

**EMOTION PREDICTED THROUGH
SOCIAL MEDIA INTERACTION
AND
GRAPHOLOGY**

Project Report Submitted in Partial Fulfilment of the
Requirements for the degree of
Master of Computer Application
Of
Jadavpur University
May, 2019

By
Rishi Dey
Master of Computer Application – III
Examination Roll Number: MCA196014
Registration Number: 12477 of 2013 – 2014

Under the guidance of
Dr. CHITRITA CHAUDHURI
Associate Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata – 700032, India
May, 2019

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

TO WHOM IT MAY CONCERN

I hereby forward the project report entitled “*Emotion Predicted Through Social Media Interaction and Graphology*” prepared by **Rishi Dey** under my supervision to be accepted in partial fulfilment for the degree of **Master of Computer Application** in the Faculty and Technology of Jadavpur University, Kolkata.

(Dr.Chitrita Chaudhuri)

Associate Professor

Project Supervisor

Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032

Countersigned:

Prof. Mahantapas Kundu

Head, Dept. of Computer Science and Engineering

Jadavpur University

Kolkata – 700032

Prof.Chiranjib Bhattacharjee

Dean, Faculty of Engineering and Technology

Jadavpur University

Kolkata – 70032

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL *

The foregoing project report is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessary endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project report only for the purpose for which it has been submitted.

Final Examination for
evaluation of the project

(Signatures of Examiners)

* Only in case the project report is approved

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this project report contains literature survey and original research work by undersigned candidate, as part of my Master of Computer Application studies.

All information in this document had 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 : Rishi Dey

Examination Roll Number : MCA196014

Registration Number : 124177 of 2013 - 2014

Project Title : Emotion Predicted Through Social Media
Interaction and Graphology

Signature with Date :

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr.Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this project.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Application course.

A special note of thanks goes to Prof. Mahantapas Kundu, Head, Department of Computer Science and Engineering, Jadavpur University.

I am also thankful to Prof.ChiranjibBhattacharjee, Dean, Faculty of Engineering and Technology, for providing an excellent environment for completion of this project.

I am also indebted to my co-researchers Mr. Anupam Baidya, Ms. Abhradita Ghosh, Ms. Debarati Bera and Ms. Chhanda Roy for their seamless co-operation and help in completion of this project. I am thankful to my fellow classmates and my family for constant help and support.

Date:_____

Rishi Dey
Master of Computer Application – III
Examination Roll No. – MCA196014
Registration No:124177 of 2013 – 2014

Contents

1	Introduction	1-2
2	Previous Research Work	3-5
3	Basic Concepts	6-13
3.1	Machine Learning	6
3.2	Natural Language Processing (NLP)	7-10
3.2.1	Brief Introduction	7-8
3.2.2	Text Analytics	8
3.2.3	Computational Linguistics	9
3.2.4	Sentimental Analysis	9
3.2.5	Tokenization and POS Tagging	10
3.3	Rule Based Classifier	10
3.4	CBR	11-12
3.5	Graphology	12-13
4	Methodology	14-25
4.1	Schematic of the System.	14
4.2	Phase-1	15-20
4.2.1	Data Collection	15-16
4.2.2	Data Pre-processing	16-18
4.2.3	Applying Lexical Analysis	18
4.2.4	Algorithm: Phase-1	18-20
4.3	Phase-2	20-23
4.3.1	Signature and Script Collection	20

	4.3.2 Manual Analysis	21
	4.3.3 Applying Rule Based Classifier	21-22
	4.3.4 Applying Lexical Analysis	22-23
4.4	Phase-3	23-25
	4.4.1 Comparison	24
	4.4.2 CBR System	24-25
5	Experimental Configuration	26-28
	5.1 Lexical Techniques	26-28
	5.5.1 Vader Lexicon	26
	5.5.2 SentiWords	26
	5.5.3 SentiWordNet	27
	5.5.4 SenticNet	27-28
	5.2 Machine Configuration	28
6	Results and Performance Analysis	29-34
	6.1 Phase-1	29-31
	6.2 Phase-2	32
	6.3 Phase-3	33-34
6	Conclusion and Future Scope	35
7	Bibliography	36-37

List of Tables

1	Data pre-processing & lexical Analysis of sample message	18
2	Graphological features & lexical analysis of sample signature	23
3	Dissimilarity and mean measurement for participant ID-1	24
4	Sample from Vader Lexicon	26
5	Sample from SentiWords	26
6	Sample from SentiWordNet	27
7	Sample from SenticNet	28
8	Polarity score for each participant in Phase-1	31
9	Polarity score for each participant in Phase-2	32
10	Mean score for each participant in Phase-3	33
11	Dissimilarity score for each participant in Phase-3	34

