Sentiment Prediction from Online Examination Scripts using Computer Aided Graphology

Thesis by

Isita Mitra

Class Roll No.: 001910504019

Examination Roll No.: M6TCT22020

Registration No.: 149853 OF 2019-2022

Under the guidance and supervision of

Dr. Chitrita Chaudhuri

Associate Professor

Department of Computer Science and Engineering Faculty of Engineering and Technology

In Partial Fulfillment of the Requirements for the Degree of

Master of Technology

JADAVPUR UNIVERSITY

Kolkata, West Bengal, India

2022

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

To Whom It May Concern

I hereby forward the thesis entitled **"Sentiment Prediction from Online Examination Scripts using Computer Aided Graphology"** prepared by **ISITA MITRA** (University Registration No: 149853 OF 2019-2022, Examination Roll No: M6TCT22020) under my guidance and supervision. It is a bona-fide piece of work that may be accepted in partial fulfillment of the requirement for awarding the degree of **Master of Computer Technology** in the Faculty of Engineering and Technology, Jadavpur University, Kolkata.

.....

Dr. Chitrita Chaudhuri
(Thesis Supervisor)
Associate Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Countersigned

.....

Prof. Anupam Sinha
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-32.

.....

Prof. Chandan Mazumdar
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata – 32.

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval*

This is to certify that the thesis entitled **"Sentiment Prediction from Online Examination Scripts using Computer Aided Graphology"** prepared by ISITA MITRA is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it has been submitted.

……..
Signature of Examiner 1:
Date:

Final Examination for evaluation of the thesis.

……..
Signature of Examiner 2:
Date:

*Only in case the thesis is approved

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by undersigned candidate, as part of her Master of Computer Technology studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

NAME	: ISITA MITRA
ROLL NUMBER	: 001910504019
REGISTRATION NUMBER	: 149853 OF 2019-2022
SIGNATURE WITH DATE	**:**

ACKNOWLEDGEMENTS

I am pleased to express my gratitude and regard to my Project Guide Dr. Chitrita Chaudhuri of CSE Department, J.U. for her invaluable guidance, constant encouragement and inspiration during the period of my project.

I would also like to thank my family for their constant support and motivation during the tenure of this course.

Any omission of acknowledgment does not reflect my lack of regard or appreciation.

DATE:

…….

ISITA MITRA

EXAMINATION ROLL NUMBER: M6TCT22020

REGISTRATION NUMBER: 149853 OF 2019-22

Department of Computer Science and Engineering,
Jadavpur University

Table of Contents

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Number *Page*

Chapter1

Introduction

Humans have always been inquisitive about correctly deciphering the uniqueness and variance in the behavioral pattern of persons surrounding them [1]. Handwriting Analysis is a scientific approach which helps them detect, assess and comprehend the personality and mentality of an acquaintance by studying their writing hand. Pressure, strokes and individualistic trends followed at the time of writing are supposed to reveal the mentality of the author. Fear, emotional crisis, strategies of defenses, and honesty are some of the genuine characteristics which can be identified from a handwritten document. A counselor who wants to judge a person will be aided greatly by handwriting analysis. This field of investigating personality traits using handwriting is known as 'Graphology'. Recent progresses in the field also help in diagnosing diseases. Computer-Aided-Graphology (CAG) contributes to this effort by extracting precise features from handwritten text images more rapidly. Gauging the condition and reaction of the human mind at the moment of stress can also be achieved with greater accuracy through CAG.

The present work attempts to assess the degree of stress and confidence issues faced by a student during an examination, by analyzing his/her handwriting from answer-script images. Such a system could in turn lead to design of models which can be used to enhance the confidence level of the students in their future learning and knowledge gathering endeavors. It can also be used by the examiners and teachers to grade the students' current progresses more accurately on the basis of their overall grasp on the subject under scrutiny.

1.1. Motivation

The motivation for probing the topic arises from the fact that computer aided graphology is a comparatively cheap, effort-less and non-evasive

mode of personality assessment. While a certain amount of pressure may be beneficial, too much of it can cause nervousness or anxiety, eventually leading to poor performance and overall demotivation.

The necessity of studying stress-related panics is grounded in the need for better handling of the growing sense of alienation faced by the student community in general during these pandemic situations. While expert graphologists can definitely be the best persons to decipher abnormalities, there are some obvious disadvantages involved with the manual process. First and foremost, amongst these is the issue of availability- competent persons are hard to find in times of need. Secondly, the cost involved to hire them may prove to be exorbitant. Training the regular teachers to decipher graphology may also not be a feasible solution, in terms of time, expense, or interest. Moreover, since the task is tedious and time-consuming, it is also bound to be error- prone. Finally, the point of human bias is always a threat in all types of manual system – graphology is no exception. On the other hand, CAG can solve all these problems of the manual system by analyzing large volumes of data accurately in much less time.

1.2. Objective

The study in this work focuses on understanding the mental state of the students in a stressful environment with the help of Computer-Aided Graphology. A dataset comprising of answer scripts from 12 University students is collected for the purpose of this study. The scripts are analyzed and four features, namely, Baseline Orientation, Line Spacing, Word Spacing and Margin, are extracted. To evaluate the mental state based on the features, two experiments are formulated. In the first experiment, all the data features are evaluated using a clustering algorithm to achieve an overall accuracy of 90.47%. An approach to remove the role of the human annotator is also proposed as the second experiment. K-Means clustering algorithm is used for extracting each feature value range from 7 randomly selected data points. The ranges are converted to feature value rules leading to prospective class categories. The remaining 5 data are used as test tuples for which the class categories can automatically be generated

using those rules. The accuracy achieved in the process is found to be 67.61% only in the present case. Such a rule-based classifier approach can be made utilizing this technique on a larger dataset to presumably achieve a better accuracy.

1.3. Thesis Layout

The thesis is organized as follows: The second chapter of this thesis deals with reviewing state-of-the-art literatures on the topic and finding out the research gaps. The third chapter studies the theoretical works on which the foundation of the tools used in the experiments are grounded. The fourth chapter deals with the actual description of the methodologies adopted in the work. This chapter also provides details of the datasets and software utilized in the process. The pre-final chapter presents the results achieved through the work in a comprehensive manner. The last chapter helps to draw the overall conclusions based on the experimental results and analysis made during the research work. Some areas of interest which may be explored in the future are also mentioned here. The dissertation finishes with a list of the researches cited in the work.

Chapter 2

Literature Review

Literature review can be defined as a survey of scholarly resources on a certain topic, e.g. here, graphology. This provides a complete overview of current work, and also provides knowledge to identify relevant methods, theories and also gaps or limitations in all this existing research. The following section provides a de- tailed analysis of the different approaches performed in the past for the purpose of automated handwriting analysis.

2.1. Past Work

The task of identifying the personality of a writer based on Handwriting had been tackled in the paper by Ghosh et al. [2]. The objective of the proposed method is to identify personal behaviors, positive and negative social behaviors of an individual. In this method several structural features such as stroke, slants, zones, aspect ratio, loops, contour shapes are extracted from handwriting samples of 5300 people of different genders and age group. This system accepts characters from a to z and reveals behaviors. Graphological rules for this system are generated on the basis of several discussions with the graphologists of Graphological Institute in Kolkata. The quantitative evaluation of this method achieves 86.70% accuracy. The framework by the authors is shown in Fig.2.1.

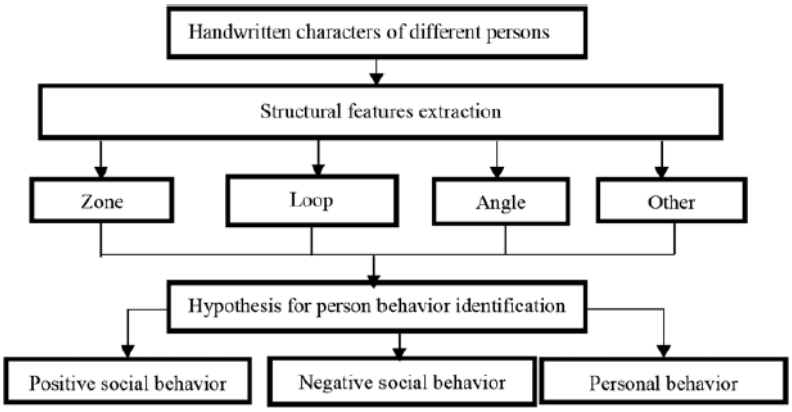


Figure 2.1: System Architecture for [2]

For behavior identification, this proposed system accepts individual characters (A-Z) as an input. So, in case of more variations or touching characters the system predicted result is poor as character segmentation is needed for producing more accurate results.

In the paper by Anand et al. [3], conscious and subconscious factors are included for deciding individuals' attitude and aptitude. The proposed system works in two different modules namely, MCQ Analysis and Handwriting Analysis. The MCQ module consists of aptitude questions and psychometric questions. The extracted features from the handwriting samples are word space, left and right margin, pen pressure, Slant of all letters, Zone and Size. The aptitude test gives the top five appropriate broad career domains. The psychometric test gives an idea about conscious personality traits and a few specific careers. The handwriting analysis gives specific careers and both the results are mapped together for career selection. A system diagram of the model is shown in Fig. 2.2.

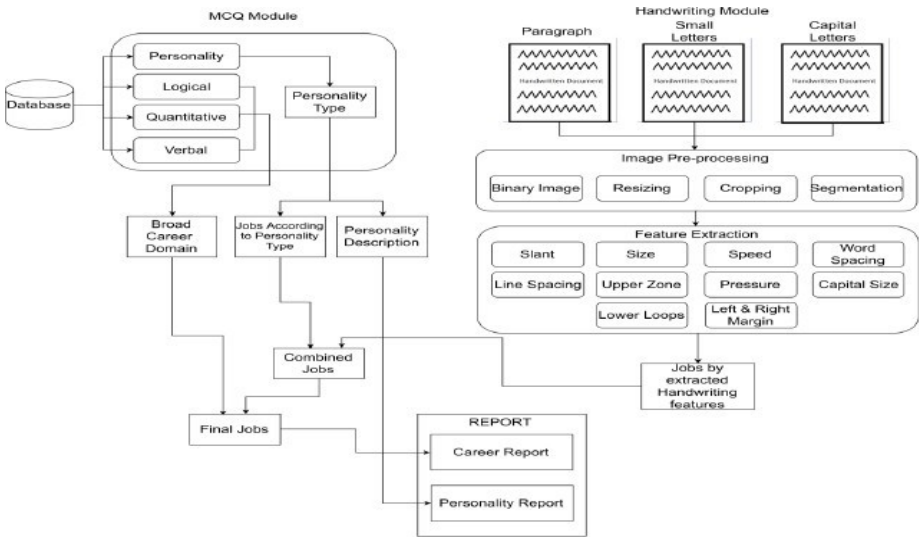
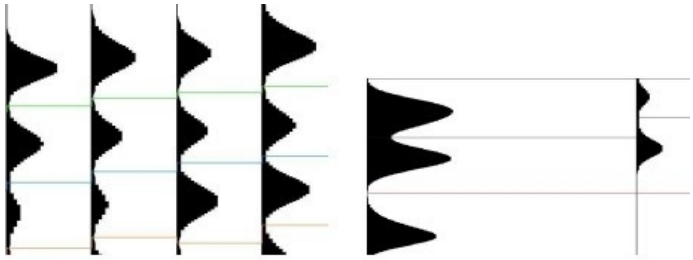


Figure 2.2: System Architecture for [3]

This is tested for 15 samples and out of 15, 13 samples are appropriate in terms of final jobs and score. It provides 93.33% precision. The system can be used for recruiting in industry so that the personality and aptitude certificate make it easy to decide if he/she is suit able for their work culture or not.

