001:2008 to computer software is a guideline developed for organizations in the application of ISO 9001 to the acquisition, supply, development, operation, and the maintenance of computer software and related support services.
3	ISO/IEC/IEEE 29148	This standard provides details for the processes and products related to the engineering of requirements for software products (including services) and systems throughout their life cycle. It defines the construct of a good requirement, provides attributes and characteristics of requirements, and discusses the iterative and recursive application of requirements processes throughout the life cycle. It also provides guidance in the 7 application of requirements engineering and management processes for requirements-related activities in ISO/IEC/IEEE 12207 and ISO/IEC/IEEE 15288.

1.7 Approved Objectives

The objectives of the Capstone Project are as follows:

- To acquire the complaint dataset from the Crime Investigation Department.
- To create a user-friendly interface where users can easily file their complaints which are received by the CID officials.
- To develop an efficient complaint email classification system that can classify incoming emails into different categories like accidents, kidnaps, murder, rape, and theft.

- To create a user-friendly interface where the CID officials can easily access the complaint email classification results and understand the categories to which each email has been classified.
- To allow the CID officials to download the segregated complaints as a csv file.

1.8 Methodology

1.8.1 Communication and Planning

We will communicate with a CID official to understand the requirements of the system to be developed.

1.8.2 Modeling

- Based on the understanding of the project, E-R diagrams, Data flow diagrams, UML diagrams will be made.
- Wireframing of UI for the project will be done and user surveys will be conducted to review the Human-Computer Interactions. The final UI of the system will be developed on the outcomes of such multiple surveys.

1.8.3 Construction

Construction of the project is divided into two major categories:

- **Data Collection and Machine Learning:**

Collection of the email dataset from the Crime Investigation Department, ensuring compliance with data privacy and security standards. Traditional, and Machine Learning-based text processing and prediction approaches will be implemented to extract information and insights. Python-based NumPy Stack, SpaCy to be used for operations.

- **Software Product Development:**

It will include the development of a modular web application for delivering the product to various stakeholders. A user-friendly interface for the stakeholders to interact with the email classification system will be developed. The frontend will be powered by React, JavaScript, and Material UI. Implementation of backend using Node.js, and MongoDB for flexibility and modularity.

1.8.4 Testing

Unit testing, Integration testing, and System testing like white and black box testing on the proposed system for obtaining results.

1.9 Project Outcomes and Deliverables

Text Classification Model: The trained text classification model capable of accurately categorizing crime-related emails into “murder”, “rape”, “kidnap”, "theft" and "accident" classes.

User Interface (UI) for General Public: A user-friendly interface for the masses where they can file their complaints.

User Interface (UI) for CID Personnel: A user-friendly interface for CID investigators and personnel to interact with the email classification system, review predictions.

Project Report and Presentation: A detailed project report summarizing the methodology, outcomes, and key findings of the project. A presentation for stakeholders highlights the project's achievements and potential impacts.

1.10 Novelty of Work

Data obtained from CID: We are working on real time data obtained directly from the CID officials.

Application of NLP Techniques in Crime Investigation: While email classification and natural language processing (NLP) are well-established areas of research, their application specifically to crime investigation and law enforcement is relatively novel. The project explores the use of NLP techniques to process and analyze crime-related emails, enabling the automation of email categorization in a law enforcement context.

Multiclass Classification for Crime Types: Instead of binary classification, which is common in some email filtering systems, this project delves into multiclass classification for crime types, encompassing “murder”, “rape”, “kidnap”, "loot" and "accident". Multiclass classification presents unique challenges in handling multiple crime categories effectively and is particularly relevant in the context of crime investigation.

Scalability for Real-time Crime Incident Identification: The project emphasizes real-time processing of incoming emails, making it a novel approach in the context of crime reporting systems. Ensuring the scalability of the system to handle the volume of incoming emails efficiently is a unique aspect of this work.

The novelty of this project lies in the combination of NLP techniques, machine learning models, ethical considerations, and real-world application in the domain of crime investigation. By addressing the challenges specific to crime reporting data and integrating cutting-edge

methodologies, the project aims to deliver a text classification system that significantly enhances the Crime Investigation Department's investigative capabilities and contributes to more effective crime management and public safety.

2. REQUIREMENT ANALYSIS

2.1 Literature Survey

2.1.1 Theory Associated with Problem Area

2.1.1.1 Introduction

The categorization of news articles which consists of understanding the topic of the articles and associating each of them to a category is talked about in [5]. It discusses the use of word embeddings for the crime categorization on an Italian dataset of 15,000 news articles. Both supervised and unsupervised categorization algorithms were explored. In the case of news articles related to crimes, the scope is to identify the type of crime (crime categorization). This task is important for many reasons. Categorization enables for further processing that are in the scope of crime analysis. From each news article, it is possible to retrieve detailed information about the event it reports: the place, the thief, the victim. If we know the type of crime, we can also retrieve information specific to that crime type, e.g., the stolen items in a theft. Moreover, Machine Learning approaches can help crime analysts to identify the connected events and to generate alerts and predictions that lead to better decision-making and optimized actions.

Detailed information can be extracted through the application of Natural Language Processing (NLP) techniques.

According to the use case, the scope of assigning a news article to a crime category can be addressed following several approaches, such as text classification, community or topic detection.

2.1.1.2 Text Classification

Text classification is the process of classifying text documents into a predefined set of classes. It is a supervised learning approach in which a training set of documents $\{D_1, D_2, \dots, D_n\}$ labelled with classes from $\{1 \dots m\}$ are used to build a classification model and predicts the class label of a new incoming document based on the training model. Text classification types include single label and multi-label classification. When a document is assigned with only one class it is called single labelled and when more than one class is assigned for a document it becomes multi-label classification. Binary classification which predicts if a document belongs to a particular class or not is the best example of a single label

classification. Text categorization has various stages such as pre-processing, indexing and dimensionality reduction, classification and performance evaluation [6].

[7] describes and focuses on the following five elements of the text classification process: (1) document pre-processing, i.e. tokenization, stop-word removal, and stemming or lemmatization, (2) document modeling, i.e. representing a document in an appropriate form, to be processed by a machine learning algorithm, (3) feature selection and projection, (4) machine learning algorithm utilization to construct a classification model or function, and (5) quality indicators and evaluation methods. Figure 1 shows a basic outline of the text classification process.

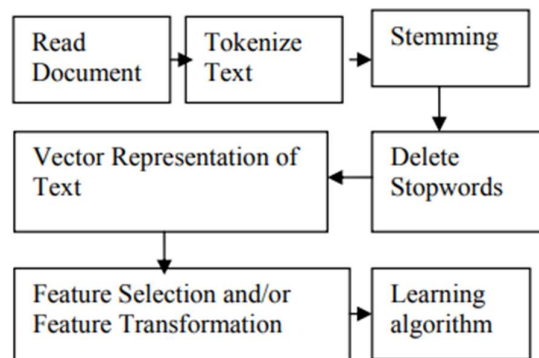


Fig. 2.1: Text Classification Process

TF-IDF algorithm was used to classify news articles in Bahasa Indonesia. This algorithm counts the weight of each word with respect to its repetition in the text and the number of files in which it exists. When a word is repeated too many times in all the texts, it means that that word is not important, and that a high precision has been achieved in classification [8].

The text classification of theft crime data based on the combination of TF-IDF and XGBoost algorithm was carried out in [9]. It achieved accurate and efficient classification of data. This was an effective attempt at a machine learning algorithm for police data mining, and a basic work for police data governance and crime prediction.

Given the high dimensions of the data involved, text classification is a challenging task. In [10], an approach was proposed to classify news texts. This approach consisted of three different steps: 1) text preprocessing, 2) feature extraction based on TF-IDF, and 3)

classification based on SVM. This approach was trained through the SVM classifier which was selected because it could support data with high dimensions.

In [11], an experiment was conducted for Sri Lankan news. Several active Twitter news groups such as ‘Ada Derana’, ‘Ceylon Today’, ‘ITN Sri Lanka’, ‘Lanka Breaking News’, ‘Lanka E News’ and ‘Sri Lanka News Now’ were chosen to extract the data. The short messages were classified by the system into 12 groups: war-terrorist-crime, economy-business, health, sports, development-government, politics, accident, entertainment, disaster-climate, education, society and international. These groups were chosen in order to cover the main areas of a general news provider.

In [12] character level and word level n-gram models were used to extract features of Turkish newspaper articles, and then three machine learning techniques namely Naive Bayes, SVM and Random Forest were applied to classify articles. When extracting features, different preprocessing techniques, namely, n-gram choice, stemming, and punctuation removal were used.

2.1.2 Existing System and Solutions

The Regional Crime Analysis Program (ReCAP) system is a computer application designed to aid local police forces (e.g., University of Virginia (UVA), City of Charlottesville, and Albemarle County) in the analysis and prevention of crime. ReCAP works in cooperation with the Pistol 2000 records management system, which aggregates and houses all of the crime information from a region [13].

Wikicrimes (<http://www.wikicrimes.org/>) is a Brazilian site that allows report crimes of different types predefined (burglary and theft) and create new ones. The user must log in to report a crime and indicate various data, including their location and type of crime. This site is a clear example of VGI platform that is not provided by an official organization. While information from sites like this is useful for querying and display statistics, it would be interesting to use it in real time to help official emergency systems. This would be possible if these capabilities were added to emergency management platforms or even a more ambitious bet would be adding these platforms to national IDEs. [14].

Spot Crime is a site that allows people to report and query data from crimes. Is an Open Data platform that uses Google Maps. The main source of the information is from police departments. The user can receive free crime alerts via email and SMS if he wants. In 2012, launched its own crime tip service, a VGI simple application, Crime Tip.us, allows users to

anonymously report crimes in their area. In 2014 it was one of the most visited crime mapping sites in the US, with over 8.5 million email alerts sent out on a monthly basis (according to data provided by the enterprise) [14].

2.1.3 Research Findings for Existing Literature

Table 2.1 Research Findings for existing literature of StreamlineCID

S. No.	Roll Number	Name	Paper Title	Tools/ Technology	Findings	Citation
1	102017070	Garima Chandna	“Using Word Embeddings for Italian Crime News Categorization”	Word2Vec, LinearSVC	The study reveals that the use of embeddings improves outcomes in many Natural Language Processing (NLP) activities, including text categorization.	[5]
2	102017070	Garima Chandna	“A comprehensive study of text classification algorithms”	KNN, Naïve Bayes, SVM, Neural Network based, Rule based classification	Proposes machine learning classification techniques along with the need of an appropriate dimensionality reduction technique to improve the expected outcome of classification.	[6]
3	102017060	Simranjit Kaur	“A recent overview of the state-of-the-art elements of text classification.”	Bag of Words, Naïve Bayes	Explores that both vector and graph representation of textual data have been presented in the literature, but among the two the vector representation seems to be more commonly used and in the literature.	[7]
4	102017060	Simranjit Kaur	“Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach”	TF-IDF	The study shows that TF-IDF algorithm could categorize online news articles in a high accuracy.	[8]

5	102017065	Rashmeet Kaur	“The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model”	TF-IDF, XGBoost Classifier	Studies the importance of the quality of sample data in affecting the accuracy of data classification and crime prediction. The improvement of data quality improves the accuracy of data classification and crime prediction more than that of the optimization algorithm.	[9]
6	102017065	Rashmeet Kaur	“A novel text mining approach based on TF-IDF and Support Vector Machine for news classification”	TF-IDF, Support Vector Machine	Highlights the basic workflow of text classification-text preprocessing, feature extraction and classification.	[10]
7	102017059	Aakanksha Pandey	“Twitter news classification using SVM”	Bag of Words, Support Vector Machine	The paper explains the benefits of using SVM, that it supports high dimensional data. SVM does not over fit to the data, as it minimizes both error and complexity. SVM also has the ability of finding the global minimum.	[11]
8	102017059	Aakanksha Pandey	“Effects of various preprocessing techniques to Turkish text categorization using n-gram features”	N-gram, Naïve Bayes, Support Vector Machines, Random Forest	The study reveals that Character level n-grams perform better than word level n-grams, however, using a bigram or a trigram model generally, has no significant effect on the results. And, SVM performs better than Naive Bayes and Random Forest in this domain.	[12]

2.1.4 Problem Identified

Major works have been done in classification of crimes but no significant work has been done in developing a web-based system for the use of crime reports in real time crime categorization and management.

