# ABSTARCT

Social media text analytics is the process of deriving information from text sources. Text analysis can be applied to any text-based dataset, including social media. Crime is a major problem faced today by society even we are finding in social media. Crimes have affected the quality of life and economic growth badly. It can identify the crime patterns and predict the crimes by detecting and analyzing the historical data. However, some crimes are unregistered and unsolved due to a lack of evidence. Thus, detecting crimes is a still challenging task. Some can use social media to detect crimes related activities. Because social media users sometimes convey messages related to his or her surrounding environment via social media message. It is proposed as a machine learning approach to detect the crimes and analyses it on its type. As the first step, we fetch the text messages using predefined keywords relating to the crimes. Then, after the preprocessing, we applied a support vector machine- based filtering approach to eliminate the noise. And then Random forest is used for classification . Then in the final stage, it analyses and categorize the crime type.

**TABLE OF CONTENTS**

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| ML | MACHINE LEARNING |
| SVM | SUPPORT VECTOR MACHINE |
| RFA | RANDOM FOREST ALGORITHM |
| NLTK | NATIONAL LANGUAGE TOOL KIT |
| NLP | NATIONAL LANGUAGE PROCESSING |
| ROI | RETURN ON INVESTMENT |

# CHAPTER 1

# INTRODUCTION

## 1.1 OUTLINE OF THE PROJECT

Security is a very necessary aspect of life. Unless we are safe, our most important needs cannot be met. Security is therefore a requirement in human life that helps us to achieve our goals collectively or individually. Crimes are a social problem, which costs our society deeply in many aspects. The ability to identify unsafe areas for crime and identify the most recent crime in a particular location has become a growing concern for both local authorities and residents. On the other hand, people are always interested in improving safety and make reliable relationships with neighbors when living in a busy society. The prevalence of crime is one of the greatest challenges for societies around the world, particularly in metropolitan areas There are more researches regarding social crimes in the world, but using social media, there are few types of research about crimes and their behavior. Therefore, the paper aims in presenting a prediction model (algorithm) by using the machine-learning technique, which is meant to possess a strong capability to predict crimes by factors of social media dataset using the Data Mining concept. Our main data source is social media. The main goal is to identify each hidden data source and predict results.

## 1.2 OBJECTIVES OF THE PROJECT

- The main objective of the project is to predict the crime rate and analyze the crime rate to be happened in future. Based on this Information the officials can take charge and try to reduce the crime rate.
- The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable)

- The system will look at how to convert crime information into a regression problem, so that it will help detectives in solving crimes faster.
- Crime analysis based on available information to extract crime patterns. Using various multi linear regression techniques, frequency of occurring crime can be predicted based on territorial distribution of existing data and Crime recognition.

## 1.3 DOMAIN INTRODUCTION

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

### *1.3.1 Basics of Python*

Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages. Python runs on an interpreter system, meaning that code can be executed as soon as it is written.

### *1.3.2 Python features*

Python is a dynamic, high level, free open source and interpreted programming language. It supports object-oriented programming as well as procedural oriented programming. In Python, we don't need to declare the type of variable because it is a dynamically typed language.

# CHAPTER 2
# AIM AND SCOPE

## 2.1 AIM OF THE PROJECT

- Crime detection is aimed for detecting the crime in social media platforms.
- It will be detecting the social media crime by using of machine learning concept. It used the data set for detecting the crime.

## 2.2 SCOPE OF THE PROJECT

- Future work, we planned to evaluate other machine learning algorithms, such as neural networks and decision trees. We also plan to implement the crime prediction approach using SVM and deep neural network approach.
- Also, we intend to expand our analysis to include spatial and temporal analysis to find out when and where crime has spread in the past, and when and where it is most likely to spread in the future.

## 2.3 SOFTWARE REQUIREMENTS

- Operating system: Windows 7.
- Coding Language: python
- Tool: anaconda, visual studio code
- Libraries: OpenCV

## 2.4 HARDWARE REQUIREMENTS

- System : Pentium i3 Processor.
- Hard Disk : 500 GB.
- Monitor : 15’’ LED

- Input Devices : Keyboard, Mouse, Ram : 2 GB

# CHAPTER – 3

# METHODS AND ALGORITHMS USED

## 3.1 DATA COLLECTION

- Social media posts are collected through the Search API .
- The search of the social media posts must be based on a set of keywords that can be used to classify the crime situations.
- Thus in the first filter, we used the main crime-related keywords according to crime categories.

## 3.2 DATA CLEANING

- Data cleaning is a critically important step in any machine learning project.
- Data Cleaning is done to prepare the data for analysis by removing or modifying the data that may be incorrect, incomplete, duplicated or improperly formatted.
- There are many different statistical analysis and data visualization techniques you can use to explore your data in order to identify data cleaning operations you may want to perform.

## 3.3 DATA PREPROCESSING

- As in fig 3.1, it is very important to apply the pre- processing techniques to the extracted data set.
- Because, there may be typos, unwanted content like URLs, and stop words in the social media post.
- Thus, data which is obtained from social media is highly unstructured and noisy.

- Pre-processing techniques will generate clean tweet data that will be used for the next process.
- First, we removed the stop words such as is, the, which, have, etc. The words do not convey any positive or negative meaning. So, we can easily remove the stop word without affecting the meaning of the message.
- Then, we removed URLs, hashtags, symbols, usernames, expressions, quotes, etc. Next, combine words are split by applying tokenizing techniques.
- Finally, we applied a stemming algorithm to reduce a word to its word stem that affixes to suffixes and prefixes or the roots of words.

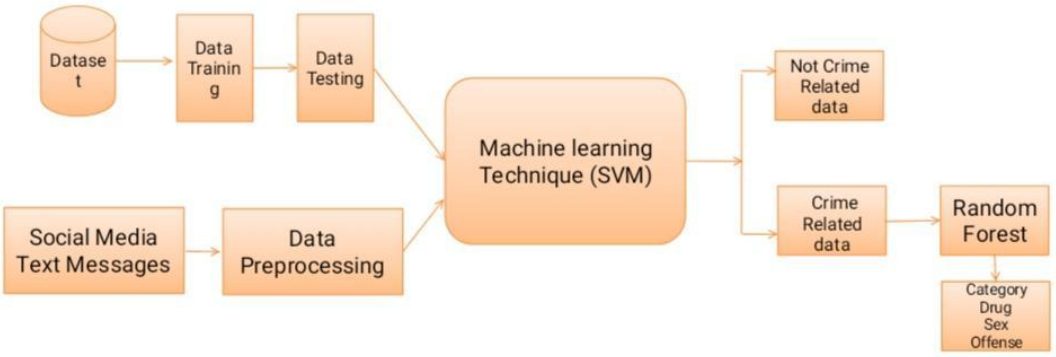


***Fig 3.1 System Architecture***

### 3.4 FEATURE EXTRACTION

- Feature Extraction is done to reduce the number of attributes in the dataset hence providing advantages like speeding up the training and accuracy improvements.
- In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.