

**VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF  
TECHNOLOGY**  
**Department of Computer Engineering**



Project Report on

**DIGITAL DATA FORENSICS**

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree  
in Computer Engineering at the University of Mumbai Academic Year  
2017-2018

**Submitted by**

Ekta Chawla (D17A , Roll no -15 )  
Sahil Jagiasi (D17A , Roll no - 34)  
Jaikumar Kukreja (D17A , Roll no - 41)

**Project Mentor**

Mrs. Pooja Nagdev

(2017-18)

# VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

## Department of Computer Engineering



## Certificate

This is to certify that ***Ekta chawla, Sahil Jagiasi, Jaikumar Kukreja*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on “***DIGITAL DATA FORENSICS***” as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Prof. Pooja Nagdev*** in the year 2017-2018 .

This project report entitled ***Digital Forensics*** by ***Ekta Chawla, Sahil Jagiasi, Jaikumar Kukreja*** is approved for the degree of Bachelor of Engineering (Computer Science).

Programme Outcomes	Grade
PO1, PO2, PO3, PO4, PO5, PO6, PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

-----

# Project Report Approval For B. E (Computer Engineering)

This thesis/dissertation/project report entitled *Digital Data Forensics* by *Ekta Chawla, Sahil Jagiasi, Jaikumar Kukreja* is approved for the degree of Bachelor of Engineering (Computer Science).

Internal Examiner

-----

External Examiner

-----

Head of the Department

-----

Principal

-----

Date:

Place:

# Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-----  
(Signature)

Ekta Chawla

-----  
(Signature)

Sahil Jagiasi

-----  
(Signature)

Jaikumar Kukreja

Date:

## ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Pooja Nagdev** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair** , for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

**Computer Engineering Department**  
**COURSE OUTCOMES FOR B.E PROJECT**

Learners will be to,

<b>Course Outcome</b>	<b>Description of the Course Outcome</b>
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solution for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

## **ABSTRACT**

The recent development in Information Communication Technology (ICT) has made changes in every aspect of our Life. These changes are taking us towards the dream of “DIGITAL INDIA”. The positive influence of Digital world on Knowledge, trade and business and Communication is no doubt remarkable. However, the dark side of it deteriorates its peaceful usage that is DigitalCrimes. Digital Crimes are defined as any illegal activities practiced by or done via digital device. Unlike “traditional “crimes Digital crimes present a real dilemma due to the fact that criminals’ identity may be hidden.

The concept of Digital Forensics has come to the existence in an attempt of formulating possible ways for digital crimes investigation and analysis process. In this report, we come across various branches of Digital Forensics along with the process of finding the digital evidence and tools used in digital forensics.

The proliferation of phones (particularly smartphones) on the consumer market has caused a growing demand for forensic examination. As a matter of fact, Law enforcement are much more likely to encounter a suspect.

The average mobile device user sends a large quantity of text and other short messages on social media such as twitter and facebook. These text message data are of great value to law enforcement investigators who may be analyzing a suspect’s social media profile for evidence of criminal activity. By Applying Natural Language Processing, machine learning principles and feature extraction Methodology one can speed up the process of analyzing and finding the sender of the text through SCAP process.

## INDEX

<b>Chapter No.</b>	<b>Title</b>	<b>Page No.</b>
	<b>Certificate of Approval</b>	2
	<b>Declaration</b>	4
	<b>Acknowledgement</b>	5
	<b>Abstract</b>	7
<b>1</b>	<b>Introduction</b>	<b>11</b>
1.1	Introduction to the project	11
1.2	Motivation for the project	11
1.3	Problem Definition	12
1.4	Drawback of the existing system	12
1.5	Relevance of the Project	12
1.6	Methodology used	13
<b>2.</b>	<b>Literature Survey</b>	<b>14</b>
2.1	Research Papers - Mentioned in IEEE format	14



2.2	Study of Patent	19
<b>3.</b>	<b>Requirement Of Proposed System</b>	<b>21</b>
3.1	Functional Requirements	21
3.2	Non-Functional Requirements	22
3.3	Constraints	22
3.4	Hardware & Software Requirements	22
3.5	Algorithms utilized in the existing systems	23
<b>4.</b>	<b>Proposed Design</b>	<b>25</b>
4.1	System Design / Conceptual Design	25
4.2	Block diagram representation of the proposed system	26
4.3	Design of the proposed system	27
4.3.a	Data Flow Diagram ( Level 0,1,2)	27
4.3.b	State Transition Diagram/ Activity Diagram	28
4.5	Project Scheduling & Tracking using Timeline / Gantt Chart	29

<b>5.</b>	<b>Implementation</b>	<b>30</b>
<b>6.</b>	<b>Testing</b>	<b>31</b>
<b>7.</b>	<b>Results obtained</b>	<b>33</b>
<b>8.</b>	<b>Conclusion</b>	<b>39</b>
<b>9.</b>	<b>References</b>	<b>40</b>
<b>10.</b>	<b>Appendix</b>	<b>41</b>
10.1	List Of Figures	41
10.2	List Of Tables	41
10.3	Paper Publications	42
10.3.1	Draft of the paper published.	45
10.3.2	Plagiarism report of the paper published /draft	47
10.3.3	Draft of paper 2	49

# CHAPTER 1: INTRODUCTION

## 1.1 What is Forensics?

- The word forensics comes from latin word “FORENSICS” ,which means “before the forum” and refers to something “of,pertaining to,used in a court of law and detection of crime” .
- Forensics is an investigative technique used to help solve crimes.
- Subdivisions of Forensics.
  - Art Forensics
  - Forensics accounting
  - Forensics anthropology
  - Forensics DNA Analysis
  - **Digital Forensics** and many more

## 1.2 MOTIVATION

- Digital forensics has become a critical component of both civil and criminal cases.
- Slowly being recognized as important by nontechnical groups
  - a) Judges and lawyers.
  - b) Law enforcement.
  - c) Business entities.
- Has been some progress in defining recognized good practices in forensics application.
- Most, aimed at collection of evidence from typical systems.
- There is still a lack of widely accepted theoretical models or principles.
- Has been some progress in defining recognized good practices in forensics application.
- Most, aimed at collection of evidence from typical systems.
- There is still a lack of widely accepted theoretical models or principles.
- Creates problems in specifying or designing systems capable of capturing digital forensics evidence.

## **1.3 PROBLEM DEFINITION**

The means of communication has changed over time according to the situation and advancements in technology. The process of transferring data from one individual to another such as audio, video and images have grown beyond texting and evolved to enable the transmission of media not only between two individuals but also in a group where huge number of people can interact and have a talent to connect worldwide.

For eg WhatsApp and twitter is such an application which is used widely for transferring media, text, files as well as audio calling. This project finds the messages which are related to crime and predicts the content is positive or negative. And basically, it focuses on the source of the text sent. It determines the author of the text being transferred by using averages and the standard deviation of the words in a sentence.

## **1.4 DRAWBACKS OF EXISTING SYSTEM**

- Hardware, software, and application diversity.
- A proliferation of data file formats, many of which were poorly documented.
- Heavy reliance on time-sharing and centralized computing facilities
- rarely was there significant storage in the home of
- The absence of formal process, tools, and training.

## **1.5 RELEVANCE OF THE PROJECT**

The project explains the hidden evidence acquisition from file system. Secondly it explains investigation on the Network. There are two types of investigation in network, live data acquisition (Packet capturing and analysis) and the hidden evidence acquisition from file system. Second section explains investigation on the Network. The input of the system is the data of the author. Firstly, we train the system by calculating the average number of prepositions, nouns, and the buts used in the system by different registered authors.