2.1.5 Survey of Tools and Technologies Used

Following technologies have been surveyed for both machine learning and software development component of our product -

2.1.5.1 Machine Learning

- Languages - Python
- NumPy Stack - NumPy, Pandas, SciPy
- Machine Learning Libraries - Sci-Kit Learn, Matplotlib, Seaborn, Re, Googletrans
- Data Collection – Real data collected from CID official
- NLP Libraries - NLTK, SpaCy

2.1.5.2 Software Development

- Web Frameworks (Backend)- NodeJS(JavaScript)
- Web Frameworks(Frontend) - CSS (Bootstrap, Material UI), JavaScript (React.js)
- Database - NOSQL(MongoDB)

2.2 Software Requirement Specification

2.2.1 Introduction

A crucial role in maintaining law and order in society is played by the Crime Investigation Department (CID), which receives a large volume of complaint emails related to various crimes, such as theft, murder, accidents, and more. The complaint classification process is currently manual and time-consuming, leading to delays and errors. Therefore, the need arises to streamline the email classification process in the CID by implementing an automated system that can accurately categorize incoming complaint emails into various categories.

2.2.1.1 Purpose

The purpose of Streamlining Complaint Classification System in CID is to

- Develop an Automated Classification System: Design and implement a cutting-edge system capable of automatically segregating various categories of crimes in real-time. The system uses advanced machine

learning algorithms and natural language processing (NLP) techniques to efficiently categorize incoming complaints.

- **Enable Efficient Complaint Handling:** Provide a user-friendly platform where individuals can file their complaints. The system automatically classifies and organizes the complaints, streamlining the process for CID officials to view and address each case promptly.

2.2.1.2 Intended Audience and Reading Suggestions

The target user group for this system comprises two main categories: Crime Investigation Department officials (CID) and the general public seeking to file complaints. Additionally, the research conducted for this project holds valuable potential for researchers specializing in Text Classification tasks, particularly those involving short documents.

2.2.1.3 Project Scope

The "Streamlining Complaint Classification System in CID" project has a well-defined scope with specific objectives and limitations. The project aims to develop an automated email classification platform for the Crime Investigation Department (CID) to efficiently segregate various categories of crime-related complaints, like “theft”, “accident”, “kidnap”, “murder” and “rape”, in real-time based on the content of the complaints. Additionally, the project includes creating a user-friendly platform for the general public to easily file complaints, promoting public engagement in crime reporting. CID officials will benefit from an efficient complaint handling process, as the system streamlines the viewing and addressing of categorized complaints, optimizing investigative workflow. The project will leverage natural language processing (NLP) techniques for data preprocessing and analysis, transforming the complaint text into numerical representations suitable for machine learning models. Various machine learning algorithms, such as Naive Bayes, SVM, etc., will be explored and employed to achieve accurate email classification. However, support for multiple languages is beyond the scope of this project. Any crime related category apart from those mentioned above would not be classified by the system.

2.2.2 Overall Description

The project is a cutting-edge initiative aimed at revolutionizing crime investigation and public engagement within the Crime Investigation Department (CID). The project's primary objective is to develop an automated email classification platform that efficiently segregates

various categories of crime-related complaints, such as “theft”, “accident”, “kidnap”, “murder” and “rape” in real-time based on the content of the complaints.

The system serves as a user-friendly platform for the general public to easily file complaints, fostering increased public participation in crime reporting and aiding the CID in promptly addressing critical incidents. CID officials will experience enhanced efficiency in complaint handling, as the system streamlines the viewing and addressing of categorized complaints, enabling investigators to focus on specific crime types more effectively.

The project incorporates state-of-the-art natural language processing (NLP) techniques to preprocess and analyze complaint text data, transforming it into suitable numerical representations for machine learning models. Various machine learning algorithms, including Naive Bayes, SVM, etc., are explored and deployed to achieve accurate email classification. The system also allows the CID officials to download the segregated complaints as a csv file. By empowering CID officials with advanced technology and methodologies, the project paves the way for a more effective and responsive crime investigation process within the department.

2.2.2.1 Product Perspective

Our goal is to develop a software solution that streamlines the process of categorizing complaints based on their associated crimes. The software will be a user-friendly web application that utilizes an expert system consisting of multiple machine learning models. The primary objective of the project is to provide a decisive outcome that shows complaints categorized into one of five categories, including kidnaps, murder, theft, rape, and accidents. By employing this software solution, we aim to improve the efficiency and accuracy of complaint categorization, allowing officials to prioritize and address complaints more effectively.

2.2.2.2 Product Features

- This system will provide an interface for collaboration between CID officials and the general public.
- A user-friendly platform allows the general public to easily file complaints, encouraging increased public participation in crime reporting.
- Offers real-time complaint classification, swiftly categorizing incoming emails into relevant crime categories, such as kidnaps, murder, theft, rape, and accidents enabling timely responses to critical incidents.

- Allows the CID officials to download the segregated complaints as a csv file.

2.2.3 External Interface Requirements

The primary input to the system is crime reports data in Hindi Language. User just need to file the complaint about the crime with proper information for better results. The Controller will then run the model to segregate it according to ‘the type of crime’ and store in the database. For the CID officials, they just need to click the ‘view complaints’ button, and the complaints segregated in their categories will be presented to them.

2.2.3.1 User Interfaces

A graphical UI is created which provides a user-friendly environment for good visualizations. The user files complaint about the crime and then the interface links to the modules required for segregating it as ‘theft’, ‘murder’, ‘rape’, ‘accident’ or ‘kidnap’ depending upon the type of crime committed. The results of the segregation and matching can be seen on the web browser. The interface also has the functionality of viewing elements for data analysis for the CID officials.

2.2.3.2 Hardware Interfaces

Not Applicable

2.2.3.3 Software Interfaces

In this system one of the major software interfaces is acting as a messenger by running models to segregate complaints into different categories, storing it in the database and viewing the results on the web browser. This acts as a software intermediary that allows users to interact with the model.

2.2.4 Other Non-functional Requirements

The system is expected to meet the following Non-Functional requirements for a smooth and seamless experience with the system.

- Simple Interface: Requirements for a UI ask that it be modern, easy to use and distinctive.
- Scalability: The system must be scalable to accommodate a large volume of complaint data and users over time.
- Reliability: Overall reliability of the system shall be achieved through the process of complaint segregation.
- Availability: The system shall be available to all the CID officials and the public.
- Maintainability: The system should be updated from time to time.

- Error Handling: The system must have proper error handling mechanisms to ensure that stakeholders are notified in case of any errors or issues during the upload, processing, or segregation of the complaint data.

2.2.4.1 Performance Requirements

The performance of our system is measured through how accurately the complaints have been classified into the various categories. It can also be justified by looking at how well the complaints have been matched. If the complaint is published with important keywords in it then only the performance requirements will be met.

- Modified data in the database should be updated for all officials accessing it within two Seconds.
- The software should be portable, moving from one OS to another OS does not create any problem.
- The website should be able to run across various platforms and screen sizes like mobile phones, tablets and full-size computers.
- The program must be able to be run concurrently by multiple users 24/7
- Evaluated complaint should be classified within 30 seconds
- Data should be updated in the database within 2 Seconds.

2.2.4.2 Safety Requirements

- Administrator should conduct a maintenance survey of the system after every six months.
- CID officials should not depend fully on this software.

2.2.4.3 Security Requirements

- The system must ensure the security and confidentiality of the uploaded data and any other sensitive information throughout the data processing and segregation process.
- The interface must be safe in terms of user data and confidentiality.
- Website should be secure according to the industry best practices.
- The admin should keep the passwords secret and not share them with anyone.

2.3 Cost Analysis

As of now there is no cost requirement as we are not using any hardware component. We only require an online platform to deploy our web-application.

2.4 Risk Analysis

One of the major risks involving our project is the misclassification of complaints as some other complaint that will eventually be discarded. This will lead to loss of important information. This can improve over time as we feed more useful data concerning crimes which would increase the accuracy of our model.

3. Methodology Adopted

3.1 Investigative Techniques

Table 3.1 Investigative Techniques of StreamlineCID

S. No.	Investigative Projects Techniques	Investigative Techniques Description	Investigative Projects Examples
1	Descriptive	<p>In a crime investigation department, email classification involves systematically analyzing and categorizing emails based on their content, metadata, and other relevant characteristics. This technique aims to gain insights into the nature of the emails, identify potential leads, and prioritize the investigation process.</p> <p>The Regional Crime Analysis Program (ReCAP) system is a computer application designed to aid local police forces (e.g., University of Virginia (UVA)) in the analysis and prevention of crime. ReCAP works in cooperation with the Pistol 2000 records management system, which aggregates and houses all of the crime information from a region.</p> <p>Spot Crime is a site that allows people to report and query data from crimes. The main source of the information is from police departments. The user can receive free crime alerts via email and SMS if he wants.</p> <p>Wikicrimes is a Brazilian site that allows reporting crimes of different types predefined (burglary and theft) and creating new ones. The user must log in to report a crime and indicate various data, including their location and type of crime.</p>	Spotcrimes, Wikicrimes
2	Comparative	<p>Any service of this kind in the past results in complaints being directed solely to the inbox of CID officials. Our project aids in obtaining a clearer understanding of the email content by categorizing them into 5 distinct categories, thereby reducing the time required to route the mails to the appropriate departments.</p> <p>Machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, are trained on a labeled</p>	Support Vector Machines usually performs better in text classification as compared to other models such as naive bayes and logistic regression etc.

		<p>dataset of emails. These models learn to classify emails based on their content and patterns. Machine learning techniques can be more accurate and adaptable as they can generalize to emails with various wordings and structures.</p> <p>Projects comparing various Supervised, and Unsupervised methods. Regular Expression based techniques by analyzing important words are repeatedly used. SVM, CNN, NLP are common text classification Machine Learning techniques. In character level and word level n-gram models were used to extract features of Turkish newspaper articles, and then three machine learning techniques namely Naive Bayes, SVM and Random Forest were applied to classify articles. When extracting features, different preprocessing techniques, namely, n-gram choice, stemming, and punctuation removal were used. Our focus is on providing the streamlining service with high accuracy and in the most secure way possible as the data is highly confidential.</p>	
3	Experimental	<p>The goal is to identify the most effective and efficient approach for accurately categorizing emails relevant to the investigation. Implementing Machine Learning Concepts for Email Classification using Python based libraries like Sci-kit Learn. Implementing Machine Learning Concepts for Email Classification using react framework with data collection, classification, presentation and collaboration modules. Steps include gathering a representative dataset of emails related to the investigation. Extracting relevant features from the emails that will be used for classification. Implementing simple baseline methods for email classification, using different classification algorithms. Evaluating the performance of each method on the testing set using appropriate evaluation metrics. Based on the experimental results, choosing the most effective email classification method and implementing it in the crime investigation department's email management system for practical use.</p>	Phishing Email Classification, Incident Response Email Classification

3.2 Proposed Solution

The proposed solution is to implement an automated email classification system that uses natural language processing to analyze and classify incoming complaint emails into specific categories accurately. The emails will be classified into appropriate categories such as accidents, kidnaps, murder, rape, and theft using natural language processing (NLP) techniques. The system will be capable of handling the high volume of emails received daily, reducing response time and errors.