A new approach for the line segmentation of handwritten image is presented in the paper, named A Statistical approach to line segmentation in handwritten documents by Manivannan Arivazhagan et al [4]. This proposed method is performed in 3 stages, initially, all the lines are drawn from left to right parallelly and also modelled using bi-variate Gaussian densities. Smoothing is done with a simple average filer of window length 5 for removing spurious valleys and peaks in the projection profile. Let consider the chunks of smoothed projection profiles as $\psi 1, \psi 2 \ldots \psi N$ from left to right where N = 20 is the total number of chunks. Connect a valley to the closest valley. If 2 or more valleys are connected to the same valley then retain the closest pair and reject the rest as shown in Fig. 2.3.

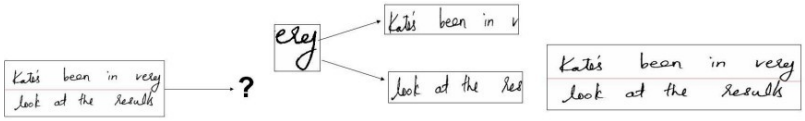


(a) ψ_{i+1} connected to ψ_i

(b) Situation where ψ_i not connected to any in ψ_{i+1}

Figure 2.3: Column Wise Connectivity

Then, probabilistic decision is to make whether it belongs to the above line or below line, in case of obstructing handwritten component. The lines are guided by the piece-wise projection profile if available. A demonstration of the model is shown in Fig. 2.4.



(a) A line drawn is obstructed by the descender of the letter 'y'.

(b) A decision need to be made whether to associate this connected component to the above or the below line.

(c) A line traversing below after deciding that the component belongs to the line above

Figure 2.4: Model Demonstration

Once the baselines of handwritten text have been detected properly, then it can be subjected to word segmentation, and other indexing steps necessary for the feature extraction.

The paper "Graphology for Farsi Handwriting Using Image Processing Techniques" by Somayeh Hashemi et. al. [5] states human behaviors related to graphology rules.

At the time of writing, the muscles of fingers, hands and arm are in control of the mind of the writer. So, the words written by him bear a direct relationship to the mind which guides their formation. The unconsciously directed movement by brain help to predict the mental state of the writer.

Farsi is written from the right to the left. So, the right margin can be aligned or irregular, convex, concave and progressive or regressive. The left margin can also be aligned or irregular. The line spacing and word spacing can be narrow, normal or wide. To calculate an index for word spacing, the image is divided into 14 equal horizontal bands and for line spacing calculation the resolution is reduced to 20 dpi. Convex hull is drawn for line spacing and the total black pixels is divided by the product of the convex hull area along with the pen width. The obtained value is considered as an index of line spacing. The median distance between the neighboring bounding boxes of connected components is calculated for word spacing.

The most important features of graphological analysis include the, line spacing, word spacing, shape of the page margins, size of letters, writing speed, text density, and regularity of writing. This paper represents several automated feature extraction methods. This rule sets are generated and verified on 150 test samples and 30 training sets.

The Paper, named as, “Personality Features Identification from Handwriting Using Convolutional Neural Networks” by Sri Hastuti Fatimah [6] proposes a method by using CAG (Computer Aided Graphology) system for analyzing the handwriting image and suggest personality traits based on the extracted feature.

This multi-structure analysis is done considering 6 features, such as, margin, line spacing, word spacing, dominant zone, slope and 4 specific alphabets (‘a’, ‘g’, ‘s’, ‘t’). CNN classification approach is used with 98.03% accuracy but structural analysis gives 82.5% to 100% accuracy. A demonstration of the model is shown in Fig. 2.5.

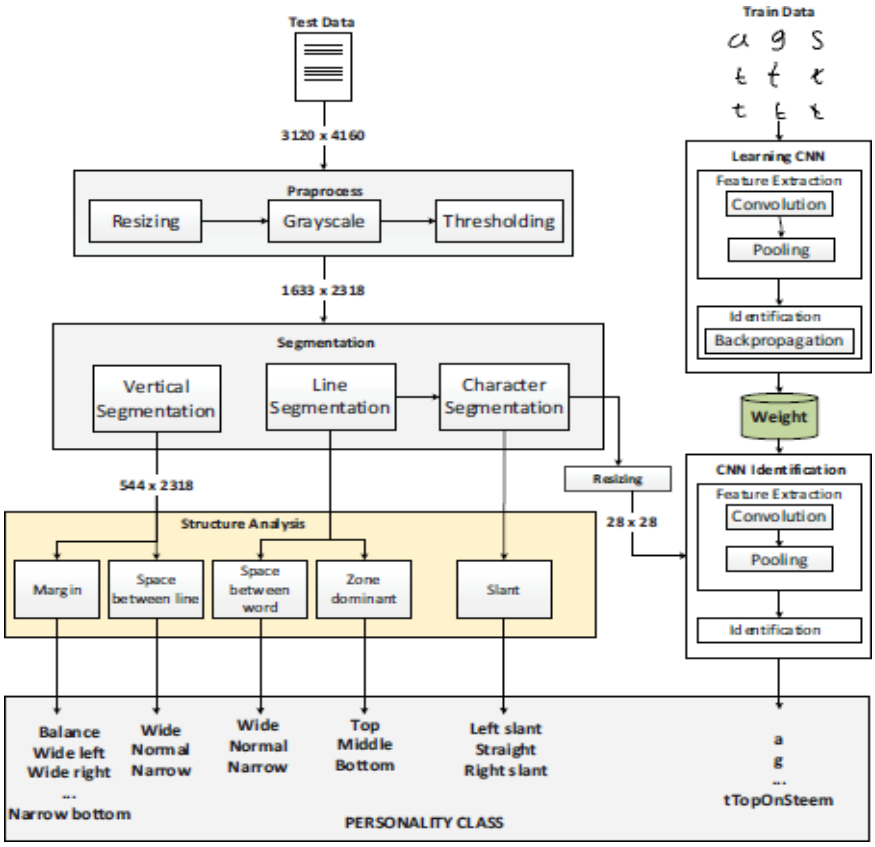


Figure 2.5: Model Demonstration

2.2. Gap Analysis

In the past research works, no Rules for wide, narrow left/right margin and proper distinct mention of word spacing and line spacing characteristics has been performed as it is mentioned as spacing as a whole. We have tried to create our own rule set on the basis of exam stress of student and taken help from several articles from Google [7].

Mostly to calculate line spacing the average value is considered. Average value of line/word space is calculated by adding all white space between lines/words and dividing by total number of rows or words [6] [3]. To identify the baseline, we have applied the method mentioned by Manivannan Arivazhagan [4] but for calculating line spacing we have taken a full exam paper of student and divided it in 3 sections and then applied clustering techniques for more accurate value. Similarly, for word spacing in most of the paper average distance between the bounding rectangles are taken [5] but here we have calculated the minimum inter-contour distances and followed the same approach as line spacing.

Chapter 3

Background Study

The human interest in handwriting analysis started approx. 400 years back. The very first graphological essay was written in 1622 by Camillo Baldi, known as father of graphology. Graphology can be stated as the study of graphic movements includ ing drawings, doodles etc. but mainly it is focused on handwriting for predicting the mental and physical states of the writer. Writing is a spontaneous action of communication and thus, it reveals several psychological and physiological facts about the writer. As per a paper by Alport, named, "A Psychological Interpretation", personality assessment is defined on the basis of an individual's view, their interaction with others, their expression of emotions, problem solving ways, how they handle stress and adapts to the new challenges of life. These are necessary piece of information wherein the health professionals assess, comprehend and treat their patients [1]. Handwriting is very unique because the writing rhythm and the pen stroke cannot be duplicated. To define graphology, it can be said that it is a projective technique which analyze the handwriting patterns and physical charac- teristics of the handwriting sample for evaluating the personality traits of the writer based on several graphic aspects (e.g., variations of shape, space, and movement). In one word, graphology is the "psychology of writing".

3.1. Computer Aided Graphology

Graphologist extract features from handwriting samples manually such as size, baseline, stroke etc and analyze those on the basis of graphology rules. Although the validity of these graphology rules is in question but this manual process is hectic and error prone. CAG or Computer Aided Graphology system helps graphologist to extract features from samples and analyze automatically. It produces results more accurately in a lesser time. CAG system takes the handwriting sample as input and gives personality traits as output. There are several modules in the CAG system [7], including:

- **Scanning:** The input handwriting sample is scanned in gray-scale resolution.
- **Pre-processing:** Binarization, thresholding and noise removal is done in this step.
- **Feature Extraction:** Based on provided information the features from the writing sample are extracted.
- **Feature Analyzer:** Then the feature mapping with the rule set is performed in this step.

3.1.1. Benefits of Graphology

Graphology can be used as a descriptive and diagnostic tool in personality assessment. Graphology aids in picturing client and formulating guidelines for psychotherapy [7]. The following advantages of graphology are emphasizing the viability of graphology:

- Handwriting sample is easily obtainable from a literate person.
- Except blank paper and pen, no expensive test material is needed.
- This analysis is based on the handwriting specimen only, no details of the writer is needed.
- Handwriting cannot become outdated or obsolete.
- It provides valuable information about personality, interaction style, thinking style as well as subconscious drives.
- Handwriting analysis gives us a clear understanding of priorities and helps in making the right decision.

3.2. Feature Extraction

For the purpose of Graphology, multiple features have been identified in the past which provide pointers to the mental state understanding of the author [3]. In the following section, a detailed study on most prominent features used in the field of Graphology is highlighted.

3.2.1. Baseline

Baseline can be stated as an imaginary or pre-printed line on which the letters reside. If no pre-printed line is there on the paper, then writer considers own baseline as per their writing style.

Table 3.1: Baseline related Personality traits

SL. No.	Type of Baseline	Personality Traits	
		Positive	**Negative**
1	Horizontal	Realistic and disciplined in nature and also straight forward, responsible and firm.	Negative: inflexibility.
2	Ascending	Optimistic, active and energetic, mostly cheerful.	Choleric behaviour, euphoria and aggression.
3	Descending	In general, these writers are negative.	Depressed with constant disap-pointment and defeat. Mental tiredness or physical weakness, pessimism which may lead to digestive trouble.

3.2.2. Spacing

Our next important feature is spacing, For handwriting analysis in graphology spacing indicates the distance between words, letters, and even lines of the writing.

Table 3.2: Line Spacing related personality traits

SL. No.	Spaces Between Lines	Personality Traits	
		Positive	Negative
1	Lines widely apart	far-sighted and self-assured	may be a little hostile.
2	Medium	balanced.	Nothing to mention
3	Lines crowding together	Nothing to mention	usually afraid of being isolated; lonely or having fear of being abandoned

Table 3.3: Word Spacing related personality traits

SL. No.	Spaces Between Words	Personality Traits	
		Positive	Negative
1	Too far apart	Nothing to mention	It implies paranoid and suspicious mentality.

Table 3.3: Word Spacing related personality traits (contd.)

SL. No.	Spaces Between Words	Personality Traits	
		Positive	Negative
2	Adequate spacing	balanced attitude towards life and socially well adjusted.	Nothing to mention
3	Inadequate spacing	Nothing to mention	cannot maintain distance and invades personal space.

3.2.3. Margins

In graphology, margins of a writing can be identified, such as on the top or to the left or the right, or at bottom. In a blank paper where the writer start writing is significant as it reveals the social standing of the individual.

Table 3.4: Margin related personality traits

SL. No.	Type of Margin	Personality Traits	
		Positive	Negative
1	Wide upper margin	formal and respectful.	Nothing to mention
2	Narrow Upper Margin	indicates familiarity.	informality
3	Wide lower margin	Nothing to mention	indicates fear of facing the fu- ture.
4	Narrow lower Margin	Nothing to mention	tendency to delay the evitable.
5	Wide right margin	Nothing to mention	indicates indecision regarding future plans
6	Narrow right Margin	indicates decisive nature in general.	Nothing to mention
7	Wide left margin	indicates detachment from past issues and moving on to new things con- fidently.	Nothing to mention
8	Narrow left Margin	Nothing to mention	Overtly preoccupied with the past which also affects future decision making.

3.3. Supervised Learning

Supervised learning is a method which provides input data along with correct output data to the ML model. The primary aim of the supervised learning algorithm is finding an appropriate mapping function for mapping the input variable with the output variable. Supervised learning can be used in the real world for Image classi- fication, Risk Assessment, spam filtering, Fraud Detection etc.