Advantages of proposed system:

- The widespread use of Microsoft Windows, and specifically Windows XP.
- Relatively few file formats of forensic interest mostly
- Microsoft Office for documents.
- Examinations largely confined to a single computer system belonging to the subject of the investigation.
- Storage devices equipped with standard interfaces (IDE/ ATA), attached using removable cables and connectors, and secured with removable screws.
- Multiple vendors selling tools that were reasonably good at recovering allocated and deleted files.
- Higher better performance algorithms are introduced to extract, collect, analyze data.

## 1.6 METHODOLOGY

A text message corpus has been developed and basic experiments were conducted in order to show information about the corpus and to demonstrate natural language processing (NLP) principles and machine classification based on supervised learning algorithms. Applicability and limitations of the corpus are discussed. A simple methodology using Naive Bayes algorithm is used for determining positive and negative sentences. Naive Bayes classifier gives upto 70% accuracy because it works on probability,

On the other hand, In stylometry, source code author profile is implemented in which in training set the average number of nouns, prepositions, verbs, articles, ands, standard deviation is calculated. Now in testing the average value of the input text is calculated and the nearest value to the average value of the training data is compared and the nearest value is considered as the sender of the text.

## CHAPTER 2: LITERATURE SURVEY

### 2.1 PAPERS

Literature survey is the most important step in software development process. Following is the literature survey of some existing technique for digital forensics :

**2.1.1 Digital Forensics and Cyber Crime Data Mining ,K. K. Sindhu , B. B. Meshram ,Journal of Information Security, 2012, 3, 196-201**

Digital forensics is the science of identifying, extracting, analyzing and presenting the digital evidence that has been stored in the digital devices. Various digital tools and techniques are being used to achieve this. Our paper explains forensic analysis steps in the storage media, hidden data analysis in the file system, network forensic methods and cyber crime data mining. This paper proposes a new tool which is the combination of digital forensic investigation and crime data mining. The proposed system is designed for finding motive, pattern of cyber attacks and counts of attacks types happened during a period. Hence the proposed tool enables the system administrators to minimize the system vulnerability.

**2.1.2 Forensic Investigation of Social Media and Instant Messaging Services in Firefox OS: Facebook, Twitter, Google+, Telegram, OpenWapp and Line as Case Studies, Mohd Najwadi Yusoff, Ali Dehghantanha, Ramlan Mahmod, Pages 41-62, Chapter 4, (Elsevier) Contemporary Digital Forensic Investigations Of Cloud And Mobile Applications, 2017.**

Mobile devices are increasingly utilized to access social media and instant messaging services, which allow users to communicate with others easily and quickly. However, the misuse of social media and instant messaging services facilitated conducting different cyber crimes such as cyber stalking, cyber bullying, slander spreading and sexual harassment. Therefore, mobile devices are an important evidentiary piece in digital investigation. In this

chapter, we report the results of our investigation and analysis of social media and instant messaging services in Firefox OS. We examined three social media services (Facebook, Twitter and Google+) as well as three instant messaging services (Telegram, OpenWapp and Line). Our analysis may pave the way for future forensic investigators to trace and examine residual remnants of forensics value in FireFox OS.

<b>GROUP</b>	<b>APPLICATION</b>	<b>MOBILE WEB</b>
<b>SOCIAL MEDIA</b>	1.Facebook 2.Twitter	1.Facebook 2.Twitter 3.Google+
<b>INSTANT MESSAGING APP</b>	1. Telegram 2. OpenWapp 3. Line	1. Telegram

Table-1:Social Media

### **2.1.3 Bringing science to digital forensics with standardized forensic corpora, Simson Garfinkel, Paul Farrell , Vassil Roussev C , George Dinolt, 2009 Digital Forensic Research Workshop. Published by Elsevier Ltd.**

Progress in computer forensics research has been limited by the lack of a standardized data sets corpora that are available for research purposes. We explain why corpora are needed to further forensic research, present a taxonomy for describing corpora, and announce the availability of several forensic data sets. Much of the work to date in digital forensics has focused on data extraction and for presentation in courts. Researchers have developed technologies for copying data from subject hard drives, storing that data in a disk image file, searching the disk image for document files, and presenting the documents to an examiner. As both the variety and scale of forensic investigations increase, forensic practitioners need tools that do more than search and present: they need tools for reconstruction, analysis,

clustering, data mining, and sense-making. Such tools frequently require the development of new scientific techniques in areas such as text mining, machine learning, visualization, and related fields. One of the hallmarks of science is the ability for researchers to perform controlled and repeatable experiments that produce reproducible results. Science is based on the principle that phenomena can be observed and results can be reproduced by anyone there are no privileged experimenters or observers (given sufficient training and financial resources, of course). Sadly, much of today's digital forensic research results are not reproducible. For example, techniques developed and tested by one set of researchers cannot be validated by others since the different research groups use different data sets to test and evaluate their techniques.

#### **2.1.4 Text Message Corpus: Applying Natural Language Processing To Mobile Device Forensics Daniel R. O'Day And Ricardo A. Calix, Multimedia And Expo Workshops (ICMEW), 2013 IEEE International Conference**

A text message corpus has been developed and basic experiments were conducted in order to show information about the corpus and to demonstrate natural language processing (NLP) principles and machine classification based on supervised learning algorithms. Applicability and limitations of the corpus are discussed. A simple methodology for extracting features from the corpus is proposed. The alternate approach taken was to extract bigrams (two-word pairs) as features and to allow the algorithm to determine which bigrams were most effective in classifying text messages as drug-related or neutral. A unigram feature set would have a positive hit on the word 'weed' in a text message containing this text:

"After school today let's go smoke some weed at my house."

However, the same unigram feature extractor would also receive a positive hit on a text message containing this text:

"Hey pull that weed in my flower garden when you get home."

While the first text message was drug-related, the second was neutral and would therefore be a false positive using the standard unigram feature extraction methodology tied to known



drug-related unigrams. The hypothesis was that drug related terms would exist in frequented bigrams, such as “smoke weed,” “mary jane,” “hit acid,” “pop pilz,” etc. and that these bigrams would increase classification accuracy. In addition, common “stopwords” were removed from consideration as features (as determined by the “stopwords” corpus within NLTK). All words were treated as case insensitive. Stemming was performed in order to identify word stems, so that words such as “smoke,” “smoker,” and “smoking” would be counted as one term rather than three. Punctuation was also removed.

### **2.1.5 Source Code Author Identification Based on N-gram Author Profiles by Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, Sokratis Katsikas, Laboratory of Information and Communication Systems Security**

Source code author identification deals with the task of identifying the most likely author of a computer program, given a set of predefined author candidates. This is usually based on the analysis of other program samples of undisputed authorship by the same programmer. There are several cases where the application of such a method could be of a major benefit, such as authorship disputes, proof of authorship in court, tracing the source of code left in the system after a cyber attack, etc. We present a new approach, called the SCAP (Source Code Author Profiles) approach, based on byte-level n-gram profiles in order to represent a source code author's style. Experiments on data sets of different programming language (Java or C++) and varying difficulty (6 to 30 candidate authors) demonstrate the effectiveness of the proposed approach. A comparison with a previous source code authorship identification study based on more complicated information shows that the SCAP approach is language independent and that n-gram author profiles are better able to capture the idiosyncrasies of the source code authors. Moreover the SCAP approach is able to deal surprisingly well with cases where only a limited amount of very short programs per programmer is available for training. It is also demonstrated that the effectiveness of the proposed model is not affected by the absence of comments in the source code, a condition usually met in cyber-crime cases.