The proposed work is divided into two components:

- Machine Learning
- Software Development

3.2.1 Machine Learning:

The Machine learning solution involves 4 major steps to build a model: Data Collection, Data Cleaning, and Pre-processing, Feature Extraction and Classification. This model will then be used to classify complaints as rape, murder, theft, kidnap, and accident.

- Data Collection: Using an official from CID for confidential data
- Data Cleaning and Preprocessing: Applying Data cleaning and preprocessing techniques to clean and structure the data. The techniques used are converting the data from Hindi to English, Removal of special symbols, Removal of stop words and Lemmatization.
- Feature Extraction: Features can be extracted from the corpus by using TF-IDF or word embeddings. Our system used TF-IDF and represented it for further processing.
- Classification Algorithms: The plan is to select a Classification Algorithm by comparative analysis of state of art algorithms for example Logistic Regression, SVM , Naive Bayes, Random Forest.
- Prediction: Using the Trained model to predict the class of emails which the stakeholder uploads.

3.2.2 Software Development

Software development includes the building of a modular web application for delivering the product to various stakeholders i.e., CID officials and the general public.

This website allows any individual to submit a complaint after creating an account and logging in to the system. The received complaints are then processed by the system, which categorizes them based on their nature. To prioritize data security, the option to create accounts on the site for the CID Officials is intentionally excluded. Instead, only an administrator is granted the authority to create or delete accounts, specifically for accessing

the lodged complaints.

All authorized officials possess the capability to access and examine both the crime trends evident in the complaints and the content that has been categorized accordingly. This ensures that they can efficiently analyze the data and gain insights into prevailing issues. Additionally, these officials are empowered to securely download the complaints in a read-only format, guaranteeing the confidentiality and integrity of the mail's content.

By offering a complaint submission option, the website promotes an inclusive approach, enabling everyone to report complaints conveniently. Meanwhile, entrusting account management solely to administrators enhances data security by minimizing the risk of unauthorized access. This approach maintains a balance between accessibility and safeguarding sensitive information.

In conclusion, the website's design ensures ease of complaint submission for all users while upholding robust security measures to protect the integrity and confidentiality of the information. Authorized officials can effectively utilize the categorized data to gain valuable insights and take appropriate actions in response to the lodged complaints.

3.3 Work Breakdown Structure

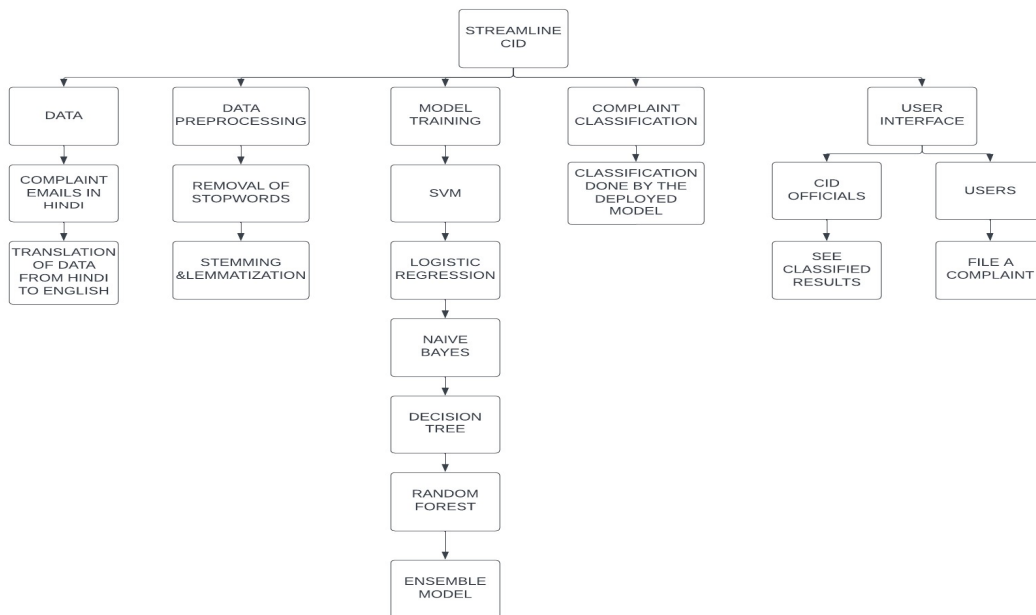


Fig. 3.1: Work Breakdown Structure

3.4 Tools and Technologies Used

3.4.1 Machine Learning

- NumPy Stack including NumPy, Pandas, Matplotlib
- Sci-Kit Learn
- Natural Language Toolkit (NLTK)
- SpaCy
- Google Translate

3.4.2 Web Development

- React for frontend
- NodeJS for Backend
- Material UI and Bootstrap CSS Frameworks for frontend
- JavaScript
- MongoDB Database.
- Chrome Web Tools for website optimization

3.4.3 UX design

- Draw.io
- Google Fonts

4. Design Specifications

4.1 System Architecture

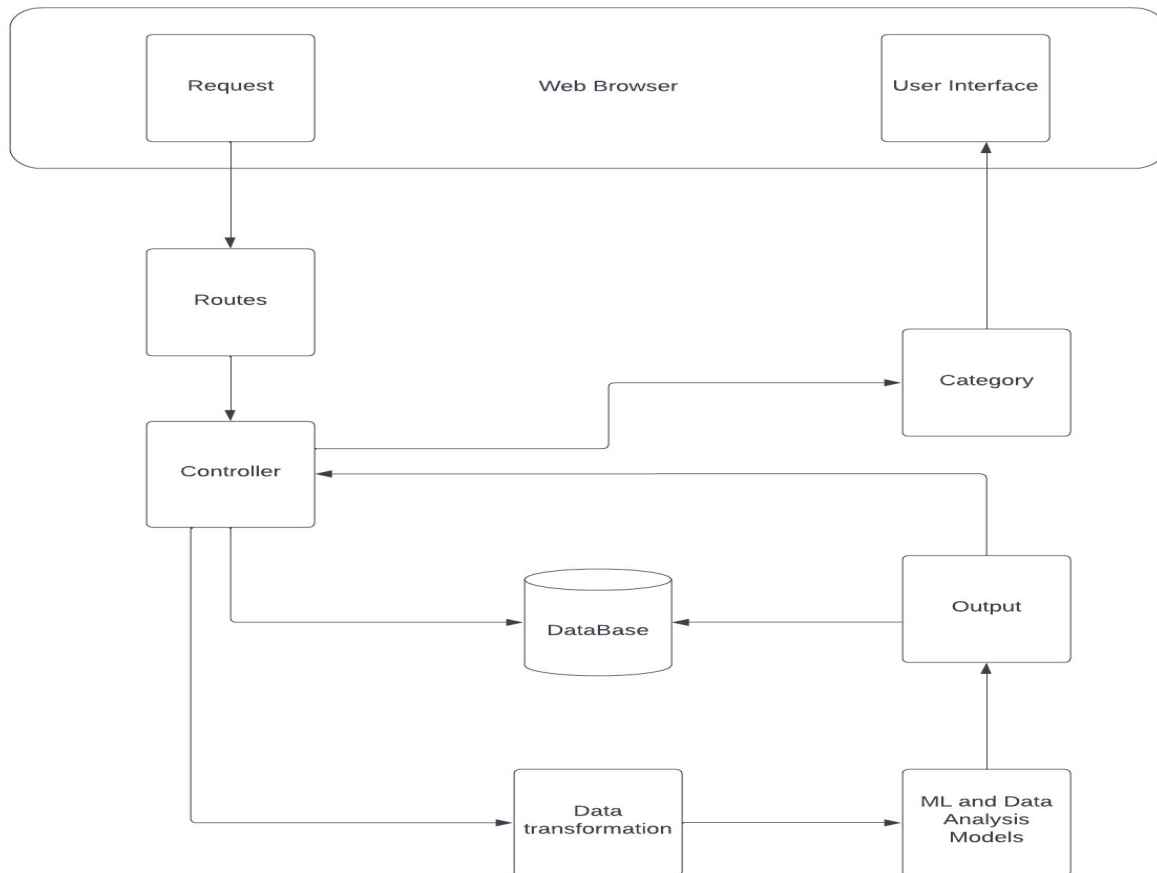


Fig. 4.1: System Architecture

The above diagram describes the overall architecture of the system. The requests will be sent to routes and corresponding action will be performed according to the requests. Models have direct communication with the database. Data is pre-processed and updated with useful information using machine learning models. Controller is given the power to do any of the three things: view information on a web browser, run any machine learning model, and extract data from the user.

4.2 Design Level Diagrams

4.2.1. State Chart Diagram

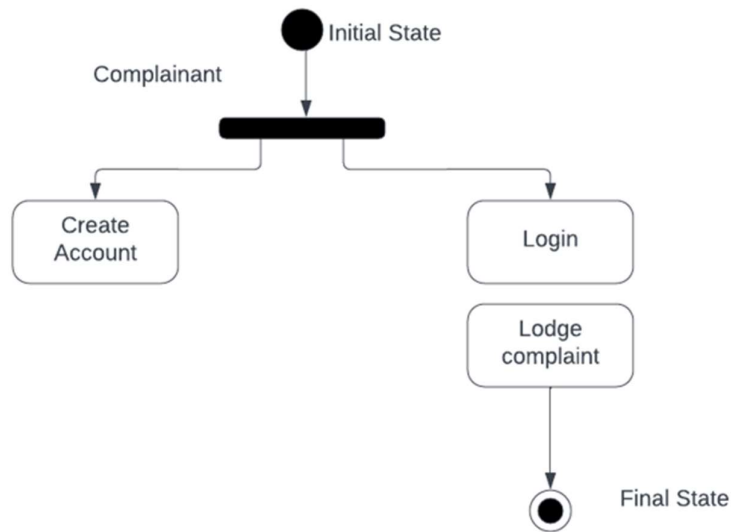


Fig. 4.2: State Chart Diagram for Complainant

The diagram depicts the workflow for a complainant. The complainant needs to first register on the website, in case already registered, the complainant can login into the system and lodge a complaint to be processed by the CID.

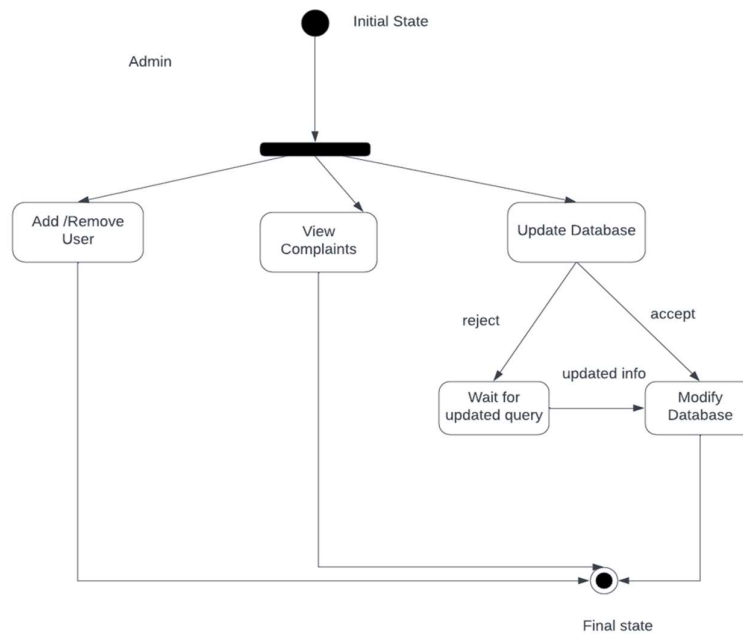


Fig. 4.3: State Chart Diagram for Admin

The administrator has 3 main functionalities- adding/deleting a user, viewing the complaints stored in the database and updating the database whenever needed.