Rule-based classifier is used here which extracts each feature value range in the form of rules from the train data set and makes the class decision on the test data set depending on the usage of the "if..else" clauses. The class values generated for the test data are compared with pre-existing class values already provided with the test data, and the matched cases are statistically counted to provide accuracy percentages for such a classifier. The advantage of such a rule-based classifier is that the model is built simply by implementing the if-else constructs readily available in all programming languages.

3.4. Unsupervised Learning

Supervised learning cannot be applied directly to a problem if no corresponding class value is provided with the input data. In that case, the only solution is to partition the data using unsupervised learning techniques. The aim of unsupervised learning is to find the underlying structure of a dataset and group those data according to the similarities, and also represent the dataset in an organized format.

Clustering is an unsupervised learning algorithm which is used to solve the clustering problems in data science or machine learning.

3.5. Clustering and Distance Measurement Techniques

Euclidean Distance is an important metric which is used for assessing the similarity or dissimilarity of data to deduce if they are inter-related to one another for comparison purposes.

Euclidean distance finds an extensive usage in data clustering literature by using the equation given below.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

3.5.1. Clustering Algorithm

One such clustering algorithm that utilizes the Euclidean distance metric is the K-Means Clustering algorithm, which is a form of Unsupervised Learning capable of grouping unlabeled data into separate clusters. The variable K is used to symbolize the number of clusters to be formed. For example, for a model with K=3, three clusters are formed for grouping the data. The demonstration of the algorithm is as highlighted next in Fig. 3.5.

Step 1: Let's decide k for identifying the dataset, here, K=2 to put them into 2 different clusters (Fig. 3.5a).

Step 2: Select some random centroid or k points to form the cluster. These points can be any other point as in it can be a part of the input or outside the dataset. Here the selected centroids are not the part of the dataset (Fig. 3.5b).

Step 3: Now assign each data point to the closest centroid. A median between the two centroids id drawn (Fig. 3.5c).

From the above image, it is clear which points are closest to which centroid. Let color them as yellow and blue for clear visualization (Fig. 3.5d).

Step 4: Repeat the process to find the closest cluster by choosing a new centroid (Fig. 3.5e).

Step 5: Reassign each datapoint to the new centroid. Repeat the process to find a median line (Fig. 3.5f).

The above-mentioned image states that 1 yellow point is on the left side of the line, and 2 blue points are right to the line. So, these 3 points need to be assigned to the new centroids (Fig. 3.5g).

Step 6: Go to the Step-4 for re-assignment for finding new centroid. We will repeat the process until the new centroid is found (Fig. 3.5h).

Draw the median line as per the new centroid and reassign the data points (Fig. 3.5i).

Model is formed as no dissimilar data points on either side of the line is noticed (Fig. 3.5j).

Step 7: Model with 2 final clusters is shown in (Fig. 3.5k).

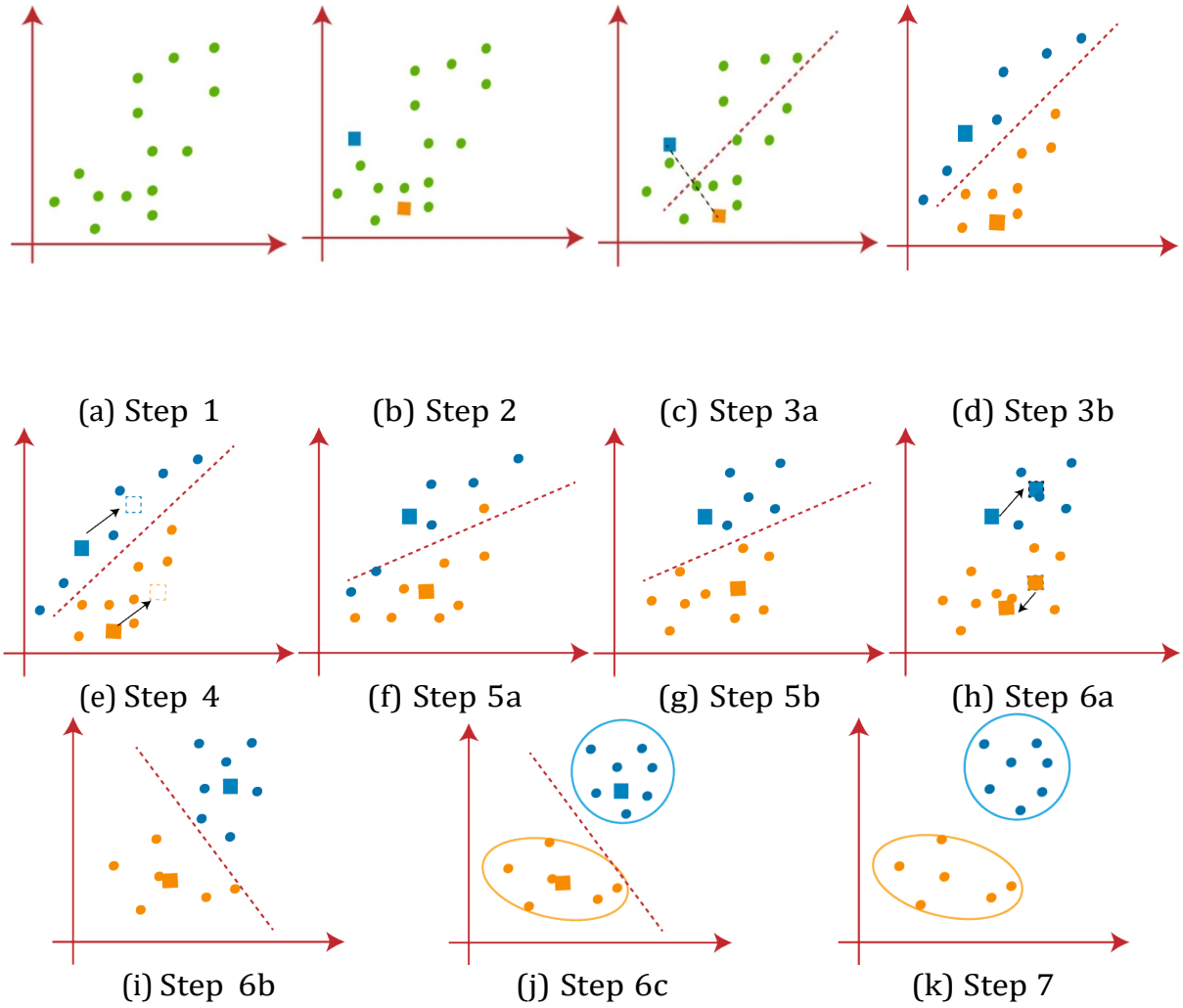


Figure 3.5: Demonstration of Clustering Algorithm

3.6. Python Programming Language

In image processing MATLAB is particularly good and it is faster than Python. But in our work, we have chosen Python as is a high-level language and MATLAB is a low-level language. Python is user friendly, more portable and more readable. Python has its own libraries as our focus was on machine learning. The most popular machine learning frameworks, for example, Scikit-learn which is used for k-means clustering are written in Python and NumPy library function is used for mathematical calculations, which is very powerful.

Chapter 4

Methodology

This section explains the entire process and analysis based on logic and psychology. A methodology is written to provide detailed information on how the entire research is conducted.

4.1. Proposed Work

An online examination answer script with multiple pages is taken which consists of scanned handwritten answers by the student as an input. Then image pre-processing is done which includes binarization and inversion of the image and noise reduction techniques. The script is divided into 3 sections i.e. start, middle and end, to analyze the fluctuations of confidence and stress level of the student during the examination. Additionally, each page of each section is divided into 3 parts, start, middle and end of the page, to improve accuracy. Then the image line segmentation is done and it is followed by important feature extraction. Here the baseline, line spacing, word spacing, top, bottom, left and right margins are considered. A rule set is created on the basis of the different graphology features indicated in Tables 3.1, 3.2, 3.3 and 3.4. K-means clustering method is next applied to evaluate the mode of writing for each feature of a student - first for parts within each page of a section, and then collectively for each section and the document as a whole.

A **ground truth** is prepared for this study by individually assessing each data manually and annotating the feature values as ascending, straight and descending for baseline and narrow, normal and wide for the rest. The quality of annotation of the ground truth, however, requires improvement as it was done by the researcher, expert graphologists being unavailable due to the onset of the pandemic. The obtained output produces the overall mentality assessment of the student during examination. The flowchart highlighting the working of the proposed algorithm is shown in Fig 4.1.

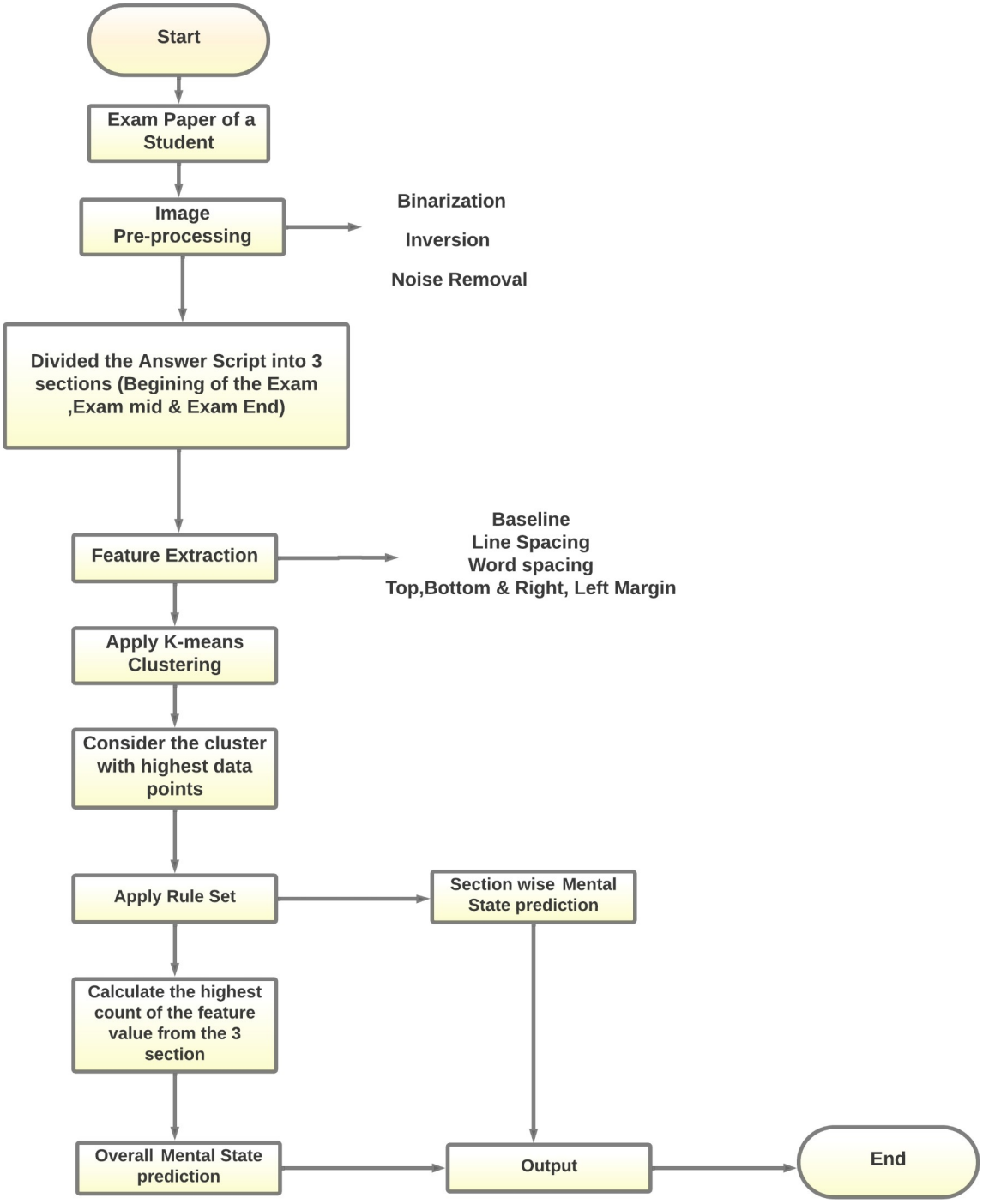


Figure 4.1: Block Diagram of Proposed Model

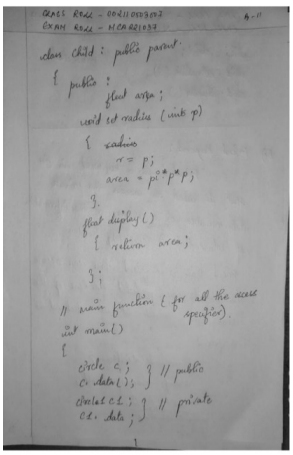
The results obtained from Experiment-1 may not be efficient owing to the lack of sufficient data and susceptibility to human bias. Additionally, a large

enough set of data might increase the time required for preparing the ground truth and might be time-consuming. Hence, in this study, the use of a rule-based classifier is also proposed as Experiment-2 by generating a set of range values for each feature of 7 randomly selected data and by testing the generated ranges on 5 data. The efficacy of the proposed approach based on analysis of 5 student data benchmarked against the annotated values.