### **2.1.6 Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method, Laboratory of Information and Communication Systems Security Department of Information and Communication Systems Engineering, University of the Aegean**

In this paper we present a new approach, which we call the SCAP (Source Code Author Profiles) method, based on byte-level n-gram –or sequential slicing– profiles representing the source code author’s style. The SCAP method extends an approach originally applied to natural language text authorship attribution by Keselj et. al. (2003). We show that the n-gram approach also suits the characteristics of source code analysis. Our methodological extension includes a simplified profile and a less complicated but more effective similarity measure. Although Frantzeskou’s doctoral research includes numerous experiments which test SCAP under multiple forensically-significant conditions, in this article we present only two experiments. These experiments show that the SCAP method functions well on different programming languages, deals surprisingly well with cases where only a limited amount of very short programs per programmer is available for training, and performs well even in the absence of comments in the source code, a condition usually met in cyber-crime cases. The rest of this paper is organized as follows. Section 2 contains a brief review of relevant research in the area of authorship attribution, focusing on Keselj et. al.’s (2003) method. Section 3 describes our approach. Section 4 presents the results two experiments using SCAP. Finally, section 5 discusses the forensic application of SCAP and our research agenda for future work.

## **2.2 PATENT SEARCH**

### **2.2.1 Digital forensic acquisition kit and methods of use thereof:**

The patent describes disclosed are compositions, methods, and kits, for issuing and conducting automated imaging and preservation for obtaining digital forensic data from active (i.e., powered-on) and non-active (i.e., powered-off) computer systems. In certain embodiments, the invention further encompasses providing a customer base a preliminary report of data. In other embodiments, the invention encompasses the option to receive a virtual machine file set of the acquired information for additional viewing and examination by the customer. The invention further encompasses methods and systems for implementing the embodiments of the invention. The invention also encompasses methods, apparatuses, and systems for secure forensic investigation of a target machine.

### **2.2.2 Stylometry--definition and development. :**

The main aspect of authorship identification is the process of selecting features to be analyzed in a text. The most immediate idea that comes to mind at the mention of linguistic analysis of a text could be misleading--that an author's identity can be revealed through complicated or specific words since they mark the author's unique style and set him apart from others. Stylometrist have proven that it is the exact opposite which matters. Authors differ in their usage of the most frequently used simple words such as with and in, because one's subconscious uses words in daily parlance, such as these prepositions, automatically and without reflection. Words such as these are in fact an authorial "fingerprint" enabling experts to identify the author. Rare and specific words have a strong impact on readers but can easily be consciously inserted into a text to imitate a specific author. It is much more difficult to imitate the usage of simple prepositions and thus it is deemed a safer technique, according to Holmes (1998).

### **2.2.3 Method for data analysis and digital forensics and system using the same**

A system and method for data analysis and digital forensics is provided. The system for data analysis and digital forensics may include: an online data forensic server for collecting and analyzing usage history information from an object device, which is subject to data collection, downloading and collecting data on the Internet based on the usage history information, requesting issuance of a timestamp token for the collected data, and receiving the issued timestamp token; a timestamp token issuing server for issuing the timestamp token for the collected data in response to the request for issuance of the timestamp token and proving the issued timestamp token to the online data forensic server; and a storage device for storing the collected data.

## CHAPTER 3: REQUIREMENTS

### What are Functional Requirements?

A functional requirement defines a function of a system or its component. Functional requirements may be calculations, technical details, data manipulation and processing and other specific functionality that define what a system is supposed to accomplish.

### What are Non-Functional Requirements?

- A non-functional requirement (NFR) is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviors.
- Non-functional requirements define how a system is supposed to be.

### 3.1 FUNCTIONAL REQUIREMENTS

- A forensic UI should have ability to recognize supported every kind of text.
- A forensic UI shall have ability to notify user.
- A forensic tool shall have ability to extract live chats and dead chats(deleted).
- Frequent updating of encryption standards that whatsapp uses to protect these backups from unauthorized access.
- The system should be able to identify the original sender of the images and text through SCAP process.

### 3.2 NON-FUNCTIONAL REQUIREMENTS

- **Performance:** System must be lightweight and be able to determine the true sender.
- **Confidential:** The resources should be confidential and should be available to specific user.
- **Security:** System should provide a secure and confidentiality of data.

- **Availability:** System must work on all mobile devices and should be available all the time.

### 3.3 CONSTRAINTS

- The end-to-end encryption of the chats in the whatsapp makes it somewhat difficult to retrieve the data.
- There are many users of the system in a whatsapp group, so finding the suspect among the large number of users is found to be a little tedious.
- The repetition of similar kind of messages becomes arduous to analyze.
- As there is no limitation in sending or receiving the text size, data collection and analysis becomes a challenging task.
- Multi-linguistic data can also cause barriers in analyzing the data.
- Lack of Physical evidence makes crime harder to prosecute.

### 3.4 HARDWARE AND SOFTWARE REQUIREMENTS

Earlier most digital forensic investigations consisted of "live analysis", examining digital media directly using non-specialist tools. In the 1990s, several freeware and other proprietary tools (both hardware and software) were created to allow investigations to take place without modifying media. Objective of using the tool is to provide the digital evidence (if any) of the device acquired for investigation.

Advantages with the open source software tools

- Zero capital cost
- Minimal maintenance cost,
- Source code is freely available for access,

#### TOOLS USED

- PYTHON VERSION 3.6/2.0
  - NLTK Package
  - WX Package

### 3.5 ALGORITHMS:

The algorithms used in the system are Decision tree. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

#### **Naive Bayes algorithm :**

In machine learning, *naive Bayes classifiers* are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s.<sup>[1]:488</sup> and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines.<sup>[2]</sup> It also finds application in automatic medical diagnosis.<sup>[3]</sup>

#### **Source code author profile Identification:**

Source code author identification deals with the task of identifying the most likely author of a computer program, given a set of predefined author candidates. This is usually based on the analysis of other program samples of undisputed authorship by the same programmer. There are several cases where the application of such a method could be of a major benefit, such as authorship disputes, proof of authorship in court, tracing the source of code left in the system after a cyber attack, etc. We present a new approach, called the SCAP (Source Code Author Profiles) approach, based on byte-level n-gram profiles in order to represent a source code author's style. Experiments on data sets of different programming language (Python) and

varying difficulty (3-6 candidate authors) demonstrate the effectiveness of the proposed approach.

#### **ALGORITHM FOR SCAP :**

- Step 1. Divide the known source code i.e dataset into training and testing data.
- Step 2. Concatenate all the programs in the training set into one large file. Leave the testing data programs in their own files.
- Step 3. For each author training and testing file, get the author profile:
- Step 3.1. Extract the n-grams at the byte-level, including all non-printing Characters. All characters, even the non-printable, such as spaces, tabs, new line characters are included in the extraction of the n-grams.
- Step 3.2. Get the frequency for each n-gram type.
- Step 4. Get the Result.