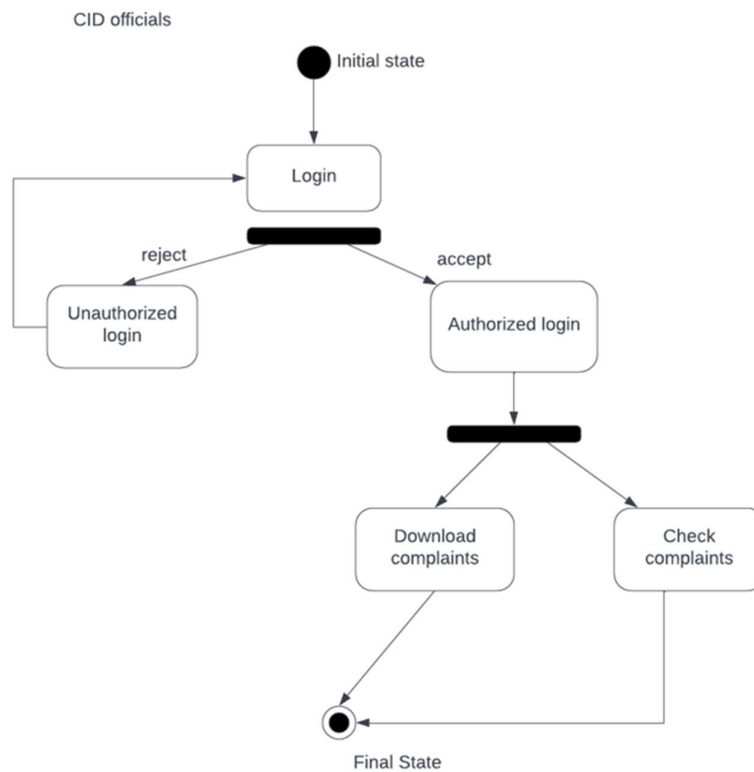


Fig. 4.4: State Chart Diagram for CID Official

The above diagram helps in understanding the workflow for a CID official. An official logs into the system, on successful authentication, he can view and download the segregated complaints.

4.2.2 Activity Diagram

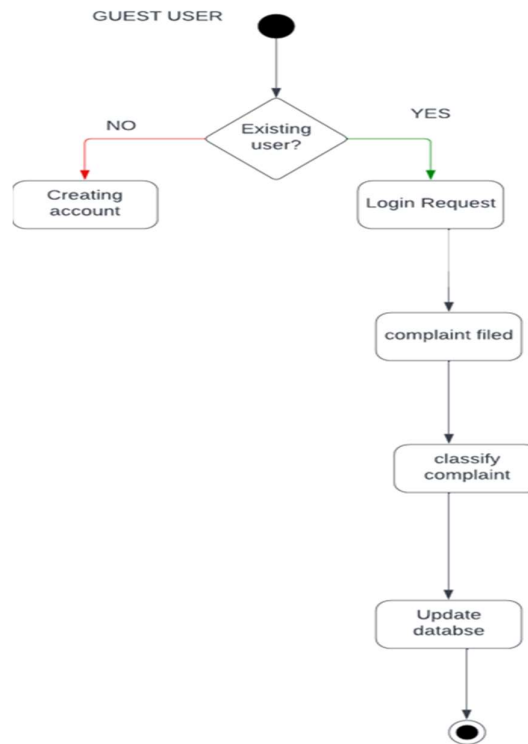


Fig. 4.5: Activity Diagram for Complainant

When a complainant logs in and files a complaint, it is classified by the model and updated in the database.

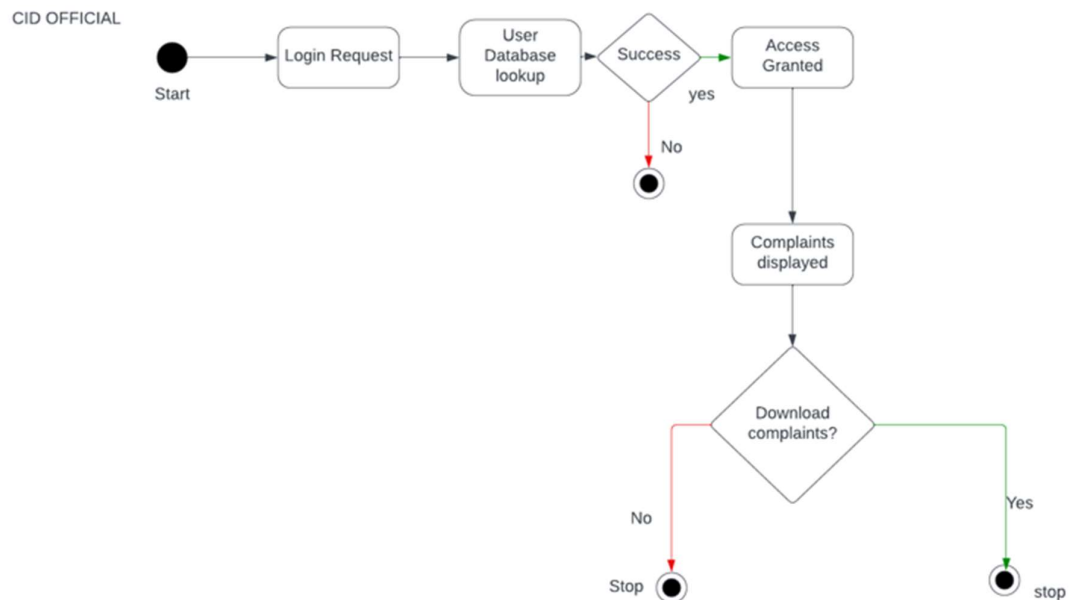


Fig. 4.6: Activity Diagram for CID Official

The above diagram helps in understanding the workflow for a CID official. An official logs in to the system, on successful authentication, the access is granted and he can view and download the segregated complaints.

4.2.3 Component Diagram

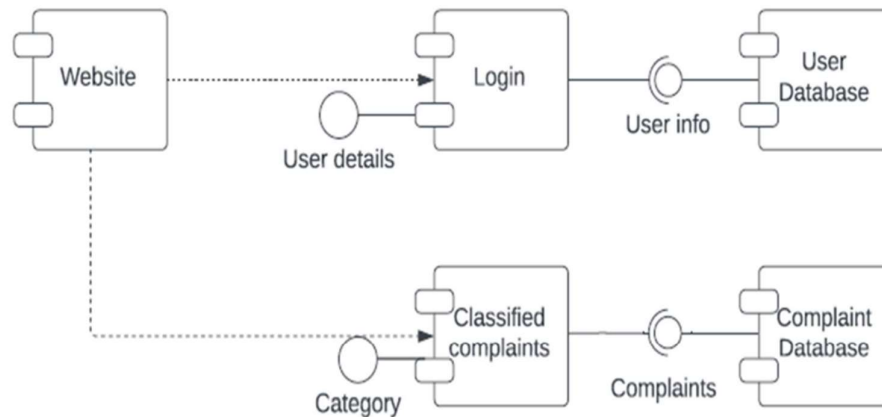


Fig. 4.7: Component Diagram

The above diagram is useful for understanding the overall structure of a system and the interactions between its various components. It aids in managing complexity by breaking down the system into modular and manageable parts, making it easier to design, implement, and maintain the system. It provides a clear and concise visualization of the components and their relationships in a system, helping stakeholders to understand how the different parts of the system interact and collaborate to achieve the overall functionality.

4.2.4 Entity Relationship Diagram

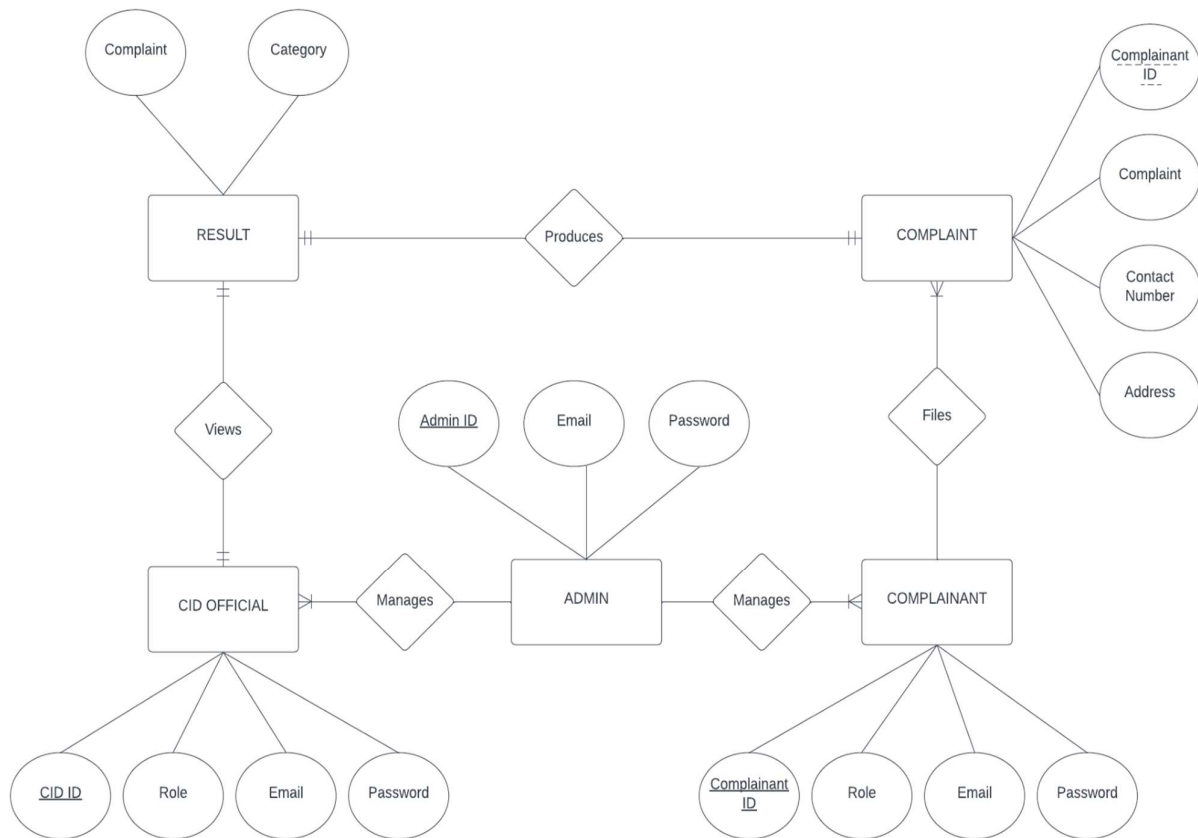


Fig. 4.8: Entity Relationship Diagram

The above diagram shows the architecture of the database which has five tables. All the Users mainly CID Official, Complainant, and Admin have an 'id' assigned which is their primary key. Complaints are stored in another table with the 'Complainant ID' of the complainant as the foreign key. Result i.e., the segregated complaints have category as an attribute which tells about the type of complaint it is (accident/theft/murder/rape/kidnap). The CID Official can view these results.