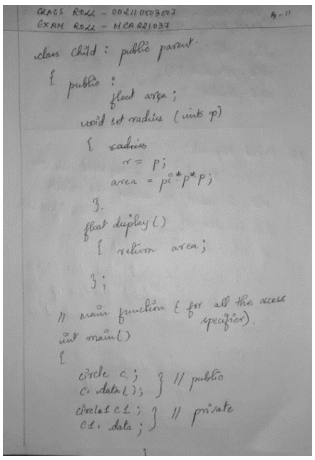
4.2. Image Pre-Processing

- **Image Area Cropping:** Image cropping is the act of improving composition or framing of the image. The input image is cropped manually.
- **Binarization:** Image Binarization is the conversion of document image into bi-level document image. Here, binarization and inversion of image is done because of the segmentation of document into foreground text and background. The grey scale images were converted to pure black and white binary images using OpenCV bilateralFilter in Python.
- **Noise Reduction:** Image noises are random variation of brightness or color information in images which produces false intensity values. We use cvtColor thresholding process provided by the Python to reduce the noise.

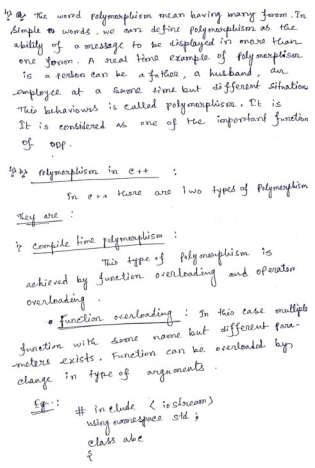
A sample pre-processing output is shown in Fig. 4.2.



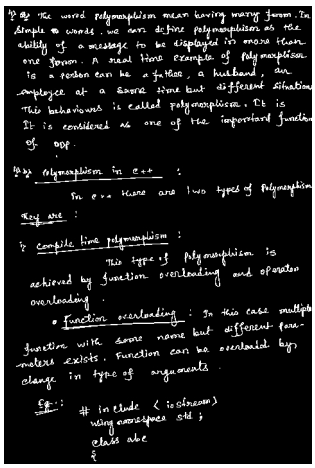
a. **Sample Student Script Page**



b. **Image area cropped**



c. **Sample Student Script Page**



d. **Binarized Image with noise reduction**

Fig. 4.2. Pre-Processing Algorithm

4.3. Feature Extraction

Feature extraction can be stated as a part of the dimensionality reduction process, where an initial set of the raw data is divided into several groups. So, the processing of these data is easier. These features can describe the actual data set with originality and accuracy.

Image processing domain is very interesting in order to understand images and here several algorithms were designed to detect under-mentioned features.

4.3.1. Baseline

Text line segmentation plays a very important role in document understanding. It is also very challenging due to various writing styles, document quality, fonts and contents of the exam papers. The limitations were analyzed and addressed. Henceforth, managed to detect each line efficiently.

Algorithm 1 Baseline Orientation

Input: An image I having pixel values ϵ (0, 1).

Output: Baseline Orientation.

1: **procedure BASELINE**

2: The I(H,W) is separated into three sections based on height of the image: (start = [h/3], mid = [2h/3], end = [h]) ⊲ H= Height of the Image and W= Width of the Image (Fig. 4.3a) and each section is analysed separately from Step 3.

#smoothing

3: I(H,W) is vertically divided into 20 windows each of length 5.

4: **while** For each 5% section of image I[H,W] along width of the image w **do**

5: Calculate the horizontal histogram ⊲ sum of all pixel intensities in I [i,0:W]

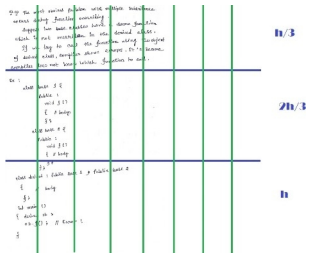
6: Find the maximum value for each index in every 5% section.

7: Calculate $x = column\_index \times 0.05 \times w$

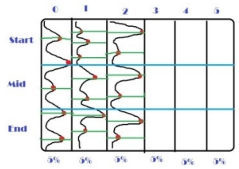
8: Draw a straight line through point (x, maximum value) in section $column\_index$.

9: Join the baselines between two adjacent 5% sections obtained using Minimum Euclidean Distance (Fig. 4.3b).

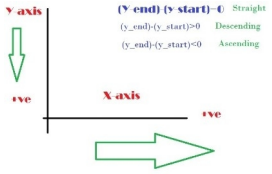
10: For each baseline, calculate slope m for the baseline (Fig. 4.3c). If (x1,y1) = starting vertex of baseline and (x2,y2) = End vertex of baseline, then, m = (y2-y1) / (x2– x1). If m>0, then baseline is Descending. If m<0, then baseline is Ascending. If m=0, then baseline is Straight.



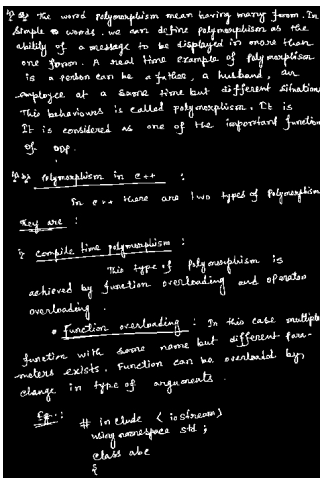
(a) In the picture it is easily noticeable that the binarized image of exam paper of a student is divided in 3 parts vertically (pointed with blue lines) start, mid and end. Each part will produce output on the basis of his/her handwriting analysis as per baseline, line and word segmentation. The green line shows that the image is divided considering the window size 5.



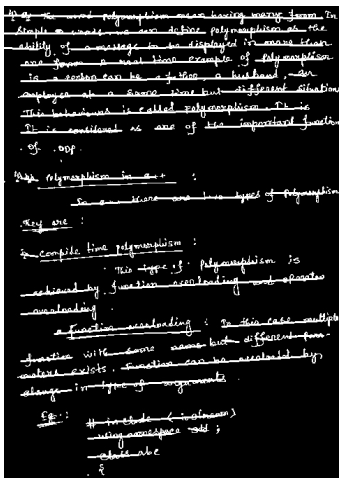
(b) For each vertical section conduct horizontal histogram which is a graph showing the number of pixels in an image at each different intensity value found in that 5%. The peak values of the histogram for every index are pointed with red dots. Considering those peak values, line plotting is done.



(c) For example, consider the data points [(11, 622), (12, 626), (13, 629)] where 11 and 13 represents the section and 622 and 629 are y-start and y-end values respectively. Hence, (629-622) =7 which is greater than 0. So, the baseline is descending.



(d) Input: Binarized Image



(e) Output: Image with Baseline

Figure 4.3: Images for Baseline Algorithm

4.3.2. Line Spacing

Algorithm 2 Line Spacing

Input : An image I (H,W) having pixel values ϵ (0, 1).

Output : Line Spacing values of the Author

1: **procedure LINE_SPACING**

2: The I(H,W) is separated into three sections based on height of the image: (start = [h/3], mid = [2h/3], end = [h]) ◅ H= Height of the Image and W= Width of the Image (Fig. 4.4a) and each section is analysed separately from Step 3.

#smoothing

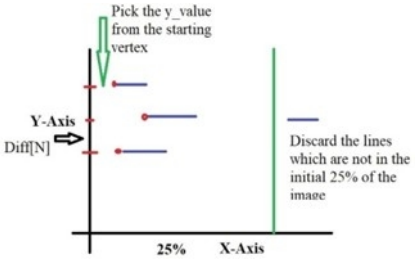
3: Consider the 25% of the image from the start and for each baseline. Pick the y_value from the starting vertex of baseline and store in an array Values [] (Fig. 4.4a). ◅ H= Height of the Image and W= Width of the Image

4: Setup DIFF[N-1] array from Values Array where, DIFF[n]=Values[n] – Values[n-1]

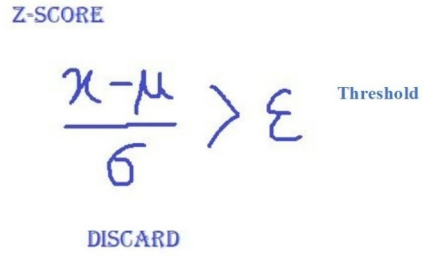
5: Sort the DIFF array obtained into ascending order.

6: Calculate the mean and std of the DIFF array and apply Z-Score for removing outliers (Fig. 4.4b).

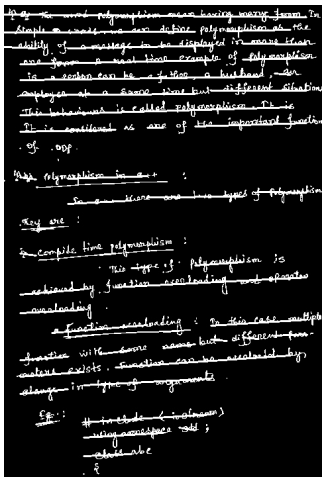
7: Remove all outliers from array DIFF and store in array LINE_SPACING[]



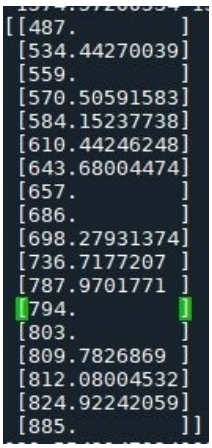
(a) The blue line presents the obtained baseline and red rod represents the starting vertex



(b) Z-Score algorithm to remove the outliers



(c) Input: Image with Baseline



(d) Output: Filtered line spacing data

Figure 4.4: Line Spacing Algorithm Flow

4.3.3. Word Spacing

Algorithm 3 Word Spacing

Input: An image I having pixel values ϵ (0, 1).

Output: Word spacing of the author

1: **procedure** WORD_SPACING

2: The I(H,W) is separated into three sections based on height of the image:

(...contd in next page)

(start = [h/3], mid = [2h/3], end = [h]) and each section is analysed separately from Step 3 to Step 8. ◅ H= Height of the Image and W= Width of the Image

3: Find all contours and store in the Contours Array.

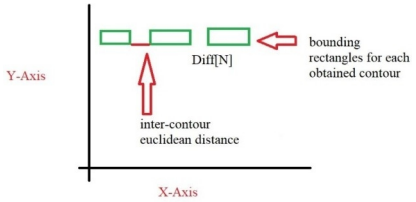
4: Find the bounding rectangles for each obtained contour in the Contours Array.

5: For each contour c in Contours, Calculate the minimum inter-contour euclidean distance between c and all other contours and store in DIFF[n].