# CHAPTER 4: PROPOSED DESIGN

## 4.1 SYSTEM DESIGN/CONCEPTUAL DESIGN

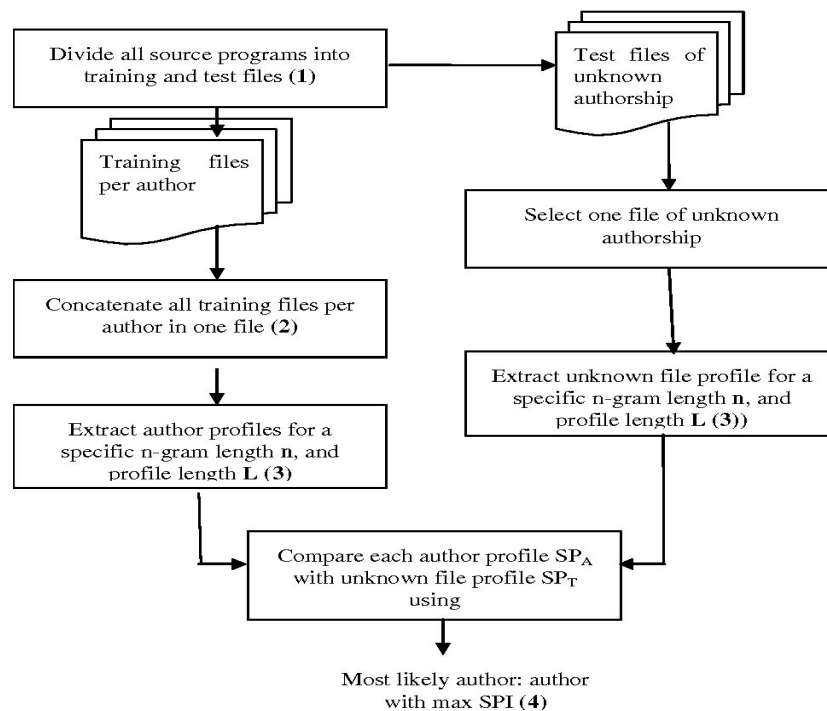


Figure-1: System Design

STEP-1: After collecting input ,divide it into two sets i that is in training and testing set per author.

STEP-2: Train all the data per author and store all training files together.

STEP-3: Extract features of all authors from training file.

STEP-4: Test data which is unknown of any author.

STEP-5: Extract features of author from testing file.

STEP-6: Compare step-3 and step-5 .

## 4.2 SYSTEM BLOCK DIAGRAM

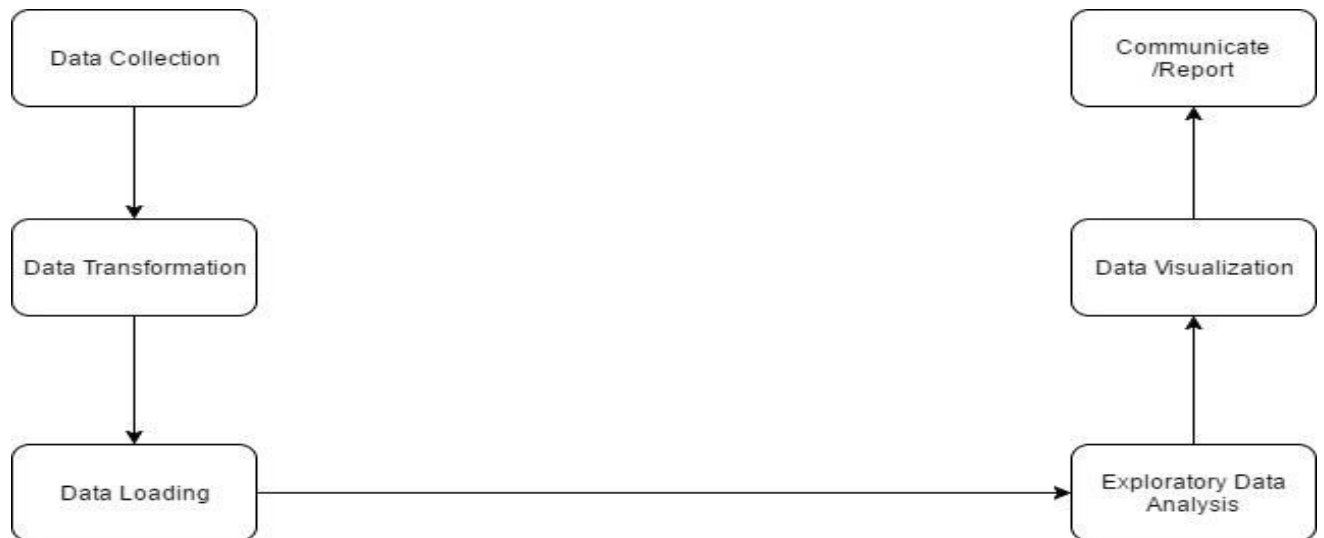


Figure-2:Block Diagram

**DATA COLLECTION:** Process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.

**DATA TRANSFORMATION:** Data transformation is the process of converting data or information from one format to another.

**DATA LOADING :**Data Load is the process that involves taking the transformed data and loading it where the users can access it.

**DATA ANALYSIS:** Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

**REPORT:** After analyzing your data and possibly conducting further research, it's finally time to interpret your results.