4.2.5 Class Diagram

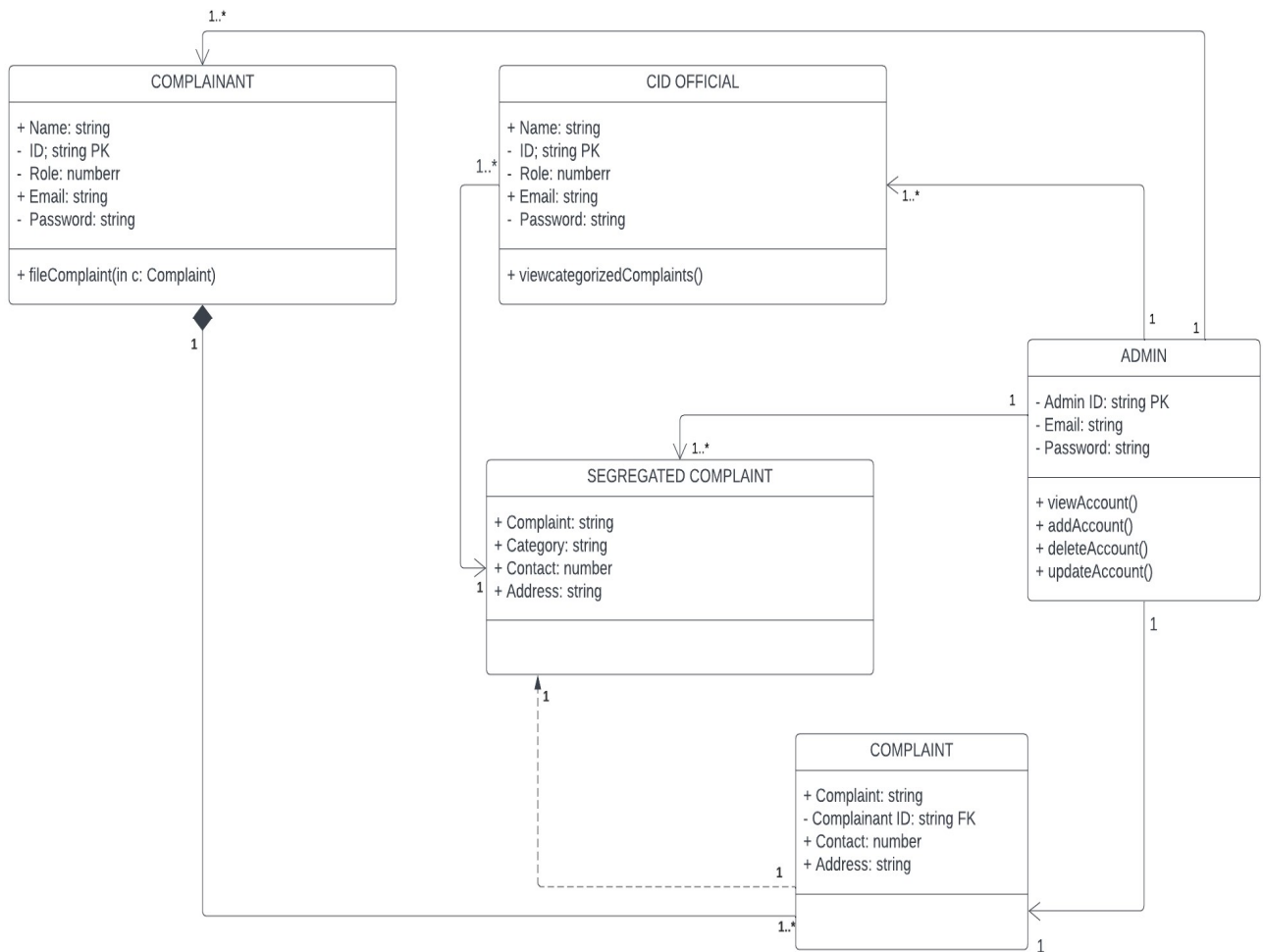


Fig. 4.9: Class Diagram

The above diagram depicts mainly five classes and their relationships along with its functionalities. There are 3 main classes: CID Official, Complainant, and the Admin with their IDs as the primary key. They all have some common features such as email and password and also contain some additional functionalities different from one another. Every complaint is associated with a complainant using the complainant's ID as a foreign key. and has many to one relationship with the complainant class. Admin and CID Officials, Complainants have one to many relationships stating one admin can manage many CID Officials, complainants. Complaints and segregated complaints have one to one relationship.

4.2.6 Data Flow Diagram

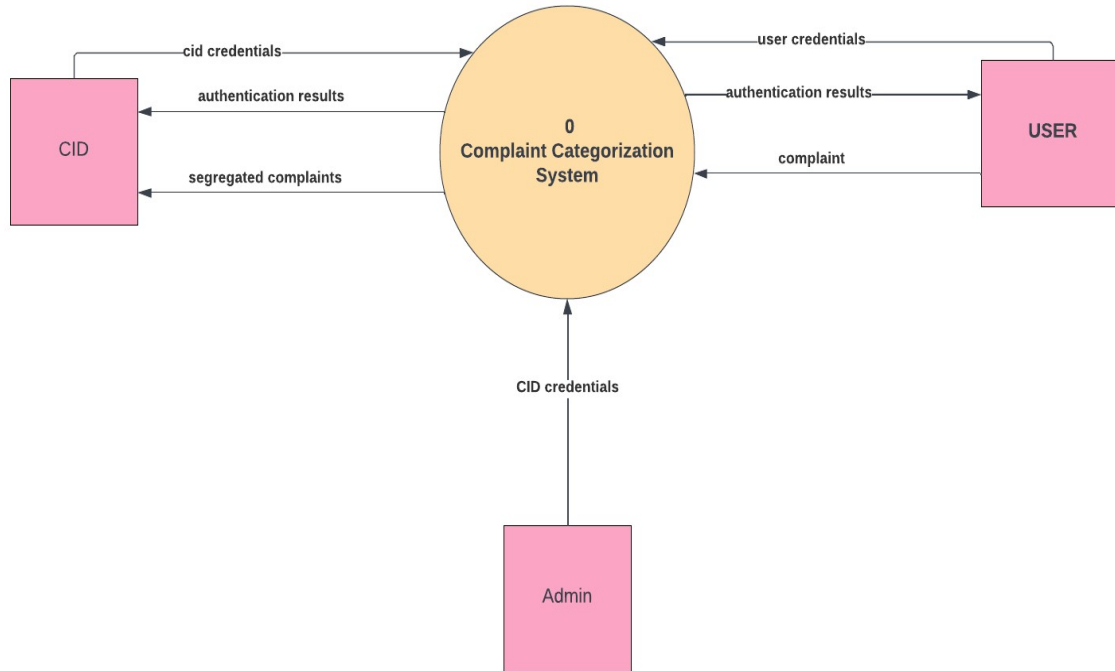


Fig. 4.10: Data Flow Diagram – Level 0

In this Level-0 DFD we can see three entities that interact with our system. The three entities are CID, User i.e., the complainant and the Admin. There is a common interaction of these entities like logging into the system. The admin is the authority which is taking care of registering a CID officer on the platform. The Users can file complaints on our system. The CID Officials can view and download the segregated complaints.

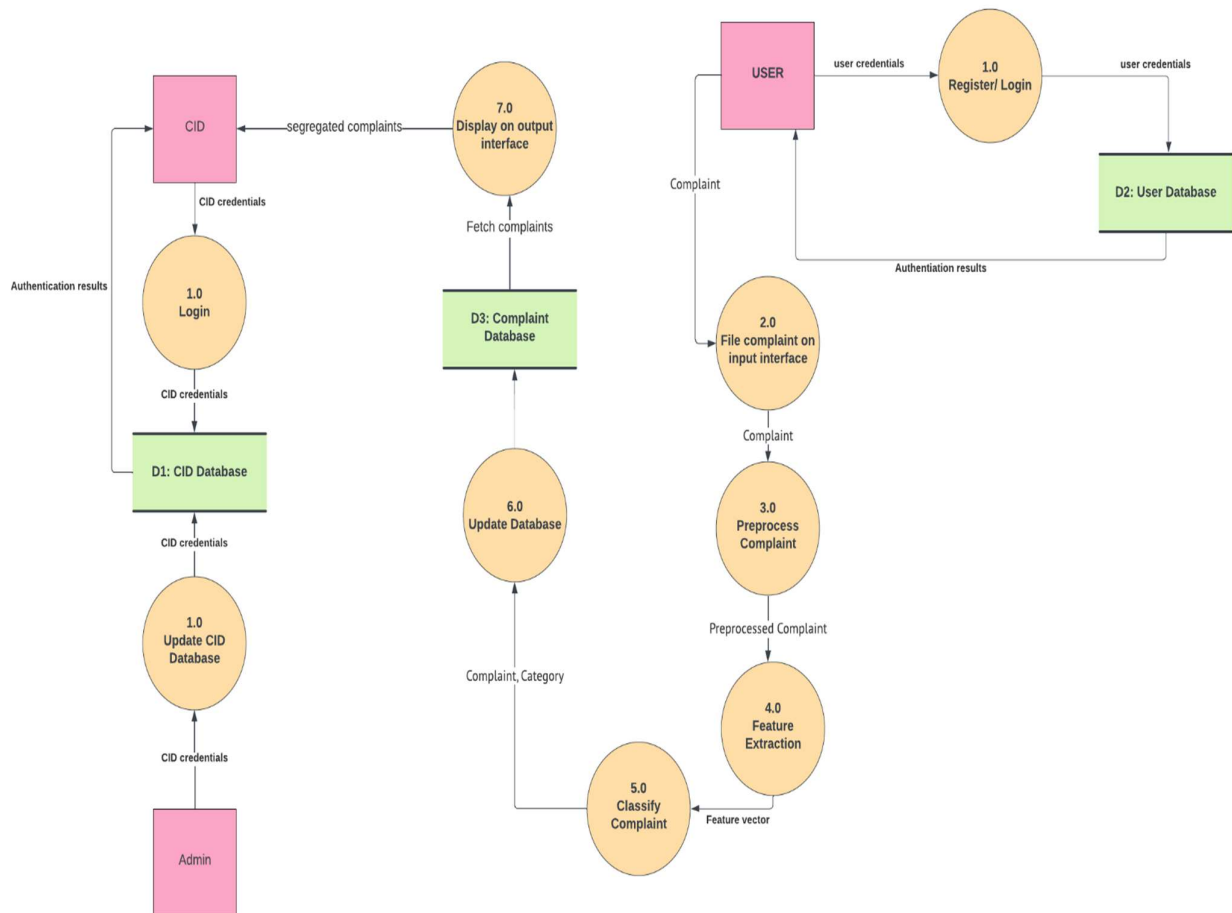


Fig. 4.11: Data Flow Diagram – Level 1

Expanding the system in Level-1 DFD we can see the flow of data inside the system. We can see the different Databases which store and organize information. The complaint filed by the user goes through various processes like Data pre-processing, Feature Extraction and Text classification and then finally stored to the Complaints Database. Then entities defined in the previous level can interact with complaints. Each Entity has its Database to store information regarding Login Credentials.