List of Figures

1	Language Encoding and Decoding	7
2	Train-Test-Evaluate cycle of machine learning	8
3	Typical Corpus Division	8
4	Simple Input-Output model	9
5	CBR life cycle	12
6	Automated system of graphological technique	13
7	Schematic diagram of System	14
8	Schematic diagram of Phase-1	15
9	Schematic diagram of Phase-2	20
10	Signature and Script sample for a participant	21
11	Rule Based Classifier for Signature	22
12	Rule Based Classifier for Script	22
13	Schematic diagram of Phase-3	23
14	Schematic diagram of CBR System	24-25

15	Sentimental polarity score for participant-4	29
16	Sentimental polarity score for all participants	30
17	Most similar & dissimilar of polarity scores in Phase-1	31
18	Most similar & dissimilar of polarity scores in Phase-2	32
19	Dissimilarity and Mean score of LA tools	34

List of Abbreviations

1. LA – Lexical Analysis
2. CBR – Case Based Reasoning
3. POS – Parts of Speech
4. NLP – Natural Language Processing
5. SA – Sentimental Analysis

Chapter 1

Introduction

Social media is a growing source of data and information spread. However, such information is convoluted with varying interests, opinions and emotions. Moreover, the form of communication lacks standardized grammar, spelling, use of slang, sarcasm and abbreviations, and more. These parameters can make extracting critical points, facts, and the sentiment of the message difficult in situations where a number of these aspects are present. With help of natural language processing (NLP) it is possible to study and analyze these messages and objectively classify sentiments presented in social media [1].

Sentiment Analysis (SA) is the task of extracting emotional sentiment with certain pre-defined polarities through analysis of the properties contained within the data. For instance, twitter messages about a local event, or blog posts on some issue, or reviews of some products, may induce SA to classify the emotions expressed in such texts through a polarity spectrum of positive - neutral – negative[3]. This can help in solving many problems and provide various indicators in election results, opinion mining, advertisement designs, health care improvements and a variety of such public domains. Applying mining techniques and sentiment analysis over unstructured data is considered a big challenge in this research area.

In the past decade, sentiment analysis has become a hot research field and a booming industry. For instance, IBM SPSS provides quantitative sentiment summaries of survey data to assist businesses in understanding consumer attitudes. Wall Street has also started to use SA in their trading algorithms with companies like OpFine providing up-to-date sentiment tracking of financial news [3].

Our second topic of interest is Graphology. Graphology is the study of handwriting. It is a scientific method of evaluating and also understanding a person's personality by identifying the strokes and patterns revealed by his/her handwriting. Handwriting is recognized as being unique to each individual, irrespective of the fact whether the person has written using hand, foot or mouth. The main reason is that the handwriting is controlled by the brain [2].

Hence, the colloquial term *handwriting* is also known as “brain writing”. Some scientists in the neuromuscular field of research state that some small neuromuscular movements are associated to the person's personality [2]. Each trait of personality is shown by a neurological brain pattern. A unique neuromuscular

movement is produced by each neurological brain pattern which is similar for every person who has that personality trait [4]. These tiny movements occur unconsciously while writing. Each stroke or written movement reveals a specific personality trait. Graphology is the discipline of identifying these strokes as they appear in handwriting and describe the corresponding personality trait.

Here we use a few methods for analysing real world handwritten text and signature samples with some technological aids. The analysis is done for specific features of the sample for determining various characteristic behavioural traits of the person. Certain parameters of the handwritten sample are considered to determine corresponding traits.

In chapter 2, we introduce some concept of Machine Learning, CBR, Natural Language Processing (NLP), Graphology techniques and definition of other parameters deemed important in the present work. In chapter 3 is described the methodologies used to determine Emotional Polarity based on different tools. In chapter 4, we discuss about the system configuration and tools utilised for our work. Chapter 5 describes the results obtained using various tools and graphology techniques. It also provides summaries on the basis of dataset and characteristic traits. Lastly chapter 6 provides the conclusion drawn on the outcome of the experiments. This chapter also hints at future scopes in this research domain.

Chapter 2

Previous Research Work

In recent years a lot of work has been done in the field of “Sentiment Analysis on Social Media” by number of researchers. Majority of the work have been performed on Twitter. In its early stage it was intended for binary classification which assigns opinions or reviews to bipolar classes such as positive or negative only. They proposed either machine learning approach or lexicon-based approach or they may even include combination of both to achieve good accuracy.

Some of literature reviews are:

□Turney et al [5](2002) used bag-of-words method for sentiment analysis in which the relationships between words was not at all considered and a document is represented as just a collection of words. To determine the sentiment for the whole document, sentiments of every word was determined and those values are united with some aggregation functions.