### 4.3 DETAILED DESIGN:

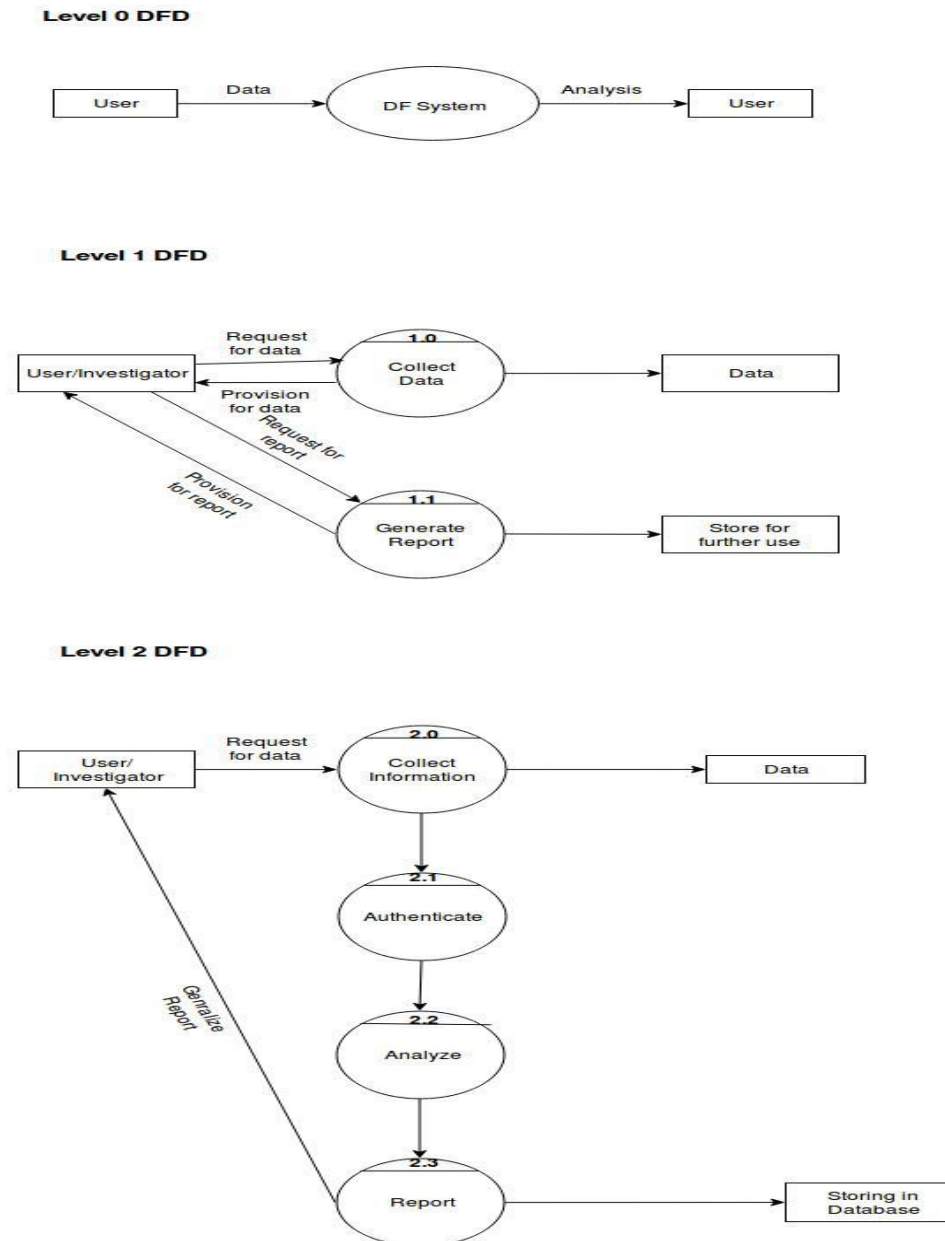


Figure-3: DFD(Level 0, Level 1, Level 2)

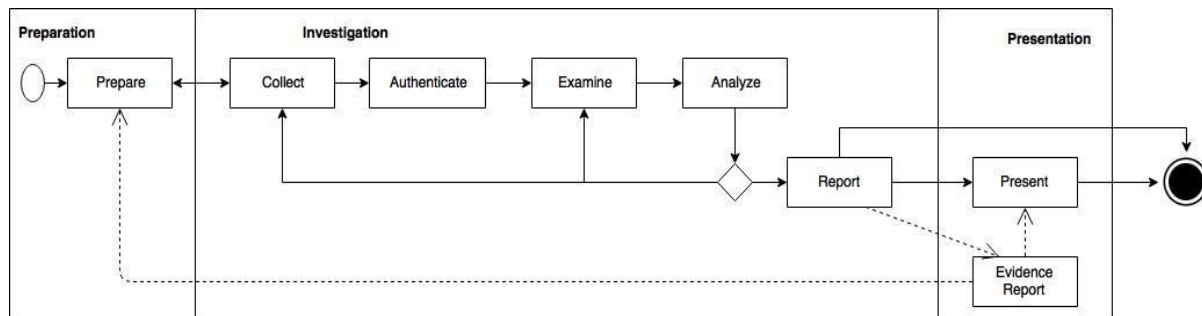


Figure -4: Activity diagram

- Activity diagram is a graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.
- In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes , as well as the data flows intersecting with the related activities.
- In our project,it basically focuses on preparation,collection,examination,analyzing,reporting and presenting activities.

## 4.4 PROJECT SCHEDULING AND TRACKING /GANTT CHART

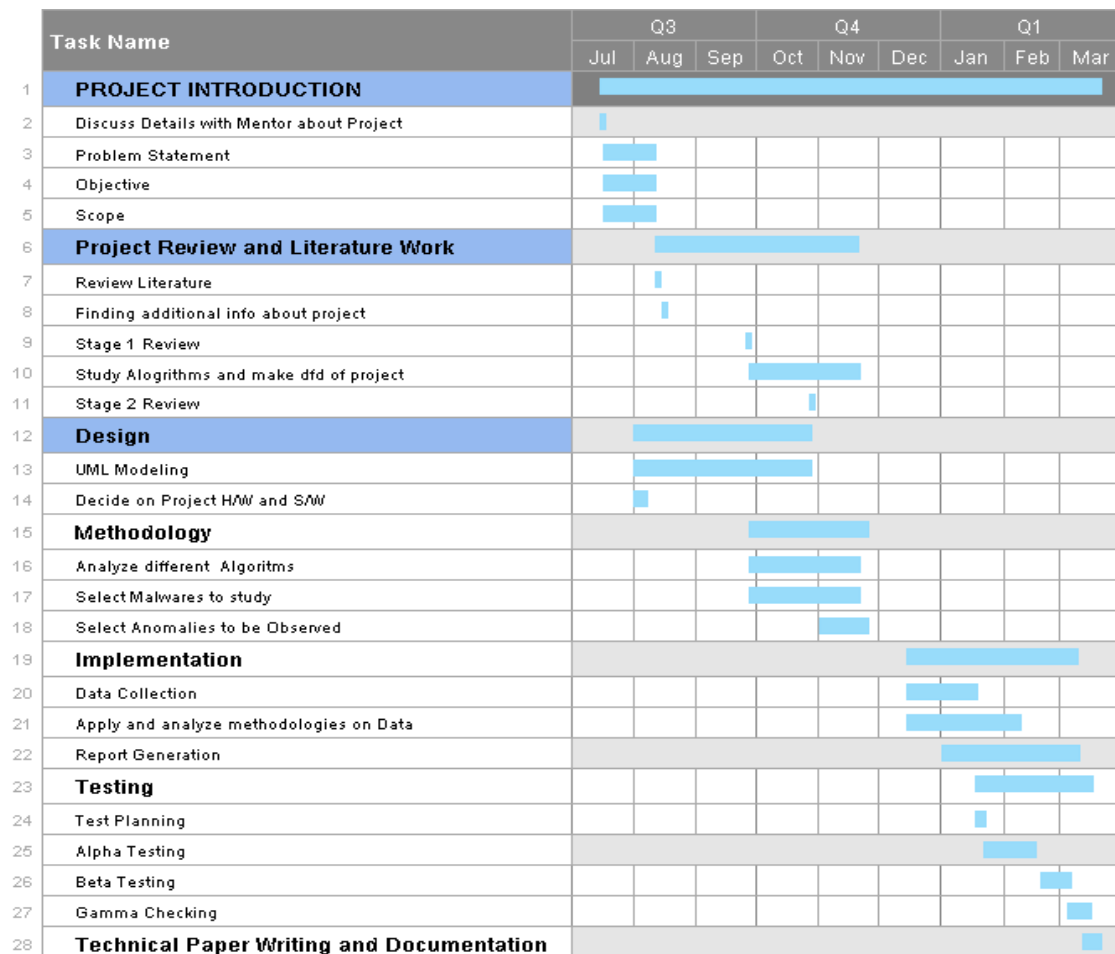


Figure-5:Gantt chart

## CHAPTER 5: Implementation

### Pseudo code for Naive Bayes

## The NLTK - Naïve Bayes Classifier

```
from nltk.classify import NaiveBayesClassifier
mr = movie_reviews
neg_examples = [(features(mr.words(i)), 'neg')
                 for i in neg_ids]
pos_examples = [(features(mr.words(i)), 'pos')
                 for i in pos_ids]
train_set = pos_examples + neg_examples
classifier = NaiveBayesClassifier.train(train_set)

# later on a previously unseed document
predicted_label = classifier.classify(new_doc_features)
```

### ALGORITHM FOR SCAP :

- Step 1. Divide the known source code i.e dataset into training and testing data.
- Step 2. Concatenate all the programs in the training set into one large file. Leave the testing data programs in their own files.
- Step 3. For each author training and testing file, get the author profile:
- Step 3.1. Extract the n-grams at the byte-level, including all non-printing Characters. All characters, even the non-printable, such as spaces, tabs, new line characters are included in the extraction of the n-grams.
- Step 3.2. Get the frequency for each n-gram type.
- Step 4. Get the Result.