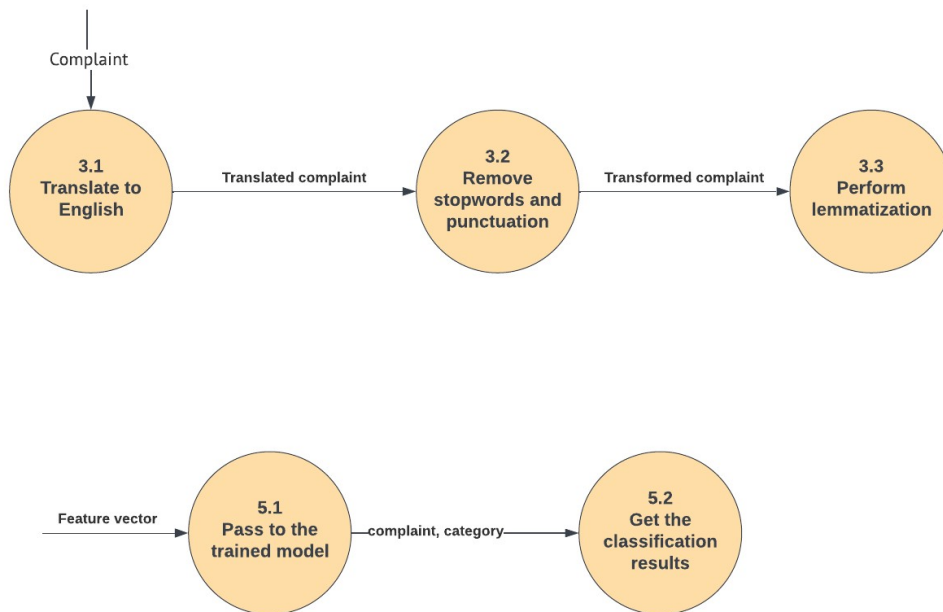
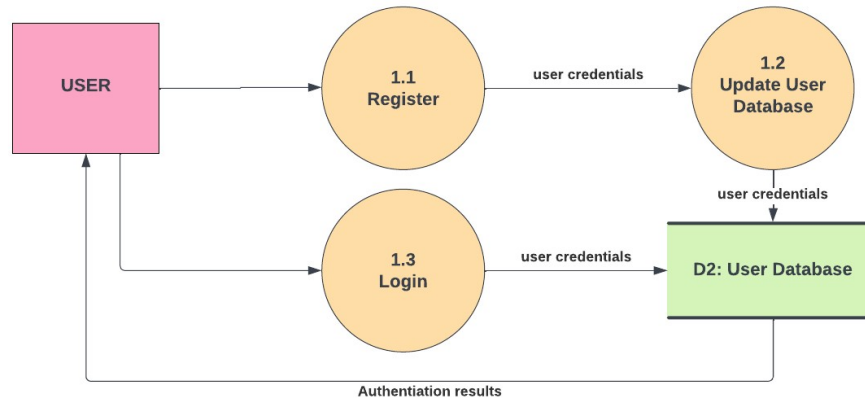


Fig. 4.12: Data Flow Diagram – Level 2

In these Level-2 DFD's we have taken processes from the Level-1 DFD and expanded them to visualize the flow of data in these processes. The processes Expanded are 1. Register/Login, 3. Data Preprocessing which includes translation, stop words removal and lemmatization, and 5. Text Classification which comprises passing feature vector to the trained model and getting the classification results.

4.3 User Interface Diagrams

4.3.1 Use-Case Diagram

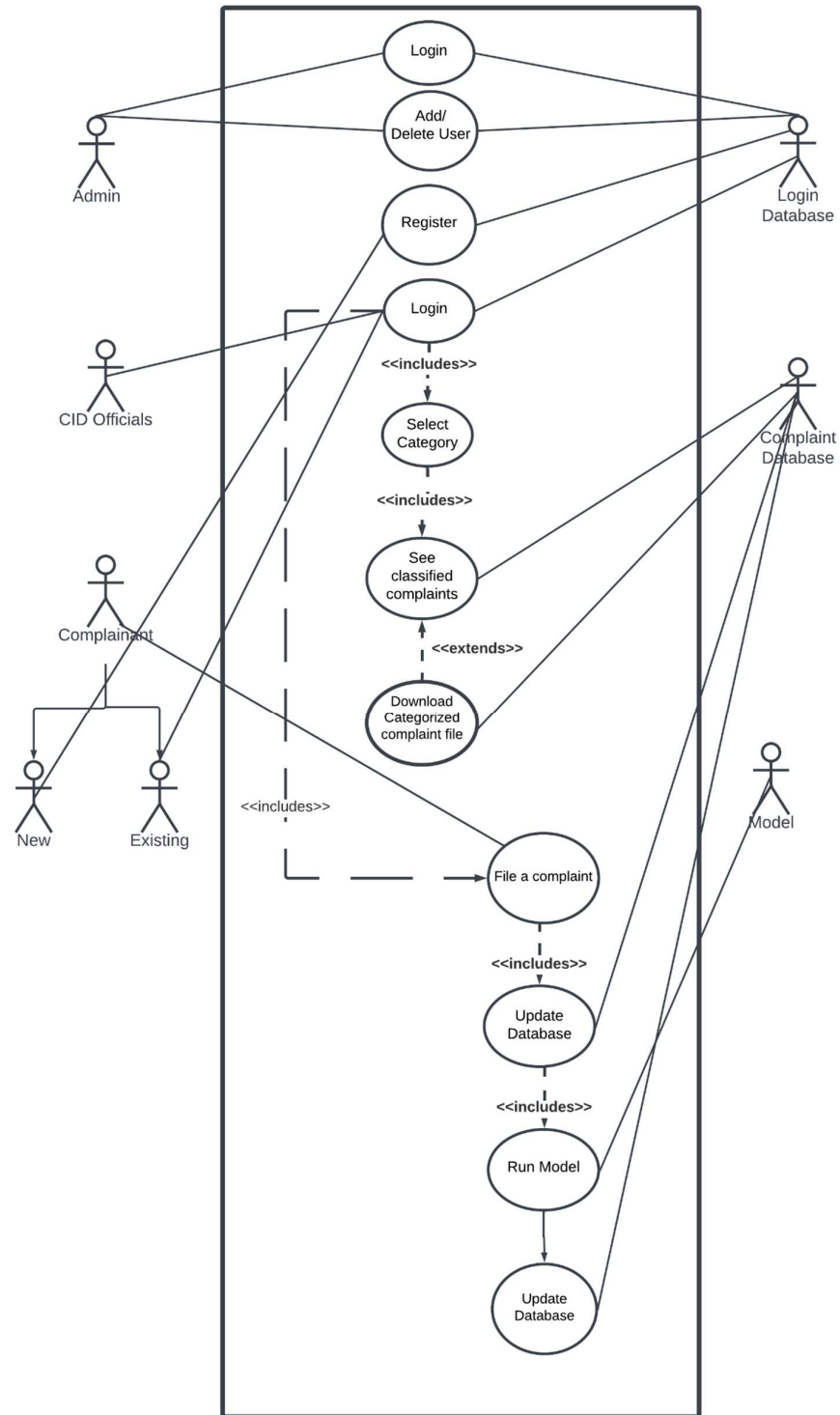


Fig. 4.13: Use Case Diagram

The use case diagram has three actors: admin, CID official and complainant. Admin has the functionality of providing access to CID officials. CID officials authorized by admin can see the

segregated complaints. A new Complainant has to first register on our website, and after login he can file a complaint. Correct category to which the complaint belongs will be assigned to it by the ML model. Login database stores the credentials of the users and complaint database stores all the complaints filed by the complainants.

4.3.2 Use-Case Template

Table 4.1 Use Case Template of Login

Name of Use Case:	Login		
Created By:	Simranjit Kaur	Last Updated By:	Simranjit Kaur
Date Created:	1/05/2023	Last Revision Date:	8/08/2023
Description:	The homepage displays the login button for users.		
Actors:	Admin, Complainant, CID official		
Preconditions:	1. User must be registered.		
Postconditions:	1. The database records are effectively maintained.		
Flow:	1. User should enter the correct credentials. 2. Credentials from the database are matched with the entered credentials. 3. If the credentials match user is logged in successfully and then he/she can perform the desired actions		
Alternative Flows	1. The User is not registered and cannot login. 2. The system displays an error message.		

Table 4.2 Use Case Template of Filing a Complaint

Name of Use Case:	File a complaint		
Created By:	Simranjit Kaur	Last Updated By:	Simranjit Kaur
Date Created:	7/08/2023	Last Revision Date:	10/08/2023
Description:	Complainant can file a new complaint		
Actors:	Complainant, Login Database, Model, Complaint Database		

Preconditions:	<ol style="list-style-type: none"> 1. Complainant must register. 2. Complainant must login into the system. 3. Complainant must submit the form by specifying the complaint.
Postconditions:	<ol style="list-style-type: none"> 1. Complaint will be saved in the database and classified according to the category it belongs to.
Flow:	<ol style="list-style-type: none"> 1. User logs into the system successfully. 2. File the complaint. 3. The system processes the uploaded complaint by cleaning the data to ensure accurate and consistent results. 4. Our complaint classification model is applied to this filed complaint. 5. Complaint is classified to the category it belongs to.
Requirements	<p>The following requirements must be met before execution of the use case</p> <ol style="list-style-type: none"> 1. User must log in into the system. 2. Complaint must be provided to the model.

Table 4.3 Use Case Template of Viewing Segregated Complaint

Name of Use Case:	View Segregated Complaints		
Created By:	Simranjit Kaur	Last Updated By:	Simranjit Kaur
Date Created:	1/05/2023	Last Revision Date:	8/08/2023
Description:	The CID Official has access to the complaint data, which is processed and segregated into different categories based on the trained machine learning model's output.		
Actors:	CID Official, Model, Database		
Preconditions:	<ol style="list-style-type: none"> 1. The stakeholder has a valid user account and login credentials. 2. The system is running and is integrated with a trained machine learning model capable of classifying the complaint data into different categories. 		

Postconditions:	<ol style="list-style-type: none"> 1. The segregated data is displayed to the CID Officials in a user-friendly manner.
Flow:	<ol style="list-style-type: none"> 1. The CID Official logs in to the system using valid user account credentials. 2. He selects the option to view the segregated complaints. 3. The system fetches the complaints from the database. 4. The system displays the segregated data to the CID Officials in an easy-to-understand and user-friendly manner. 5. The CID Official logs out of the system.
Alternative Flows	<ol style="list-style-type: none"> 1. If the CID Official does not have a valid user account credentials, the system prompts them to contact an administrator to request access. 2. If the system fails to integrate with the machine learning model or encounter any other errors during the processing or segregation of the complaint data, the system notifies the stakeholder and prompts them to try again or contact an administrator for assistance.

Table 4.4 Use Case Template of Downloading Segregated Complaint Data

Name of Use Case:	Download Segregated Complaint Data		
Created By:	Garima Chandna	Last Updated By:	Garima Chandna
Date Created:	6/05/2023	Last Revision Date:	6/05/2023
Description:	The CID Official downloads a file containing the segregated complaint data in a supported format from the system.		
Actors:	CID Official, Database		
Preconditions:	<ol style="list-style-type: none"> 1. The CID Official has a valid user account and login credentials. 2. The CID Official has logged into the system. 		

Postconditions:	<ol style="list-style-type: none"> 1. The segregated complaint data is successfully downloaded by the CID Official in a supported format.
Flow:	<ol style="list-style-type: none"> 1. The CID Official logs in to the system using valid user account credentials. 2. The CID Official selects the option to view the segregated complaint data. 3. The CID Official selects the option to download the segregated complaint data. 4. The system generates the file containing the segregated complaint data in the supported format. 5. The file is downloaded to their device. 6. The CID Official logs out of the system.
Alternative Flows	<ol style="list-style-type: none"> 1. If the CID Official does not have valid user account credentials, the system prompts them to contact an administrator to request access.

4.4 Snapshots of Working Prototype

Home Page:



Fig. 4.14: Home Page of StreamlineCID

The main focal point of our home page is a dynamic carousel slider that showcases the diverse range of categories our system adeptly manages. These categories encompass murder, rape, theft, accidents, and kidnappings. Embedded within this page is also comprehensive information about our core functions and services.