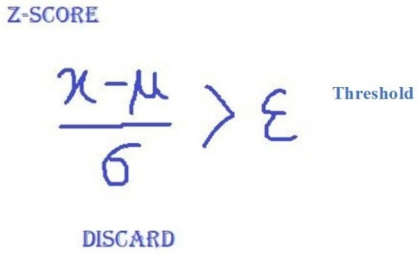
6: Sort the DIFF array obtained into ascending order.

7: Calculate the mean and std of the DIFF array and apply Z-Score for removing outliers.

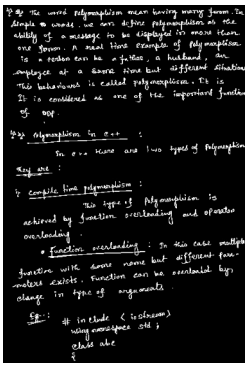
8: Remove all outliers from array DIFF and store in array WORD_SPACING[]

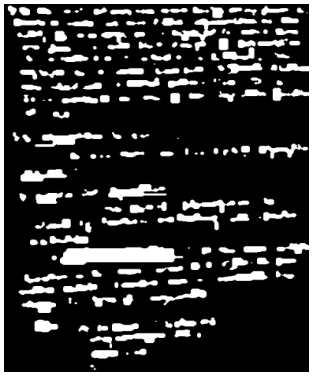


(a) Green rectangle presents the word contour. The red line shows the inter-contour distance.



(b) Z-Score algorithm to remove the outliers





(c) Input: Binarized Image

(d) Word Contour

```
Outlier in dataset is:  [1292.0, 1319.496873812136, 1321.0, 1340.839289400486, 1351.6752568572083, 1372.0,
1393.0918849810303, 1440.2836526184694, 1458.0, 1494.202128227637, 1514.0, 1569.9888534636161, 1571.0,
1574.5720053398638, 1592.4525110658717, 1614.0, 1706.15503398724, 1724.214893799494]
```

(e) Sample output: outliers with ε set to 3

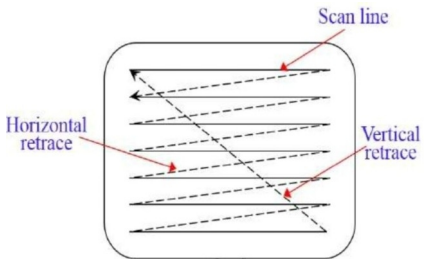
```
[[  60.          67.          69.42621983   78.          81.
    84.58132182  98.         107.87029248  110.29052543  112.36102527
   122.33151679 132.43866505 143.8401891   148.51935901  167.45148551
   195.88772294 245.         250.          254.46807265  260.9233604
   272.         294.1088234  321.67219339  355.          410.57155284
   438.         442.8205054  457.23188865  477.          487.
   534.44270039 559.         570.50591583  584.15237738  610.44246248
   643.68004474 657.         686.          698.27931374  736.7177207
   787.9701771  794.         803.          809.7826869   812.08004532
   824.92242059 885.         940.          984.         1032.53668216
  1037.08871366 1061.15267516 1067.       1096.         1124.18148001
  1132.         1137.52714253 1162.73857767 1169.02865662 1213.
  1262.00237718 1292.         1319.49687381 1321.         1340.8392894
  1351.67525686 1372.         1393.09188498 1440.28365262 1458.
  1494.20212823 1514.         1569.98885346 1571.         1574.57200534
  1592.45251107 1614.         1706.15503399 1724.2148938 ]]
```

(f) Sample output filtered data: word space after removing outliers

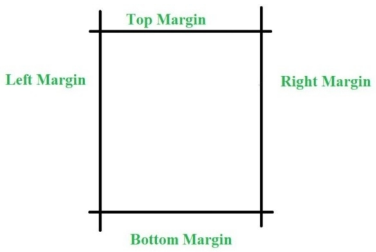
Figure 4.5: Word Spacing Algorithm Flow

4.3.4. Margin

Margin can be explained as the space left with no handwriting either to the left or to the right of the handwriting. Top and bottom margin states the starting of the handwriting from a page and end of the writing. Margins could either be narrow or wide which shows certain personality traits about the writer.



(a) Raster Scan of Document



(b) Margin Boundary

Figure 4.6: Margin Algorithm

Algorithm 4 Top Margin

Input: An image I having pixel values ϵ (0, 1).

Output: Top Margin of the author (Wide, Narrow)

1: **procedure** TOP_MARGIN

2: The I(H,W) is separated into three sections based on height of the image:

(start = [h/3], mid = [2h/3], end = [h]) and each section is analysed separately from Step 3 to Step 6. ⊲ H= Height of the Image and W= Width of the Image

3: Initialize the pixel values to 0, let Sum=0 and TopMargin = -1.

4: Add all the pixel intensities and store in Sum, Sum = Sum + I'(h,w).

5: Raster scan the image from left to right [h=0:H and w=0:W] until a non-zero pixel is found. Break the loop If (TopMargin == -1 and Sum>0).

6: Set TopMargin = h. Draw a straight line through point in section row_index.

Algorithm 5 Bottom Margin

Input: An image I having pixel values ∈ (0, 1).

Output: Bottom Margin of the author (Wide, Narrow)

1: **procedure** BOTTOM_MARGIN

2: The I(H,W) is separated into three sections based on height of the image:

(start = [h/3], mid = [2h/3], end = [h]) and each section is analysed separately from Step 3 to Step 6. ◅ H= Height of the Image and W= Width of the Image

3: Initialize the pixel values to 0, let Sum=0 and BottomMargin = -1.

4: Add all the pixel intensities and store in Sum, Sum = Sum + I'(h,w).

5: Raster scan the image from left to right [h=H:0 and w=W:0] until a non-zero pixel is found. Break the loop If (BottomMargin == -1 and Sum>0).

6: Set BottomMargin = h. Draw a straight line through point in section $rowindex$.

Algorithm 6 Right Margin

Input: An image I having pixel values ϵ (0, 1).

Output: Right Margin of the author (Wide, Narrow)

1: **procedure** RIGHT_MARGIN

2: The I(H,W) is separated into three sections based on height of the image:

(start = [h/3], mid = [2h/3], end = [h]) and each section is analysed separately from Step 3 to Step 8. ⊲H= Height of the Image and W= Width of the Image

3: Initialize the pixel values to 0, let Sum=0 and Create an array Right[H].

4: Add all the pixel intensities and store in Sum. Sum = Sum + I'(h,w).

5: Raster scan the image from right to left.

6: If Sum > 0, then, Set the array Right[H] = W.

7: Find the value in the Right array having the largest value in Right[] and return min (Right[H]).

8: Draw a straight line through point (x, minimum value) in section $column\_index$.

Algorithm 7 Left Margin

Input: An image I having pixel values ϵ (0, 1).

Output: Left Margin of the author (Wide, Narrow)

1: **procedure** LEFT_MARGIN

2: The I(H,W) is separated into three sections based on height of the image:

(start = [h/3], mid = [2h/3], end = [h]) and each section is analyzed separately from Step 3 to Step 8. ⊲ H= Height of the Image and W= Width of the Image

3: Initialize the pixel values to 0, let Sum=0 and create an array Left[H].

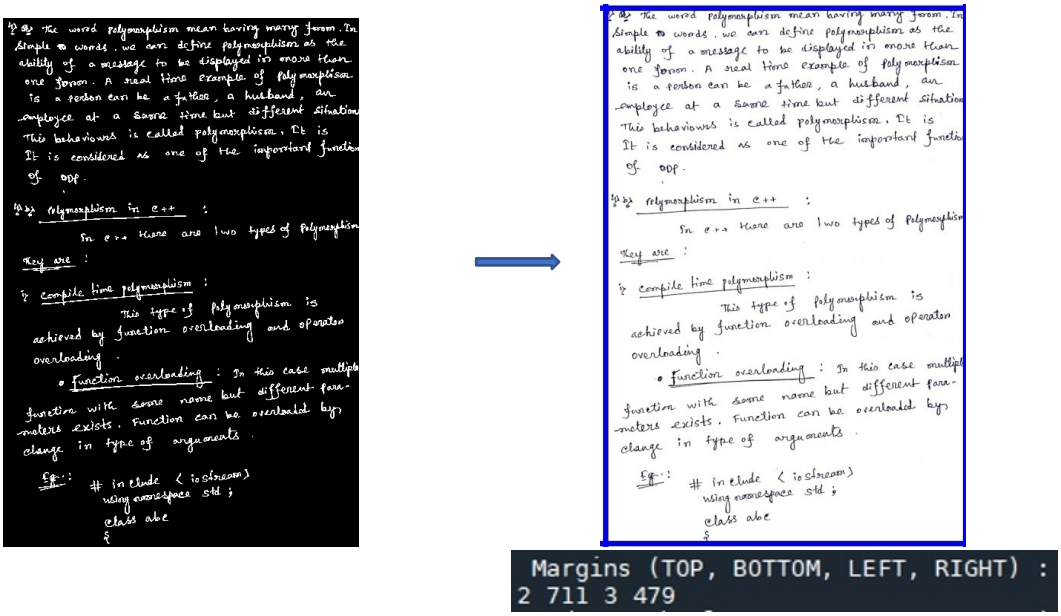
4: Add all the pixel intensities and store in Sum. Sum = Sum + I'(h,w).

5: Raster scan the image from left to right.

6: If Sum > 0, then set the array Left[H] = W.

7: Find the value in the Left array having the largest value in Left [] and return max (Left [H])

8: Draw a straight line through point (x, minimum value) in section $column\_index$.



(a) Input: Binarized Image (b) Output: Top, Bottom, Left, Right Margin

Fig 4.7. Margin Output

Algorithm 8 K-Means Clustering

Input: Any Feature value from Algorithms (1-7) for a given section from [start, mid, end]

Output: Cluster with highest data points

1: procedure K_Means_Clustering

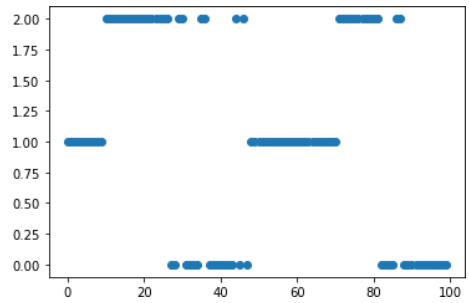
2: while for each feature f present in the section do

3: Apply K-Means clustering using Scikit-learn python library on values of the feature f for number of clusters, K=3

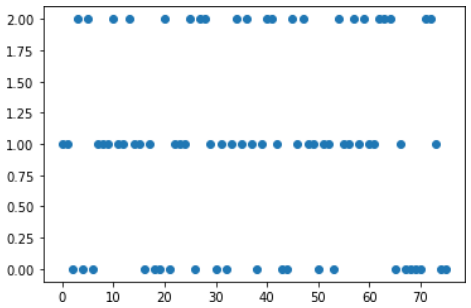
4: Set the array NPOINTS[] to be equal to the number of data points present in each cluster found in Step 3.

5: Set the variable CLUSTER to be equal to the feature value of the cluster with the largest number of data points in array NPOINTS[], found in Step 4.

6: PRINT the variable CLUSTER



(a) Sample Cluster output from algorithm 8



(b) Sample Cluster output from algorithm 8

Fig 4.8: K-Means Clustering Output

```
At the beginning of the examination:
Baseline: Straight
Left Margin: Normal
Bottom Margin: Normal
Right Margin: Narrow
Top Margin: Normal
Line Spacing: Small
Word Spacing: Normal

At the middle of the examination:
Baseline: Straight
Left Margin: Normal
Bottom Margin: Normal
Right Margin: Narrow
Top Margin: Normal
Line Spacing: Normal
Word Spacing: Narrow

At the end of the examination:
Baseline: Straight
Left Margin: Normal
Bottom Margin: Normal
Right Margin: Narrow
Top Margin: Normal
Line Spacing: Normal
Word Spacing: Wide
```

Fig 4.9: Sample Output: Section Values after applying k-means clustering

4.4. Experiment-1: Overall Assessment

An algorithm is presented next that is used to compute the accuracy of the proposed approach based on analysis of 12 student data benchmarked against the annotated values.