## CHAPTER 6 : TESTING

Testing is the process of evaluating a system or its component(s) with the intent to find whether it satisfies the specified requirements or not. Testing is executing a system in order to identify any gaps, errors, or missing requirements in contrary to the actual requirements. Testing can be defined as - A process of analyzing a software item to detect the differences between existing and required conditions (that is defects/errors/bugs) and to evaluate the features of the software item.

### UNIT TESTING

**UNIT TESTING** is a level of software testing where individual units/ components of a software are tested. The purpose is to validate that each unit of the software performs as designed. A unit is the smallest testable part of any software. It usually has one or a few inputs and usually a single output. In procedural programming, a unit may be an individual program, function, procedure, etc. In object-oriented programming, the smallest unit is a method, which may belong to a base class, abstract class or derived/ child class. (Some treat a module of an application as a unit. This is to be discouraged as there will probably be many individual units within that module.)

### BENEFITS OF UNIT TESTING

- Unit testing increases confidence in changing/ maintaining code. If good unit tests are written and if they are run every time any code is changed, we will be able to promptly catch any defects introduced due to the change.
- Codes are more reusable. In order to make unit testing possible, codes need to be modular. This means that codes are easier to reuse.
- The cost of fixing a defect detected during unit testing is lesser in comparison to that of defects detected at higher levels.
- Debugging is easy.

## INTEGRATION TESTING

**INTEGRATION TESTING** is a level of software testing where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in Integration Testing. When two or more units are ready, they are assembled and Integration Testing is performed. For example, adding author name and feature extraction modules must be integrated in our system.

## PERFORMANCE TESTING

**PERFORMANCE TESTING** is the process of determining the speed or effectiveness of a computer, network, software program or device. This process can involve quantitative tests done in a lab, such as measuring the response time or the number of MIPS (millions of instructions per second) at which a system functions. Qualitative attributes such as reliability, scalability and interoperability may also be evaluated. Performance testing can verify that a system meets the specifications claimed by its manufacturer or vendor. The process can compare two or more devices or programs in terms of parameters such as speed, data transfer rate, bandwidth, throughput, efficiency or reliability.

## TYPES

- **Load Testing** is a type of performance testing conducted to evaluate the behavior of a system at increasing workload.
- **Stress Testing** a type of performance testing conducted to evaluate the behavior of a system at or beyond the limits of its anticipated workload.
- **Endurance Testing** is a type of performance testing conducted to evaluate the behavior of a system when a significant workload is given continuously.
- **Spike Testing** is a type of performance testing conducted to evaluate the behavior of a system when the load is suddenly and substantially increased.



## CHAPTER 7: RESULT ANALYSIS

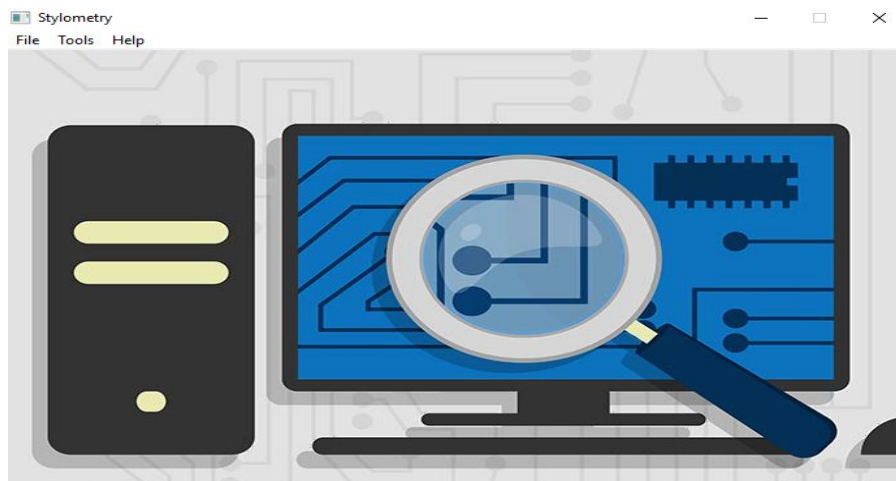
The following benchmarks were decided upon at the end of discussions and research.

1. Proper collection of the relevant data is to be collected from the entire social media sites.
2. Analysis of the collected data is to be done explicitly to ensure the collected data is related to crime.
3. Collection of digital evidence from cross geographically placed servers is important in case of transnational cyber and electronic crime.
4. Extraction of Computer crime related data from the social media to determine crime patterns.
5. Training the system and the source of the text is identified through different averages and SCAP process.
6. Author of the text that is sent on twitter is identified.