**For Admin:
Register a CID Official:**

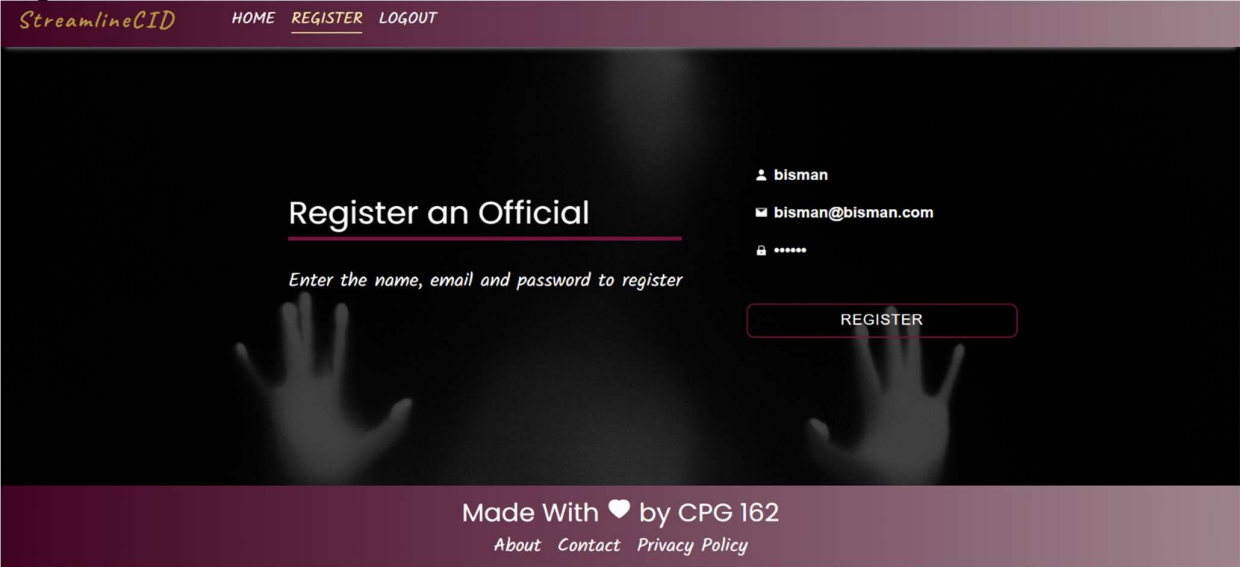


Fig. 4.15: CID Official Registration Page of StreamlineCID

Exclusive authorization to register a CID Official on our website is a distinct privilege reserved solely for our admin. This capability is solely vested in the admin, ensuring a controlled and secure process.

**For Complainant
Register:**

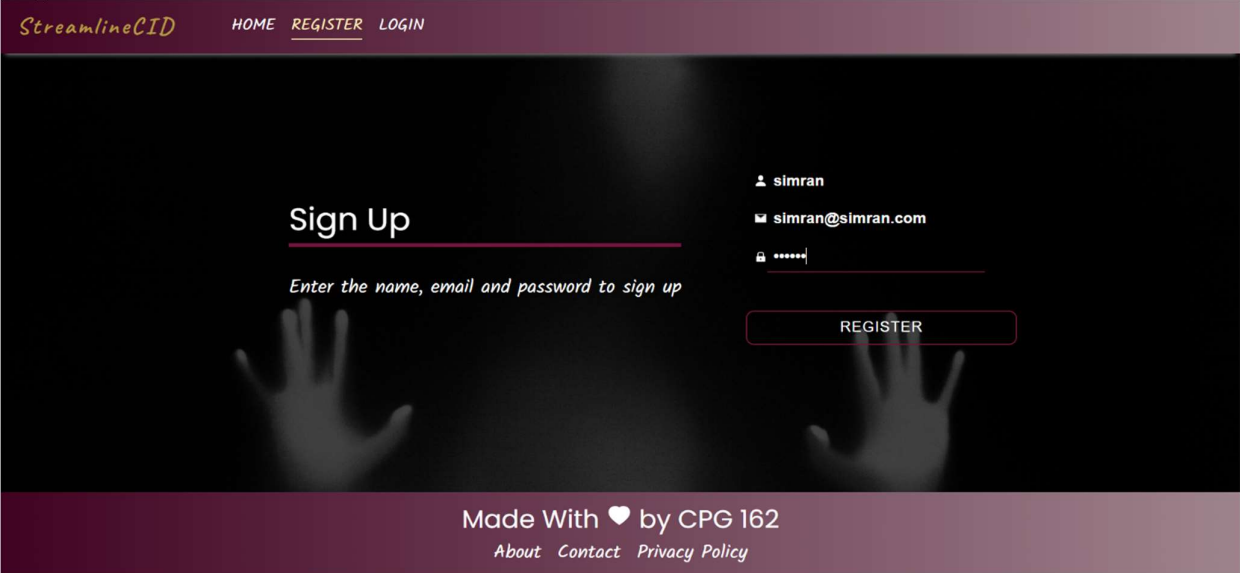


Fig. 4.16: Complainant Registration Page of StreamlineCID

We have a dedicated "Sign Up" page tailored for complainants, offering a straightforward avenue for individuals seeking to file complaints and transmit them to our officials. This user-friendly interface ensures seamless interaction and streamlined communication between complainants and

the designated authorities.

File complaint:

The image displays two screenshots of the 'StreamlineCID' web application's 'File Complaint' page. The page has a dark purple header with the logo 'StreamlineCID' and navigation links: 'HOME', 'LODGE A COMPLAINT', and 'LOGOUT'. The main content area features a background image of a wooden gavel on a wooden surface. The text 'File your complaint here!' and 'Enter the details' is centered at the top of the form area.

Top Screenshot (Before Submission): The form contains four input fields with placeholder text: 'Enter your name', 'Enter your contact number', 'Enter your address', and 'Enter your complaint'. A red dashed border outlines the form area. At the bottom of the form is a purple button labeled 'FILE COMPLAINT'.

Bottom Screenshot (After Submission): The form is filled with the following details: Name: 'Aakanksha', Contact Number: '9947348231', Address: 'Delhi, India', and Complaint: 'भीषण गर्मी होने के कारण गाड़ी का खलासी अपने ही गाड़ी के नीचे सोये हुआ था. वहीं दोपहर के वक्त चालक अपना गाड़ी ज्योहि स्टार्ट कर आगे बढ़ाया कि गाड़ी के नीचे सोये खलासी पर ही गाड़ी चढ़ गया. जिससे उसकी मौत मौके स्थल पर ही हो गई.' The 'FILE COMPLAINT' button is still present at the bottom. A white modal box is overlaid on the page, displaying the message: 'localhost:3000 says Complaint registered successfully!!!!' with an 'OK' button.

Fig. 4.17: Complaint Filing Page of StreamlineCID

Complainants benefit from an intuitive 'File Complaint' page that empowers them to effortlessly upload the pertinent file containing the details of the complaint they wish to register. This feature simplifies the process of formally submitting complaints and ensures that the relevant information is seamlessly conveyed.

For CID Official:

View segregated complaints:

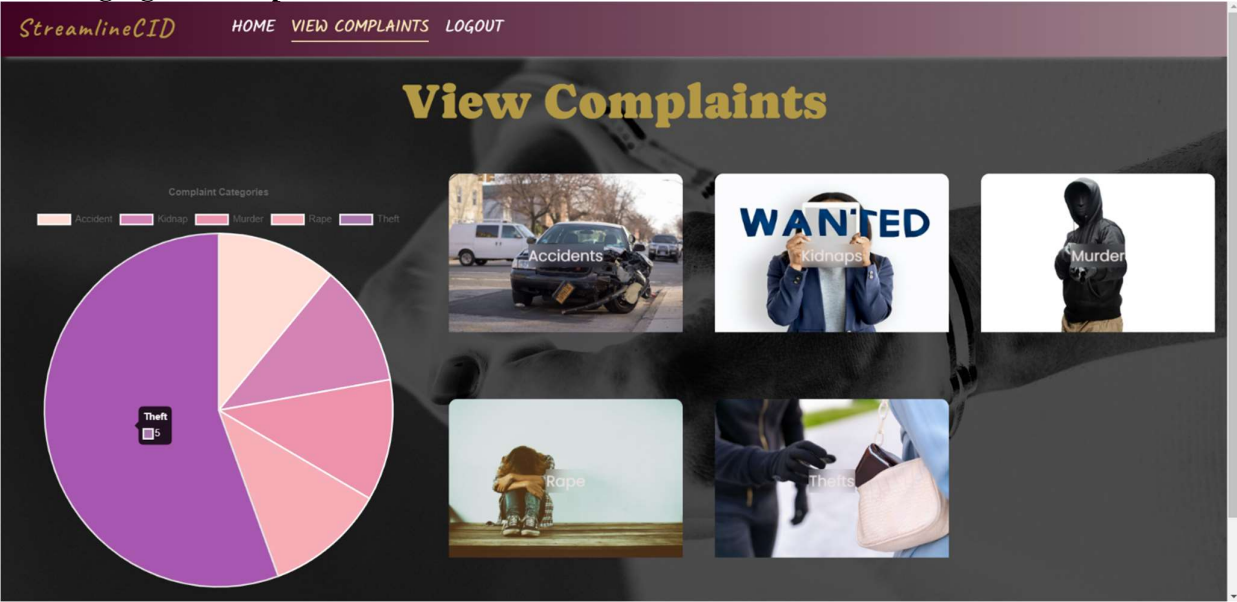


Fig. 4.18: Segregated Complaints Page of StreamlineCID

CID Officials are equipped with a convenient ‘View Complaints’ page that empowers them to visually analyse the number of complaints belonging to each category in the form of a pie chart, and choose the category they wish to explore.

View particular category:

StreamlineCID HOME VIEW COMPLAINTS LOGOUT

Priority Thefts

Sort by: Newest to Oldest

S. No.	Complainant Name	Complaint Address	Complainant Phone	Complaint	Priority
1	Simran	Banga, Punjab	8734231234	हदियाणा के जिन शहर में एक गराज से साइकिल की चोरी हो गई है, चोर ने गराज के दरवाजे को तोड़कर प्रवेश किया है	<input checked="" type="checkbox"/>
2	Rashmeet	Rajpura, Punjab	9947384126	आरोपी कंपनी के कैशियरों की हड्दरी एटिया में टहकर रेकी करता था. आरोपी ने एनआईटी 3 नंबर में जूस की दुकान पर बैठकर रेकी की थी.	<input checked="" type="checkbox"/>

DOWNLOAD

Thefts

Sort by: Newest to Oldest

Thefts

Sort by: Newest to Oldest

S. No.	Complainant Name	Complaint Address	Complainant Phone	Complaint	Priority
1	Simran	Banga, Punjab	8734231234	बिहार के पटना नगर में, मेरे घर के समीप एक वाणिज्यिक कार्यालय से की गई डॉक्यूमेंट्स की बड़ी लूट हुई है, चोरों ने इन दस्तावेजों का अवैध प्राप्त किया है।	<input type="checkbox"/>
2	Simran	Banga, Punjab	8734231234	अमृतसर, पंजाब में, मेरी दुकान से नकद रकम लूट ली गई है, चोरों ने बिड़की के माध्यम से प्रवेश किया और पैसे लूट लिए।	<input type="checkbox"/>
3	Rashmeet	Rajpura, Punjab	9947384126	5 जून को हाइवे पर हुई लूट मामले में क्राइम बॉच टेक्स्ट-48 की टीम ने 4 आरोपियों को गिरफ्तार कर लिया है. आरोपियों से 35 लाख रुपये, सोने की चेन, देखी कट्टा, एक निंद गैद और बाइसात में प्रयोग 2 मोटरसाकिल बरामद हुए हैं. मुख्य आरोपी विकास वर्ष 2020 में 2.5 लाख की लूट को पहले भी अंजाम दे चुका है, जिसमें आरोपी जमानत पर है.	<input type="checkbox"/>

[DOWNLOAD](#)

Made With by CPG 162

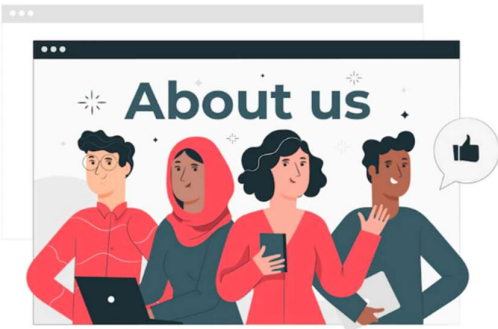
[About](#) [Contact](#) [Privacy Policy](#)

Fig. 4.19: Category Page of StreamlineCID

CID Officials are equipped with a convenient ‘View Category’ page that empowers them to access and analyze segregated complaint data. This functionality enables them to comprehensively review information, delete complaints, set certain complaints as priority, sort complaints and, if necessary, download the data for further examination and action.