Algorithm 9 Overall Assessment

Input : Feature Values

Output : Accuracy

1: **procedure** OVERALL_ASSESSMENT

2: **while** For each feature f in the combined three sections **do**

3: Apply K_MEANS_CLUSTERING on feature f from Algorithm 8 and store in the array RESULT[].

3: Calculate the occurrence of each distinct feature value present in array RESULT[] and store in array COUNT[].

4: For all feature value occurrences c in array COUNT[], if the number of occurrences of c is greater than all other feature value occurrences, then overall value for feature f is c.

5: For all feature value occurrences c, if the number of occurrences of c is equal to all other feature value occurrences, then overall value for feature f is considered to be Indecisive.

6: For each feature f, output the mentality trait based on the overall value for f.

7: For each feature f, calculate the Accuracy of the feature f.

8: Print the Accuracy.

4.5. Experiment-1: Psychological Assessment Rule Set

Based on the qualitative aspects studied as part of this work, the following mental states have been derived from the feature values for this work.

Table 4.1: Baseline Oriented mentality prediction of the Student's During Exam

Serial No.	**Type of Baseline**	**Student's Mentality During Exam**
1	Straight	indicates high confidence
2	Ascending	is optimistic, in good mood but may be over-confident
3	Descending	May be depressed and suffering from lack of self-confidence

Table 4.2: Line Spacing related mentality prediction of the Student's During Exam

Sl No.	**Space in between lines**	**Student's Mentality During Exam**
1	Lines Widely Apart	Confident.
2	Medium	Is a balanced person, does not take too much exam stress.
3	Lines crowding to- gether	indicates fear of faring poorly in the exam.

Table 4.3: Word Spacing related mentality prediction of the Student's During Exam

Serial No.	**Space in between words**	**Student's Mentality During Exam**
1	Too far apart	Not quite confident with the content of own answer, whatever the source!

Serial No.	**Space in between words**	**Student's Mentality During Exam**
2	Adequate spacing	is confident about the portion of the answer under consideration
3	Inadequate Spac-ing	may have disturbed others and constantly asked for help.

Table 4.4: Page Margin related personality traits

Sl No.	**Margin Type**	**Student's Mentality During Exam**
1	Wide upper margin	wants to create good impression.
2	Narrow Upper Margin	seems very casual and informal in approach.
3	Wide lower margin	seems to be afraid of the outcome.
4	Narrow lower Margin	indicates overall lack of confidence.
5	Wide right margin	seems uncertain about completeness of answer.
6	Narrow right Margin	indicates complete confidence.
7	Wide left margin	Self-confident.
8	Narrow left Margin	Nervous and tense.

Some points relevant for Answer scripts specifically [where the documents are broken up in pages, with 3 sections within each page]:

Table 4.5: Varying Baseline Type

Varying Baseline type within a single page or even section:	
Straight and Ascending	The student starts the exam confidently and continues in the same mood.
Straight and Descending	The student begins confidently, but becomes more uncertain with time.
Ascending and Descending	Indicates fluctuating mood.
Ascending and Straight	The student is optimistic and continues the exam confidently
Descending and Straight	The student starts the exam with lack of self-confidence and regains confidence with time.

Table 4.6: Varying Line Spacing Type

Varying Line Spacing within a single page or even section:	
Uneven Line Spacing	indicates lack of proper planning and overall unpreparedness.
Uneven Line Spacing	indicates a balanced mind well-prepared to tackle the questions methodically.
Even Word Spacing	cautious, but handling the situation adequately.
Uneven Word Spacing	haphazard approach indicative of nervousness, may be due to lack of preparatory efforts.

Table 4.7: Varying Margin Type

Variable Margin within page or even section:	
Uneven Left Margin	indicates careless behavior.
Uneven Right Margin	implies self-consciousness.
Uneven Top Margin	indicates lack of good taste.

..Contd in next page

Variable Margin within page or even section: (Table 4.7: Varying Margin Type)	
Uneven Bottom Margin	reveals the laziness, indecision.
Even Margin	indicates good manners and regularity.
Narrow Margin Through-out	implies determination to achieve a goal.

The algorithms presented above provide a means to establish the sentiment analysis of the student during examination. Based on the feature values assigned, the final assessment is generated which is indicative of the mental condition of the student during the examination. For the purpose of evaluation, the metric of Accuracy is calculated based on the output obtained from the algorithm.

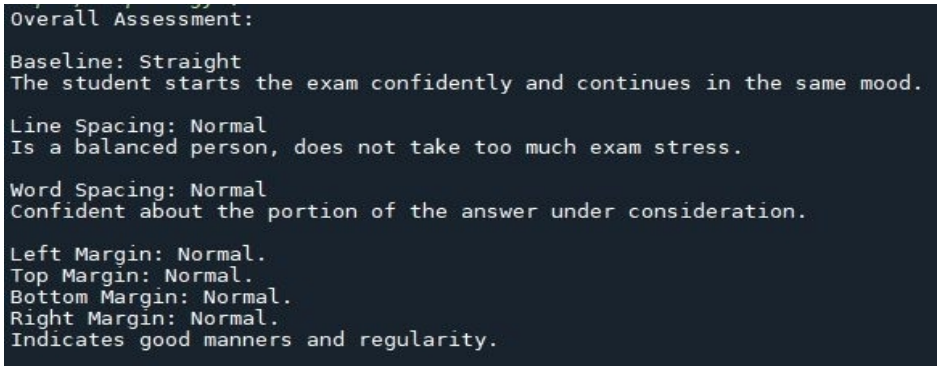


Fig 4.10: Sample Output: Overall Psychological Assessment

4.6. Experiment-2: Range Calculation for Rule-based Classifier

Additionally, as a $2nd$ part of the experiment, seven features namely Baseline (Algorithm 1), Line Spacing (Algorithm 2), Word Spacing (Algorithm 3) and the Margins (Algorithm 4,5,6,7) are extracted to generate ranges for each feature.

To evaluate the proposed experiment, the obtained dataset is divided into training set and test set. 7 answer scripts are randomly picked from the initial dataset and placed into training set. The train dataset is used to obtain a set of Decision Rules to identify the Class Label of the feature value. To obtain the feature range values for each class, a K-Means clustering model is utilized (Explained as Algorithm 8 in Experiment-1). Firstly, each set of feature values is passed as an input to a K-Means Clustering model with number of clusters set to 3, namely, Narrow, Wide and Normal (for baseline, the class labels are Straight, Ascending and Descending).

After formulating the clustering model, the range of values that belong to the Normal cluster is calculated of which, a limit of 80% of mean is imposed on either side of the range of values to obtain the minimum and maximum limit of feature values for the Normal cluster (or Straight feature value for Baseline feature). The Decision Rule for each feature thus obtained based on the train dataset is presented in Table 5.11 – 5.17 in the next chapter.

Algorithm 10 Calculating Range

1: ***procedure*** *CALCULATING_RANGE*

2: while For each feature f in the combined three sets do

3: Set Feature_List = List of feature values present in the normal region for feature f.

4: Sort Feature_List in ascending order.

5: Set Mean = Average(Feature_List)

6: Set $Range_f^{LOW} = Mean - 0.8 * Mean$, $Range_f^{HIGH} = Mean + 0.8 * Mean$

7: Print $Range_f^{LOW}$, $Range_f^{HIGH}$

Algorithm 9, already explained in Experiment-1 is also used here to compute the accuracy of the proposed approach based on analysis of 5 student data benchmarked against the annotated values.

Chapter5

Result and Analysis

In this chapter the results of the experiments on the student dataset are presented for evaluating different feature sets, namely, baseline, line spacing, word spacing and margin.

5.1. Experimental Setup

The experimental setup of the proposed algorithm is presented in this section. The experiments have been performed on an Intel I5 processor with 8GB RAM taking an average of 32 seconds to process each folder for a student.

5.1.1. Experiment-1

A comparative analysis with the prepared ground truth (as shown in left panel named Human Annotated Section Values in Table 5.1-5.7) and obtained from Algorithms 1, 2, 3, 4, 5, 6, 7 in Chapter 4 are reported in the tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6 and 5.7 below for all 12 available student datasets. The accuracy values presented at the end of each table are the percentages of test values that match with the human annotated values.

Table 5.1: Baseline Output

Baseline						
Data #	**Human Annotated Section Values**			**Section Values from Test Output (Algorithm 1)**		
	Start	**Middle**	**End**	**Start**	**Middle**	**End**
1	Ascending	Ascending	Straight	Ascending	Straight	Straight
2	Ascending	Straight	Straight	Ascending	Straight	Straight
3	Straight	Straight	Straight	Straight	Straight	Straight
4	Descending	Straight	Straight	Descending	Straight	Straight

Table 5.1: Baseline Output (contd.)

Baseline						
Data #	**Human Annotated Section Values**			**Section Values from Test Output (Algorithm 1)**		
	Start	**Middle**	**End**	**Start**	**Middle**	**End**
1	Ascending	Ascending	Straight	Ascending	Straight	Straight
2	Ascending	Straight	Straight	Ascending	Straight	Straight
3	Straight	Straight	Straight	Straight	Straight	Straight
4	Descending	Straight	Straight	Descending	Straight	Straight
5	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing
6	Straight	Straight	Straight	Descend-ing	Straight	Straight
7	Ascending	Straight	Ascending	Ascending	Straight	Ascending
8	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing
9	Ascending	Ascending	Ascending	Ascending	Ascending	Ascending
10	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing	Descend-ing
11	Ascending	Straight	Ascending	Ascending	Straight	Ascending
12	Straight	Straight	Straight	Ascending	Straight	Straight
	Accuracy:			**83.33%**	**91.67%**	**100%**

Inference: Average Accuracy for Baseline is 91.67% (Table 5.1). Rating above 90% may be taken as proof of reasonable success, specially as the bulk of each student script was heavy.

Table 5.2: Line Spacing Output

Line Spacing						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 2)		
	Start	Middle	End	Start	Middle	End
1	Normal	Narrow	Normal	Normal	Normal	Normal
2	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
3	Narrow	Normal	Normal	Wide	Wide	Wide
4	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
5	Normal	Normal	Normal	Normal	Normal	Normal
6	Normal	Normal	Narrow	Normal	Normal	Narrow
7	Normal	Normal	Narrow	Normal	Normal	Narrow
8	Normal	Normal	Narrow	Narrow	Narrow	Narrow
9	Wide	Narrow	Narrow	Wide	Narrow	Narrow
10	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
11	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
12	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
Accuracy:				**83.33%**	**83.33%**	**91.67%**

Inference: Average Accuracy for Line Spacing is 86.11% (Table 5.2). Here success rate lags to some extent. Reasons that may be considered is that the scripts contained programs and calculations and diagrams, on which line spacing can be better assessed by a human annotator. Frequent occurrences of uneven line sizes in a page poses problems for the image processing tools used for extracting feature values (Fig. 5.1).

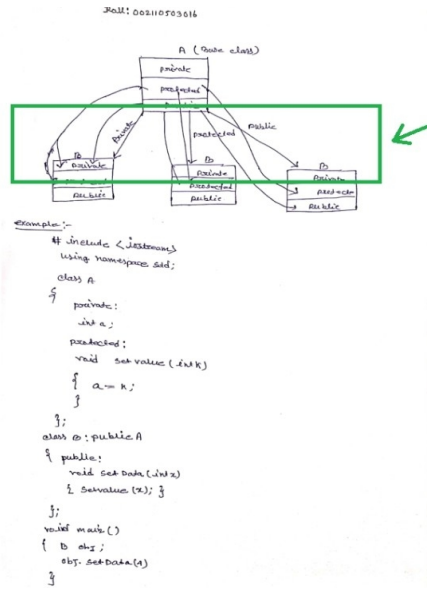


Figure 5.1: Sample Student Script Page: The image contains a diagram because of which Line Spacing algorithm (Algorithm 2) fails. The zone is marked with green arrow.