### Results and Screenshots

#### Stylometry:

##### 1.Start Window



## 2. Training Window

---

 TRAINING — □ ×

AUTHOR NAME

▼

+ AUTHOR

FILE NAME

▼

^

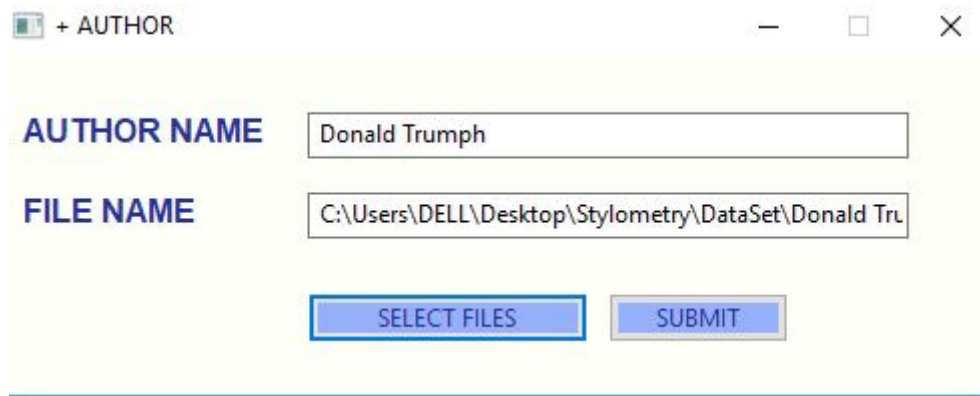
▼

EXTRACT FEATURES

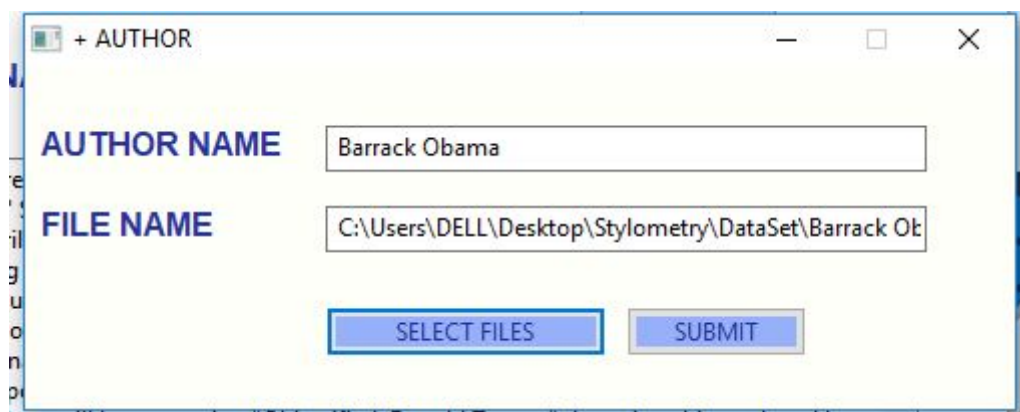
START TRAINING

---

### 3. Adding Author

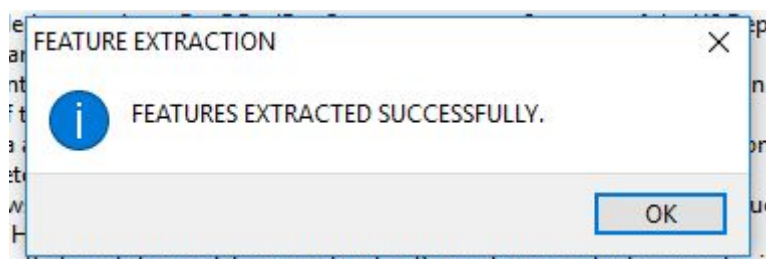


The screenshot shows a window titled "+ AUTHOR" with a yellow background. It contains two text input fields. The first field, labeled "AUTHOR NAME", contains the text "Donald Trump". The second field, labeled "FILE NAME", contains the path "C:\Users\DELL\Desktop\Stylometry\DataSet\Donald Tru". Below these fields are two buttons: "SELECT FILES" and "SUBMIT".

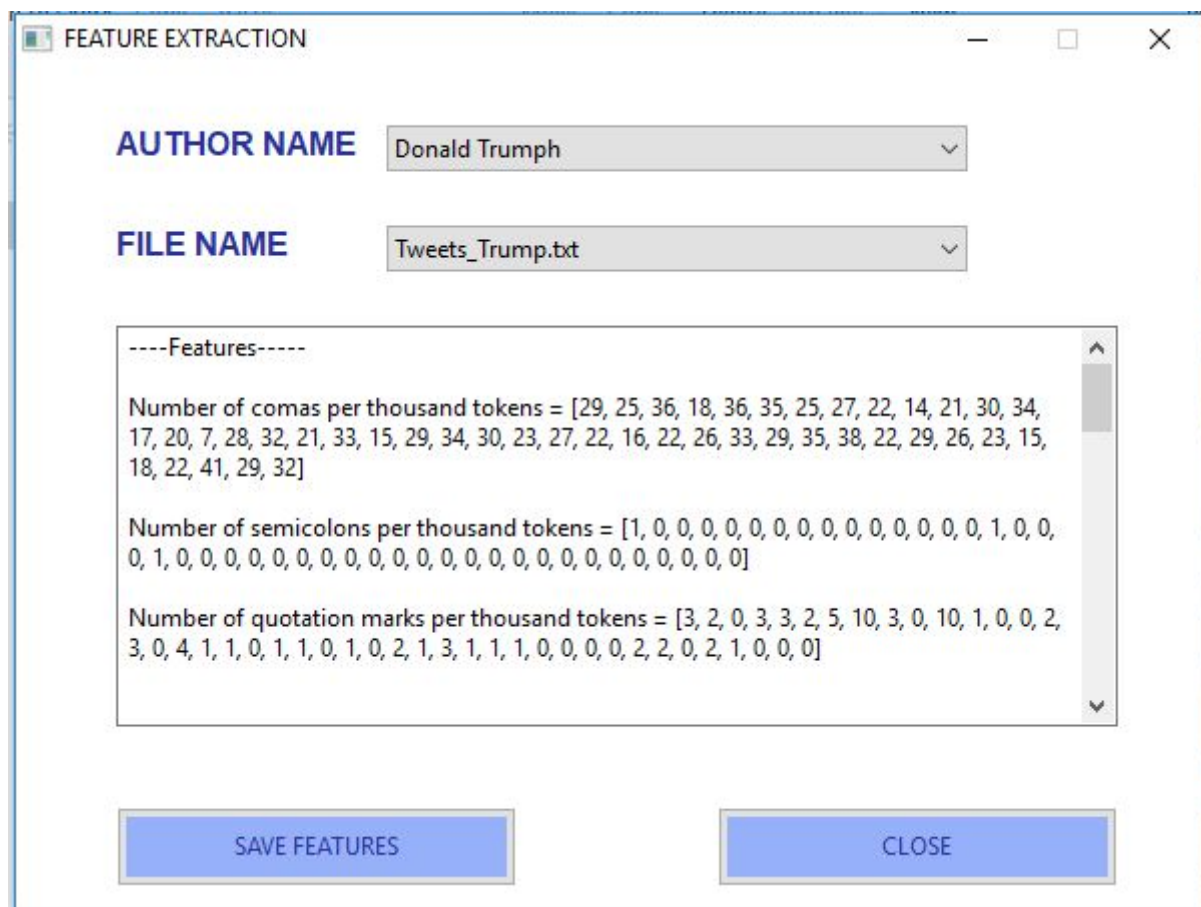


The screenshot shows the same "+ AUTHOR" window, but with different data. The "AUTHOR NAME" field now contains "Barrack Obama" and the "FILE NAME" field contains "C:\Users\DELL\Desktop\Stylometry\DataSet\Barrack Ob". The "SELECT FILES" and "SUBMIT" buttons remain at the bottom.

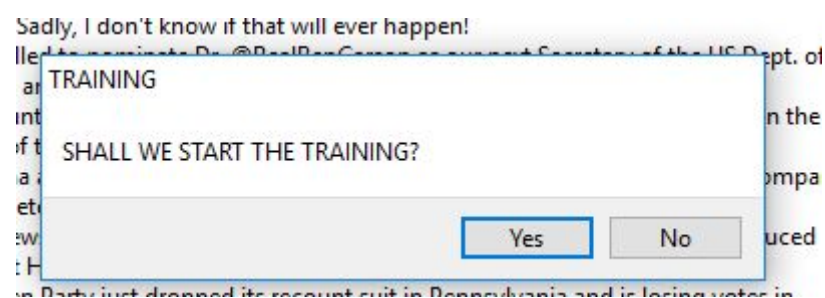
### 4. Extracting Features



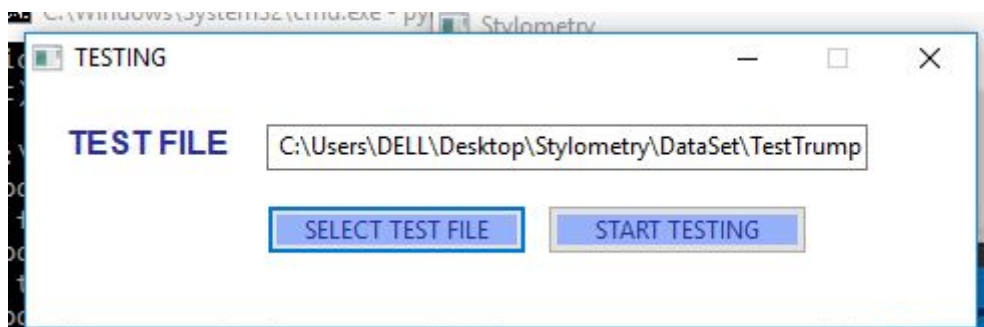
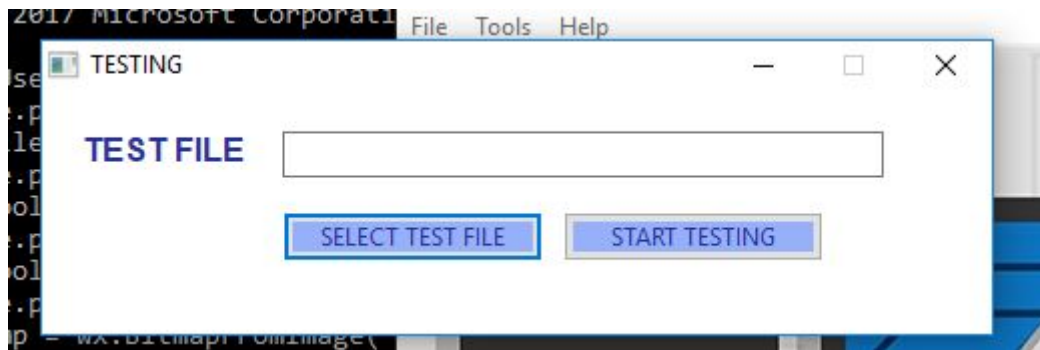
## 5. Saving Features