About Page:

StreamlineCID [HOME](#) [VIEW COMPLAINTS](#) [LOGOUT](#)



About Us

StreamlineCID is dedicated to revolutionizing the complaint classification process within the Criminal Investigation Department (CID). We understand the critical role accurate complaint classification plays in ensuring effective law enforcement and investigative outcomes. Our mission is to streamline and optimize the classification system, leveraging cutting-edge technologies to deliver faster, more accurate, and data-driven solutions.

For further information, collaboration opportunities, or inquiries, please reach out to us via our contact form or email us at admin@admin.com. Follow us on social media channels to stay updated with the latest developments and industry insights.

Made With by CPG 162

[About](#) [Contact](#) [Privacy Policy](#)

Fig. 4.20: About Page of StreamlineCID

Our "About" page provides a concise overview of the mission and purpose behind StreamlineCID. It encapsulates the essence of our organization's activities and initiatives.

Contact Page:

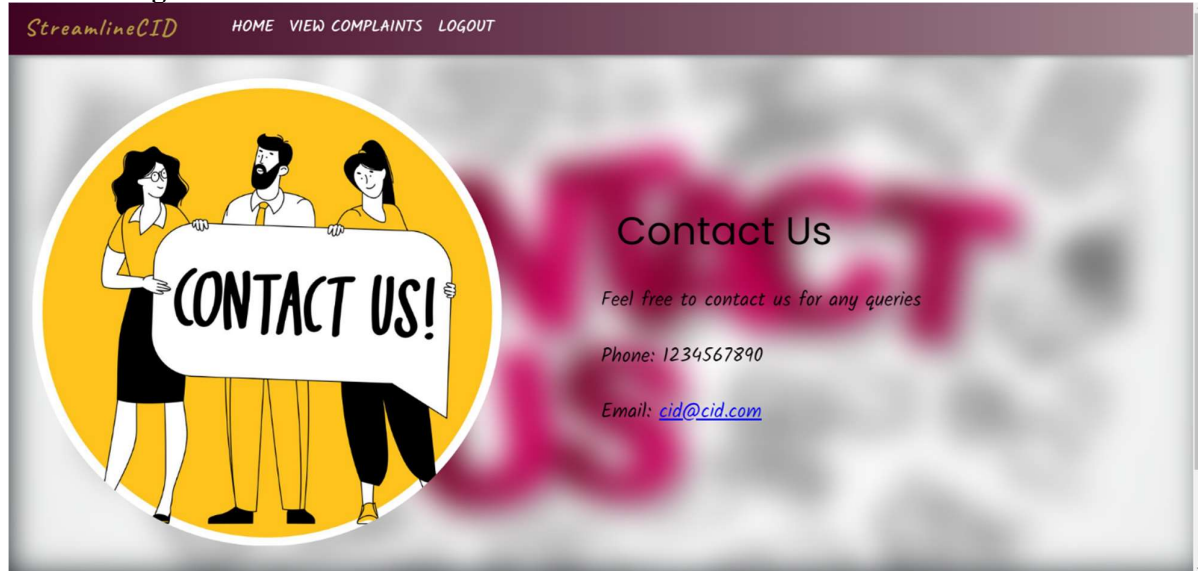


Fig. 4.21: Contact Page of StreamlineCID

Our "Contact" page provides details such as email and phone number for any queries people may wish to contact for.

Privacy Policy Page:

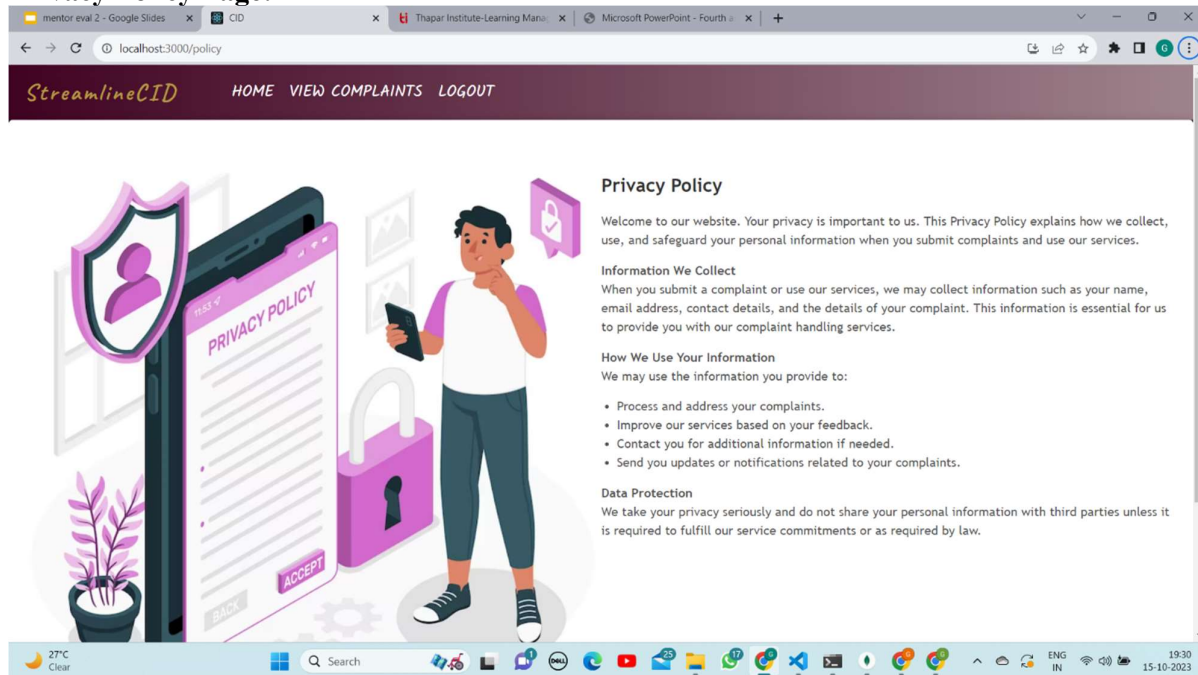


Fig. 4.22: Policy Page of StreamlineCID

Our "Privacy Policy" page outlines how we collect and use the data submitted through our platform.

5. Conclusions and Future Scope

5.1 Work Accomplished

The following objectives have been accomplished which are listed below:

1. The complaint data for training the model has been received from the CID.
2. The administrator can register CID officials on the website.
3. The user is able to login into the system.
4. The users can file complaints on the system.
5. The frontend is linked to the backend where the uploaded complaints are stored.
6. Various models were tested and Support Vector Machine was used as the final model which gave 83% accuracy on the test dataset.

The previously defined objectives have been successfully accomplished through ongoing testing and continuous improvement efforts. Currently, we are in the process of developing a website interface to facilitate complaint lodgment and display them in an organized manner. Moreover, we are actively enhancing the model's accuracy by training it with various algorithms such as SVM, Logistic Regression, Naive Bayes, Random Forest, and Decision Tree using our data. To ensure efficiency and speed, we are eliminating unnecessary logics and adopting more suitable data structures as per the requirements. Our ultimate goal is to create a highly efficient and streamlined system for an enhanced user experience.

5.2 Conclusions

We conclude that all the objectives listed above have been achieved successfully but needs testing to satisfy all the QA checks. All modules have been implemented successfully. The basic version of the website is under improvisation. The team paid great attention to learning new technologies as soon as possible and adopted rapid prototyping strategies of Software Engineering.

5.3 Environmental (/ Economic/ Social) Benefits

As technology continues to advance, society must adapt and evolve accordingly. Whenever a more dependable and convenient technology is available, it should be embraced, and this system achieves just that. The inclusion of a complaint lodging feature through a website form not only lessens an individual's carbon footprint but also does so without any cost. Ensuring affordability is of utmost importance to make this system accessible to all, allowing people to fully enjoy the advantages of modern technology in this era.

5.4 Future Work Plan

The future work intentions are to work on increasing the accuracy of the models implemented as well as making the user interface as simple and user friendly as possible. We intend to link the model in our website so that the segregated complaints are accessible to the CID.

APPENDIX A: References

- [1] National Crime Records Bureau. "Crime in India 2021 (Statistics Volume I)." (2021). Available: https://ncrb.gov.in/sites/default/files/CII-2021/CII_2021Volume%20I.pdf
- [2] Government Data Roundup: Crime in India, Accidental Deaths & Suicides, Prison Statistics, India's External Debt among the data & reports released recently (12 September 2022). Available: <https://factly.in/12-september-2022-government-data-roundup-crime-in-india-accidental-deaths-suicides-prison-statistics-indias-external-debt-among-the-data-reports-released-recently/>
- [3] Yin, Chunyong, et al. "A new SVM method for short text classification based on semi-supervised learning." 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS). IEEE, 2015.
- [4] Yuan, Pingpeng, et al. "MSVM-kNN: Combining SVM and k-NN for multi-class text classification." IEEE international workshop on Semantic Computing and Systems. IEEE, 2008.
- [5] G. Bonisoli, F. Rollo and L. Po, "Using Word Embeddings for Italian Crime News Categorization," 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, Bulgaria, 2021, pp. 461-470, doi: 10.15439/2021F118.
- [6] V. K. Vijayan, K. R. Bindu and L. Parameswaran, "A comprehensive study of text classification algorithms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 1109-1113, doi: 10.1109/ICACCI.2017.8125990.
- [7] Mironczuk, Marcin Michał, and Jarosław Protasiewicz. "A recent overview of the state-of-the-art elements of text classification." Expert Systems with Applications 106 (2018): 36-54.
- [8] A. A. Hakim, A. Erwin, K. Eng, M. Galinium and W. Muliady, "Automated document classification for news article in bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach", 6th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1-4, 2014.

- [9] Z. Qi, "The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2020, pp. 1241-1246, doi: 10.1109/ICAICA50127.2020.9182555.
- [10] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, India, 2016, pp. 112-116, doi: 10.1109/ICETECH.2016.7569223.
- [11] I. Dilrukshi, K. De Zoysa and A. Caldera, "Twitter news classification using SVM," 2013 8th International Conference on Computer Science & Education, Colombo, Sri Lanka, 2013, pp. 287-291, doi: 10.1109/ICCSE.2013.6553926.
- [12] A. Deniz and H. E. Kiziloz, "Effects of various preprocessing techniques to Turkish text categorization using n-gram features," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 2017, pp. 655-660, doi: 10.1109/UBMK.2017.8093491.
- [13] D. E. Brown, "The Regional Crime Analysis Program (ReCAP): a framework for mining data to catch criminals," SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218), San Diego, CA, USA, 1998, pp. 2848-2853 vol.3, doi: 10.1109/ICSMC.1998.725094.
- [14] F. Herrera, R. Sosa and T. Delgado, "GeoBI and Big VGI for Crime Analysis and Report," 2015 3rd International Conference on Future Internet of Things and Cloud, Rome, Italy, 2015, pp. 481-488, doi: 10.1109/FiCloud.2015.112.