Table 5.3: Word Spacing Output

Word Spacing						
Data #	**Human Annotated Section Values**			**Section Values from Test Output (Algorithm 3)**		
	Start	**Middle**	**End**	**Start**	**Middle**	**End**
1	Normal	Normal	Normal	Normal	Normal	Normal
2	Narrow	Normal	Wide	Narrow	Narrow	Narrow
3	Normal	Normal	Normal	Normal	Normal	Normal
4	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
5	Normal	Normal	Narrow	Normal	Normal	Narrow
6	Narrow	Narrow	Normal	Narrow	Narrow	Normal
7	Wide	Normal	Normal	Wide	Normal	Normal
8	Normal	Narrow	Normal	Normal	Narrow	Normal
9	Normal	Wide	Normal	Normal	Wide	Normal
10	Normal	Wide	Normal	Normal	Wide	Normal
11	Narrow	Narrow	Normal	Narrow	Narrow	Normal
12	Normal	Narrow	Narrow	Normal	Narrow	Narrow
Accuracy:				**100%**	**91.67%**	**91.67%**

Inference: Average Accuracy for Word Spacing is 94.44% (Table 5.3). This value is promising indeed!

Table 5.4: Top Margin Output

Top Margin						
Data#	Human Annotated Section Values			Section Values from Test Output (Algorithm 4)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Narrow	Normal	Normal	Narrow
2	Normal	Normal	Normal	Normal	Normal	Normal
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
5	Normal	Normal	Narrow	Normal	Normal	Narrow
6	Narrow	Narrow	Narrow	Normal	Narrow	Narrow
7	Normal	Narrow	Narrow	Normal	Narrow	Narrow
8	Narrow	Narrow	Narrow	Narrow	Narrow	Wide
9	Normal	Normal	Narrow	Normal	Normal	Narrow
10	Normal	Normal	Normal	Normal	Normal	Normal
11	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
12	Normal	Normal	Narrow	Normal	Normal	Narrow
Accuracy:				**83.33%**	**100%**	**91.67%**

Inference: Average Accuracy for Top Margin is 91.67% (Table 5.4). This one can also be assumed to be reasonably good.

Table 5.5: Bottom Margin Output

Bottom Margin						
Data#	Human Annotated Section Values			Section Values from Test Output (Algorithm 5)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Narrow	Narrow	Normal	Narrow
2	Narrow	Narrow	Normal	Narrow	Narrow	Normal
3	Narrow	Normal	Narrow	Narrow	Normal	Narrow
4	Normal	Narrow	Normal	Normal	Narrow	Normal
5	Normal	Narrow	Narrow	Normal	Narrow	Narrow
6	Narrow	Normal	Narrow	Narrow	Normal	Narrow
7	Normal	Normal	Wide	Normal	Normal	Wide
8	Normal	Narrow	Narrow	Normal	Narrow	Narrow
9	Normal	Normal	Normal	Normal	Normal	Narrow
10	Normal	Normal	Normal	Normal	Normal	Normal

Bottom Margin						
Data#	Human Annotated Section Values			Section Values from Test Output (Algorithm 5)		
	Start	Middle	End	Start	Middle	End
11	Narrow	Narrow	Normal	Narrow	Narrow	Normal
12	Normal	Narrow	Normal	Normal	Narrow	Narrow
Accuracy:				**100%**	**100%**	**83.33%**

Inference: Average Accuracy for Bottom Margin is 94.44% (Table 5.5). This one is another promising marker.

Table 5.6: Right Margin Output

Right Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 6)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Normal	Narrow	Normal	Normal
2	Normal	Normal	Normal	Normal	Normal	Normal
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Normal	Narrow	Narrow	Normal
5	Normal	Narrow	Normal	Normal	Narrow	Normal
6	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
7	Normal	Normal	Narrow	Normal	Normal	Narrow
8	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
9	Narrow	Normal	Narrow	Normal	Normal	Narrow
10	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
11	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
12	Normal	Normal	Normal	Narrow	Narrow	Narrow
Accuracy:				**83.33%**	**91.67%**	**91.67%**

Inference: Average Accuracy for Right Margin is 88.89% (Table 5.6). It is just short of the 90% marker which may be due to the difficulty in deciphering the true right margin nature due to the sample containing programming texts, which are often indented perceptibly. Fig. 5.2 below demonstrates a sample student script page containing program code. This illustrates why the right margin may be difficult to detect if a page contains varying right margin as shown using blue arrows.

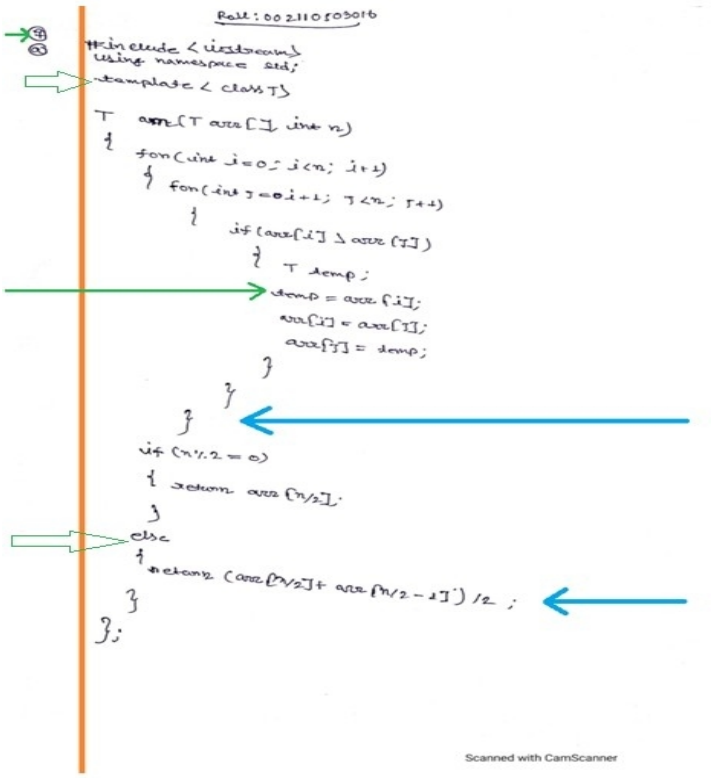


Figure 5.2: Sample Student Script Page

Table 5.7: Left Margin Output

Left Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 7)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Normal	Narrow	Normal	Normal
2	Wide	Wide	Normal	Normal	Normal	Normal
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Normal	Narrow	Narrow	Normal
5	Normal	Normal	Normal	Normal	Normal	Normal
6	Narrow	Narrow	Normal	Narrow	Normal	Narrow
7	Narrow	Normal	Narrow	Narrow	Normal	Narrow
8	Narrow	Narrow	Normal	Narrow	Narrow	Normal
9	Normal	Narrow	Narrow	Normal	Narrow	Narrow
10	Narrow	Normal	Normal	Narrow	Normal	Wide
11	Narrow	Normal	Narrow	Narrow	Normal	Narrow
12	Wide	Wide	Normal	Wide	Wide	Normal
Accuracy:				**91.67%**	**83.33%**	**83.33%**

Inference: Average Accuracy for Left Margin is 86.11%. The reason behind the poor rating may be indicating that jagged ends in handwritten text lines are a hurdle for the image processing tools to decipher. While a human annotator can base their reasoning only on the non-program texts, the machine learning tools used here are still not adept at deciphering contextual exclusion. Fig. 5.2 above illustrates this point too by marking the jagged left margin with green arrows. The left margin should have been aligned with the orange line shown in the picture.

The cumulative analysis on all the foregoing accuracy values are represented in the summary Table 5.8 below. It also carries an Overall Accuracy 90.47% for all the features taken together.

After assessing each part of the entire answer script for each student, the final objective is to provide an overall assessment on the mentality state of the student during the duration of the examination.

Table 5.8: Overall Assessment (Algorithm 8)

Sl No.	**Feature Name**	**Average Accuracy Rating for 3 Sections**			**Average Accuracy**
		Start	**Middle**	**End**	
1	Baseline	83.33%	91.67%	100%	**91.67%**
2	Line Spacing	83.33%	83.33%	91.67%	**86.11%**
3	Word Spacing	100%	91.67%	91.67%	**94.44%**
4	Top Margin	83.33%	100%	91.67%	**91.67%**
5	Bottom Margin	100%	100%	83.33%	**94.44%**
6	Left Margin	91.67%	83.33%	83.33%	**86.11%**
7	Right Margin	83.33%	91.67%	91.67%	**88.89%**
Overall Accuracy for All features:					**90.47%**

The obtained 4 feature values for all the 12 answer scripts in all the 3 sections is calculated and presented below in Table 5.9. The blue color is used to mark the dominant feature value in all the 3 sections.

Table 5.9: Data Wise Feature Extraction

Sl No.	Baseline	Line Spacing	Word spacing	Margin			
				Top	Bottom	Left	Right
1	Ascending/ Straight	Normal	Normal	Normal /Narrow	Normal /Narrow	Normal /Narrow	Normal /Narrow
2	Ascending/ Straight	Narrow	Narrow	Normal	Normal /Narrow	Normal	Normal
3	Straight	Wide	Normal	Narrow	Normal /Narrow	Narrow	Narrow
4	Descend-ing/ Straight	Narrow	Narrow	Narrow	Normal /Narrow	Normal /Narrow	Normal /Narrow
5	Descending	Normal	Normal	Normal /Narrow	Normal /Narrow	Normal /Narrow	Normal
6	Descending /Straight	Normal /Narrow	Narrow /Normal	Normal /Narrow	Normal /Narrow	Narrow	Normal /Narrow
7	Ascending / Straight	Normal /Narrow	Wide/ Normal	Normal /Narrow	Normal /Wide	Normal /Narrow	Normal /Narrow
8	Descending	Narrow	Normal /Narrow	Narrow /Wide	Normal /Narrow	Narrow	Normal /Narrow
9	Ascending	Wide/ Narrow	Normal/ Wide	Normal	Normal /Narrow	Normal /Narrow	Normal /Narrow
10	Descending	Narrow	Normal/ Wide	Normal	Normal	Narrow	Narrow / Normal / Wide
11	Ascending / Straight	Narrow	Narrow /Normal	Narrow	Normal /Narrow	Narrow	Normal /Narrow
12	Ascending/ Straight	Narrow	Narrow /Normal	Normal /Narrow	Normal /Narrow	Narrow	Wide/ Normal

Based on the above table overall mental state is predicted for each student according to the rule table provided for students in the Methodology Chapter (Table 4.2,4.3,4.4, 4.5, 4.6, 4.7).

Table 5.10: Data Wise Overall Assessment

Sl No.	Overall Assessment
1	• The student starts the exam confidently and continues in the same mood. • Is a balanced person, does not take too much exam stress. • Confident about the portion of the answer under consideration. • Indicates good manners and regularity.
2	• The student starts the exam confidently and continues in the same mood. • indicates fear of faring poorly in the exam. • May have disturbed others and constantly asked for help. • indicates good manners and regularity. • reveals the laziness, indecision.
3	• Indicates high confidence • Confident. • Confident about the portion of the answer under consideration. • Implies determination to achieve a goal.
4	• The student starts the exam with lack of self-confidence and regains confidence while writing. • indicates fear of faring poorly in the exam. • May have disturbed others and constantly asked for help. • indicates careless behavior, • implies self-consciousness, • reveals the laziness, indecision.
5	• May be depressed and suffering from lack of self-confidence. • Is a balanced person, does not take too much exam stress. • Confident about the portion of the answer under consideration. • indicates good manners and regularity. • indicates careless behavior, • implies lack of good taste, • reveals the laziness, indecision.
6	• The student starts the exam with lack of self-confidence and regains confidence while writing. • Indicates lack of proper planning and overall unpreparedness. • haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • Implies determination to achieve a goal • .. contd (next page)

Table 5.10: Data Wise Overall Assessment
(contd.)