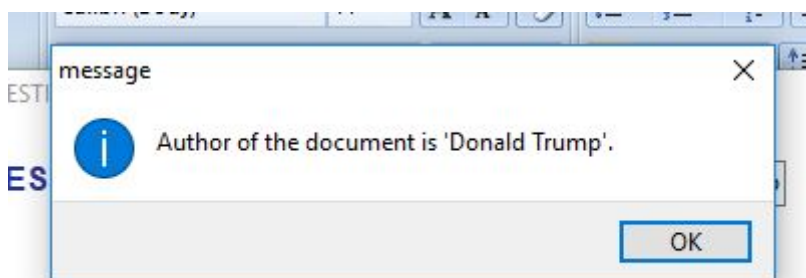
## 6. Start training



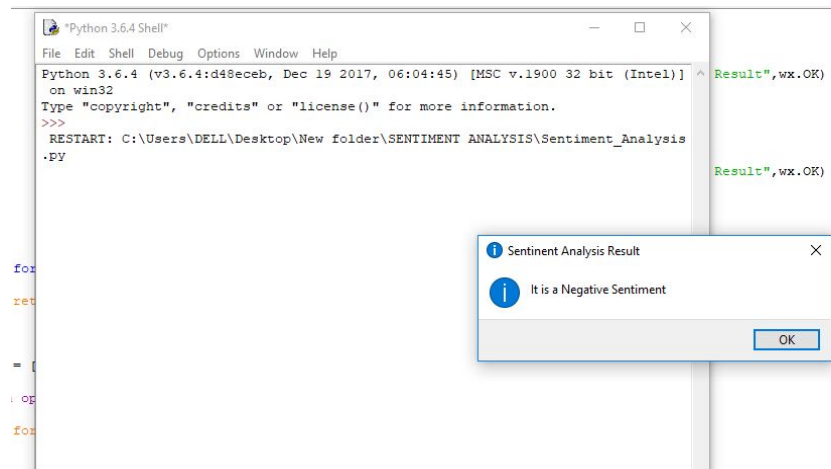
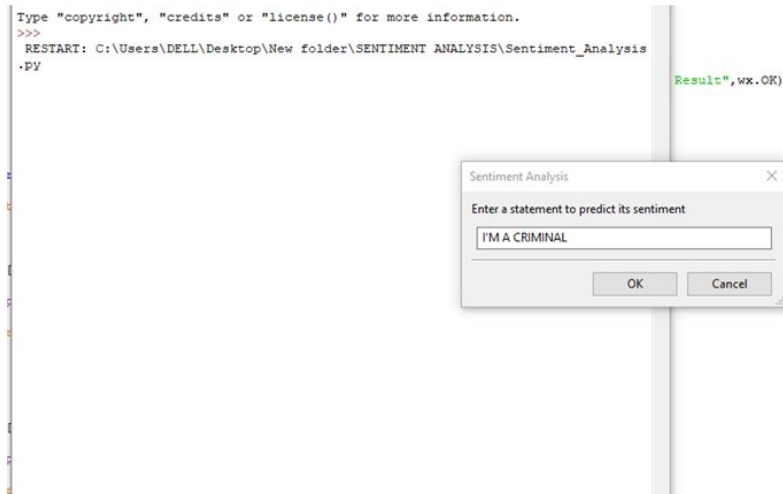
## 7. Testing Window



## 8. Testing complete(Author found)



## Sentiment Analysis :



## CHAPTER 8: CONCLUSION

We Indians are on the way to achieve the dream of Digital India!. Digital Crimes, on the other side of the coin, are increasing on the faster pace. Thus, Boosting the need of Digital Forensics. It has evolved a lot with time and hence the DF tools also. To make the process effective the discipline is subdivided into various branches where specialized tools are available for a particular branch. As every Forensic tool is associated with some or other limitation proper tool should be used as per the requirement of case in hand.

Twitter is a demandable micro blogging service which has been built to discover what is happening at any moment of time and anywhere in the world. In the survey, we found that social media related features can be used to predict sentiment in Twitter. We will use three machine learning algorithms. So, our proposed system concludes the sentiments of tweets which are extracted from twitter. The difficulty increases with the nuance and complexity of opinions expressed. Product reviews, etc are relatively easy. Books, movies, art, music are more difficult. We can also implement features like emoticons, neutralization, negation handling and capitalization/internationalization as they have recently become a huge part of the internet.

In this project, the SCAP approach to source code authorship analysis has been presented. It is based on byte-level n-gram profiles, a technique successfully applied to natural language author identification problems. This method was applied to data sets of varying difficulty demonstrating surprising effectiveness. The SCAP approach includes a new simplified profile and a less-complicated similarity measure that better suit the characteristics of the source code authorship analysis problem.

## CHAPTER 9 : REFERENCES

- [1] A quantitative approach to Triaging in Mobile Forensics by Fabio Marturana, Gianluigi Me, Rosamaria Bertè, Simone Tacconi, 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICSS-11/FCST- 11.
- [2] Text Message Corpus: Applying Natural Language Processing To Mobile Device Forensics Daniel R. O'Day And Ricardo A. Calix, Multimedia And Expo Workshops (ICMEW), 2013 IEEE International Conference
- [3] Cyber Forensics by LabSystem (I) pvt. Ltd.
- [4] Mobile Device Forensics: Extracting and Analysing Data from an Android-based Smartphone Normaziah A. Aziz, Fakhrulrazi Mokti, Mohd Nadhar M. Nozri, 2015 Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic
- [5] Smartphone Forensics Analysis: A Case Study, Mubarak Al-Hadadi and Ali AlShidhani, International Journal of Computer and Electrical Engineering, Vol. 5, No. 6, December 2013.
- [6] [https://en.wikipedia.org/wiki/Digital\\_forensics](https://en.wikipedia.org/wiki/Digital_forensics)
- [7] <https://en.wikipedia.org/wiki/forensics>.
- [8] <https://en.wikipedia.org/wiki/MOBILedit>
- [9] [https://en.wikipedia.org/wiki/List\\_of\\_digital\\_forensics\\_tools](https://en.wikipedia.org/wiki/List_of_digital_forensics_tools)



## CHAPTER 10 :Appendix

### 10.1.List of figures

Figure Number	Heading	Page Number
1	System design	25
2	Block diagram	26
3	DFD	27
4	Activity	28
5	Gantt chart	29

### 10.2.List of Tables

Table Number	Heading	Page Number
1	Social Media	15