Sl No.	Overall Assessment
7	• The student is optimistic and continues the exam confidently. • indicates lack of proper planning and overall unpreparedness. • Haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • indicates careless behavior, • implies self-consciousness, • reveals the laziness, indecision. • implies lack of good taste.
8	• May be depressed and suffering from lack of self-confidence . • indicates fear of faring poorly in the exam. • Haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • Implies determination to achieve a goal.
9	• is optimistic, in good mood but may be over-confident. • indicates lack of proper planning and overall unpreparedness. • Haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • Indicates good manners and regularity. • indicates careless behavior, • implies self-consciousness, • reveals the laziness, indecision.
10	• May be depressed and suffering from lack of self-confidence . • indicates fear of faring poorly in the exam. • haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • Nervous and tense. • implies self-consciousness.
11	• The student is optimistic and continues the exam confidently. • indicates fear of faring poorly in the exam. • Haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • Implies determination to achieve a goal.
12	• The student starts the exam confidently and continues in the same mood. • indicates fear of faring poorly in the exam. • Haphazard approach indicative of nervousness, may be due to lack of preparatory efforts. • implies lack of good taste. • seems uncertain about completeness of answer. • reveals the laziness, indecision.

5.1.2. Experiment-2

The Decision Rule for each feature obtained at this stage based on the train dataset helps to detect the class values for the test dataset. The rules obtained are presented in the following Tables 5.11 – 5.17.

Table 5.11: Baseline Rule

Baseline		
Feature	**Range**	**Class label**
Ascending	Baseline <-2 degree	0
Straight	-2 degree< Baseline< +2 degree	1
Descending	Baseline >+2 degree	2

Table 5.12: Line Spacing Rules

Line Spacing		
Feature	**Range**	**Class label**
Small	Spacing <106 pixels	0
Normal	106 pixels <Spacing< 500 pixels	1
Large	Spacing> 500 pixels	2

Table 5.13: Word Spacing Rules

Word Spacing		
Feature	**Range**	**Class label**
Small	Spacing <216 pixels	0
Normal	216 pixels <Spacing <502 pixels	1
Large	Spacing <502 pixels	2

Table 5.14: Top Margin Rules

Top margin		
Feature	**Range**	**Class label**
Small	Margin <5 pixels	0
Normal	5 pixels <Margin <16 pixels	1
Large	Margin >16 pixels	2

Table 5.15: Bottom Margin Rules

Bottom Margin		
Feature	**Range**	**Class label**
Small	Margin <20 pixels	0
Normal	20 pixels <Margin <89 pixels	1
Large	Margin >89 pixels	2

Table 5.16: Left Margin Rules

Left Margin		
Feature	**Range**	**Class label**
Small	Margin< 14 pixels	0
Normal	14 pixels <Margin< 54 pixels	1
Large	Margin >54 pixels	2

Table 5.17: Right Margin Rules

Right Margin		
Feature	**Range**	**Class label**
Small	Margin <4 pixels	0
Normal	4 pixels <Margin <49 pixels	1
Large	Margin >49 pixels	2

The class labels of test data generated by the rule-based classifier are reported and benchmarked in the Tables 5.18 – 5.24 below.

Table 5.18: Baseline Test Data Output

Baseline						
Data #	**Human Annotated Section Values**			**Section Values from Test Output (Algorithm 1)**		
	Start	**Middle**	**End**	**Start**	**Middle**	**End**
1	Ascending	Ascending	Straight	Ascending	Straight	Straight
2	Ascending	Straight	Straight	Ascending	Straight	Straight
3	Straight	Straight	Straight	Straight	Straight	Straight
4	Descending	Straight	Straight	Descending	Straight	Straight
5	Descending	Descending	Descending	Descending	Descending	Descending
Accuracy:				**100%**	**80%**	**100%**

Table 5.19: Line Spacing Test Data

Line Spacing						
Data #	**Human Annotated Section Values**			**Section Values from Test Output (Algorithm 2)**		
	Start	**Middle**	**End**	**Start**	**Middle**	**End**
1	Normal	Narrow	Normal	Narrow	Narrow	Narrow
2	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
3	Narrow	Normal	Normal	Narrow	Narrow	Narrow
4	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
5	Normal	Normal	Normal	Narrow	Narrow	Narrow
Accuracy:				**60%**	**60%**	**40%**

Table 5.20: Word Spacing Test Data

Word Spacing						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 3)		
	Start	Middle	End	Start	Middle	End
1	Normal	Normal	Normal	Normal	Normal	Normal
2	Narrow	Normal	Wide	Normal	Wide	Wide
3	Normal	Normal	Normal	Normal	Normal	Normal
4	Narrow	Narrow	Narrow	Normal	Normal	Normal
5	Normal	Normal	Narrow	Normal	Normal	Normal
Accuracy:				**60%**	**60%**	**60%**

Table 5.21: Top Margin Test Data

Top Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 4)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Narrow	Normal	Narrow	Narrow
2	Normal	Normal	Normal	Narrow	Narrow	Narrow
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
5	Normal	Normal	Narrow	Normal	Normal	Normal
Accuracy:				**80%**	**60%**	**60%**

Table 5.22: Bottom Margin Test Data

Bottom Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 5)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Narrow	Narrow	Narrow	Normal
2	Narrow	Narrow	Normal	Narrow	Narrow	Narrow
3	Narrow	Normal	Narrow	Narrow	Narrow	Narrow
4	Normal	Narrow	Normal	Narrow	Narrow	Narrow
5	Normal	Narrow	Narrow	Narrow	Narrow	Narrow
Accuracy:				**60%**	**60%**	**40%**

Table 5.23: Left margin Test Data

Left Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 7)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Normal	Narrow	Narrow	Normal
2	Normal	Wide	Normal	Normal	Wide	Normal
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Normal	Narrow	Narrow	Narrow
5	Normal	Normal	Normal	Normal	Narrow	Normal
Accuracy:				**100%**	**60%**	**100%**

Table 5.24: Right Margin Test Data

Right Margin						
Data #	Human Annotated Section Values			Section Values from Test Output (Algorithm 6)		
	Start	Middle	End	Start	Middle	End
1	Narrow	Normal	Normal	Narrow	Narrow	Narrow
2	Normal	Normal	Normal	Wide	Wide	Normal
3	Narrow	Narrow	Narrow	Narrow	Narrow	Narrow
4	Narrow	Narrow	Normal	Narrow	Narrow	Narrow
5	Normal	Narrow	Normal	Narrow	Narrow	Normal
Accuracy:				**60%**	**60%**	**60%**

Based on the formulated ground truth, the test set data is evaluated, on the basis of each Section, for each feature and the output is obtained. The Accuracy metric has been utilized in this study for the purpose of evaluation and the results are presented in Table 5.25.

Table 5.25: Overall Assessment

Sl No.	Feature Name	Accuracy			Average Accuracy
		Section-1	Section-2	Section-3	
1	Baseline	100%	100%	80%	**93.33%**
2	Line Spacing	60%	60%	40%	**53.33%**
3	Word Spacing	60%	60%	60%	**60%**
4	Top Margin	80%	60%	60%	**66.67%**
5	Bottom Margin	60%	60%	40%	**53.33%**
6	Left Margin	100%	60%	100%	**86.67%**
7	Right Margin	60%	60%	60%	**60%**
Overall Accuracy:					**67.61%**

Inference: From Table 5.25, it can be observed that the proposed algorithm performed well on Baseline Orientation detection feature (93.33%) and left margin (86.67%) whereas it performed relatively poorly for all the rest, Line Spacing and Bottom Margin features faring the lowest with 53.33% Accuracy. As a result, the accuracy achieved for Experiment 2 proves insufficient to establish the efficacy of our algorithm. The primary cause for the poor accuracy can be attributed to lack of sufficient data. A larger set of data can be collected to improve the overall accuracy obtained.

With the availability of sufficiently large set of data, incremental learning can be implemented and tested as well to improve the accuracy. Individual ranges of feature for all data can also be considered for evaluation, however, the resultant model might not have sufficient variation of feature values which might generate heavily biased results. Hence, in future, new data points may be added to the model incrementally, as and when they become available, to constantly update the range limit of values for each feature. The model would be considered optimized when accuracy rate stabilizes.

Chapter 6

Conclusion and Future Work

6.1. Conclusion

In this study, an overall approach for "Sentiment Prediction from Online Examination Scripts using Computer Aided Graphology" has been proposed that can be used to analyze the mental state of students in a controlled stress induced environment of examination. A detailed study of past research work in the field of Graphology has been presented in in Chapter 2 of this dissertation. Here, a dataset comprising of 12 students' answer scripts is collected and two experiments are conducted. The first experiment analyses the data to deduce the efficacy of the features used for the study. The seven features involved in this task are Line Spacing, Word Spacing, Baseline and Top Margin, Bottom Margin, Left Margin and Right Margin. An overall accuracy of 90.47% has been achieved in this task which helps to support the success of Experiment-1.

Chapter 4 in this study highlights the methodology used for feature extraction, feature evaluation and building up a rule-based classifier model. This last task required a second experiment involving 7 randomly chosen datapoints out of the original 12 for training purpose. K-Means clustering algorithm has been applied separately on this training set to determine the border values of the middle most cluster. This forms a set of decision rules to evaluate class values of new data points, thus constructing the basis of a rule-based classifier. The remaining 5 data points comprised the test set, whose class values are determined using the rule-based classifier and then compared with the annotated values to obtain accuracy of the model This has been found to be 67.61% for Experiment-2 in this study. A detailed analysis of the Result is presented as part of Chapter 5.

The actual implications of the feature values are also presented in this chapter in the form of student characteristic tables based on the rules of graphology. As already discussed, this technique of evaluation is completely non-evasive and provides a handy mechanism to detect and prevent way-ward mentality within the student community. The additional advantage of the experiments lie in the dataset used as it indicates the mental process of

the author under stressed condition – the examination in this instance. This allows the researcher to extrapolate the results for other real life situations and create ground for fruitful assistance early in life.

6.2. Future Work

Further extension of the work presented in this study may be incorporated by considering features such as Pen Pressure, Character tilts and size Secondly, the arena of deep learning techniques may be explored for forming the rule-based classifier. Additionally, outputs of such a classifier can be benchmarked more effectively by improving the quality of the prepared ground truth with the assistance of experienced graphologists.

BIBLIOGRAPHY

1. T. Howells, G. Allport, and P. Vernon, “Studies in expressive movement,” The American Journal of Psychology, vol. 46, no. 4, p. 667, 1934. doi: 10.2307/ 1415516.

2. S. Ghosh, P. Shivakumara, P. Roy, U. Pal, and T. Lu, “Graphology based hand- written character analysis for human behavior identification,” CAAI Transac- tions on Intelligence Technology, vol. 5, no. 1, 2020. doi: 10.1049/trit. 2019.0051.

3. A. Anand, D. Patil, S. Bhaawat, S. Karanje, and V. Mangalvedhekar, “Auto- mated career guidance using graphology, aptitude test and personality test,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1–5. doi: 10.1109/ICCUBEA.2018. 8697642.

4. M. Arivazhagan, H. Srinivasan, and S. Srihari, “A statistical approach to line segmentation in handwritten documents,” in Proceedings of SPIE - The International Society for Optical Engineering 6500, 2007. doi: 10.1117/12. 704538.

5. S. Hashemi, B. Vaseghi, and F. Torgheh, “Graphology for farsi handwriting using image processing techniques,” 2015.

6. S. H. Fatimah, E. C. Djamal, R. Ilyas, and F. Renaldi, “Personality features identification from handwriting using convolutional neural networks,” in 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2019, pp. 119– 124. doi: 10.1109/ ICITISEE48480.2019.9003855.

7. D. Dahiya, “Personality profile through handwriting analysis a textbook of handwriting analysis,”.

8. Chapman and G. Allport, “Personality: A Psychological Interpretation.”,

Sociometry, vol. 1, no. 34, p. 420, 1938. Available: 10.2307/2785590.