□Preceded by Pang Lee et. al [6], they classify documents not by topics but by sentiments, e.g. determining whether the review is positive or negative. For negation handling, if a word x follows the negation word then a new feature ‘NOT_x’ created tag every word from x until first punctuation mark. But this method cannot model the scope of negation, because it is heuristically tagging all word until it finds the mark, without concerning with negation words or not. Addition in pre-processing task, mostly the punctuation marks is removed; this is for simplification in pre-processing stage.

□As research in Indonesia, Bojar [7] who conducted research about the resources of the lexicon for Indonesian sentiment also did the negation handling. By adapting the technique from Das and Chen.[8] handled the negation of sentiment caused by a negation word. Bojar uses negation words such as ‘tidak’, ‘tak’, ‘tanpa’, ‘belum’, and ‘kurang’. The words that occur between the negation words and the first punctuation after the negation word are tagged with ‘NOT’. Example, there is a sentence: ‘kameranya kurang bagus gambarnya’ became ‘kameranya kurang NOT bagus NOT gambarnya’.

□Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level

classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004)[9] and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009)[10].

□ Parikh and Movassate(2009) [11] implemented two models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model.

□ For the negation in sentiment, there are some of the researchers that focus on the impact of the negation in sentiment sentences. A survey conducted by Wieghan et. al [12], they survey for negation role in sentiment analysis. They state that effective negation model for sentiment analysis usually requires the knowledge of polar expression. Jia et. al. [13] studied the impact of each occurrence of a negation term in a sentence on its polarity and introduced the concept of scope of the negation term t.

□ Hogenboom et. al.(2011) [14] they state that for English review sentences, the best performing method is considering 2 words following the negation to be negated.

□ Several studies also concern with the scope of negation, Moral Dadvar et. al.(2011)[15], conduct a study dealing with different negation scopes to investigate how it affects the polarity identification of the sentences and assume that opinions are mostly expressed by the use of adjective and adverbs.

□ Xia et al.(2011) [16] used an ensemble framework for Sentiment Classification which is obtained by combining various feature sets and classification techniques. In thier work, they used two types of feature sets (Part-of-speech information and Word-relations) and three base classifiers (Naive Bayes, Maximum Entropy and Support Vector Machines). They applied ensemble approaches like fixed combination, weighted combination and Meta-classifier combination for sentiment classification and obtained better accuracy.

□ In 2014,Calvinet. al.[17]proposed a model where sentiment extremity of Twitter surveys are measured utilizing Naïve Bayes classifier strategy. The model demonstrates a promising come about on characterizing the ubiquity in light of consume satisfaction and along these lines characterizing the best supplier to be utilized.

□ In 2014, Aizhan Bizhanovaet. al. [18] proposed a model for naturally characterizing the opinion of Twitter messages toward item/mark, utilizing emoticons and by enhancing preprocessing steps keeping in mind the end goal to accomplish high exactness.

□ In 2016, Sanjana Woonnaet. al.[19] proposed a framework that examinations tweets from Twitter into three classifications which are positive,

negative and neutral utilizing supervised learning approach After the execution, the outcomes demonstrated which viewpoints individuals like or aversion and how feelings on motion pictures changes over a timeframe.

□ In 2017, Kai Yang et. Al [20] proposed a highly effective hybrid model combining different single models to overcome their weaknesses. They build the sentimental dictionary from exterior data. As single model have many limitations and weakness. That why they build a hybrid model by combing many single approaches to overcome those limitations of single model. The experimental results show that our hybrid model shows very great performance. In hybrid model 2 approaches that are SVM and GDBT (Gradient boosting decision tree) are combined together that are based on stacking approach.

□ Pablo et. al. [21] presented variations of Naive Bayes classifiers for detecting polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Neutral tweets neglected). The features considered by classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters.

Chapter 3

Basic Concepts

3.1 Machine Learning

Machine learning, which is an application of artificial intelligence (AI), provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to enhance its knowledge base. The process of learning begins with data in order to look for patterns and make better decisions in the future based on the data that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly [22].

Machine learning algorithms can be broadly classified into two categories, namely Supervised learning and Unsupervised learning. Supervised learning is a learning in which we train the machine using data (training data) which is well labelled. After that, the trained system is provided with a new set of data (test data) in order to evaluate the correct label of the data. Unsupervised learning is the process of building a system using data that is neither classified nor labelled and allowing the algorithm to act on that data without guidance. Here the job of the machine is to group the data according to similarities, patterns and differences without any prior knowledge of class value.

According to Tom Mitchell, a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

Some areas where machine learning is used are biometric identification, computer vision, game playing, Natural Language Processing (NLP), recommendation system, financial market analysis to name a few.

3.2 Natural Language Processing (NLP)

3.2.1 Brief Introduction

NLP is the study of the computational treatment of natural (human) language. In other words, teaching computers how to understand (and generate) human language.

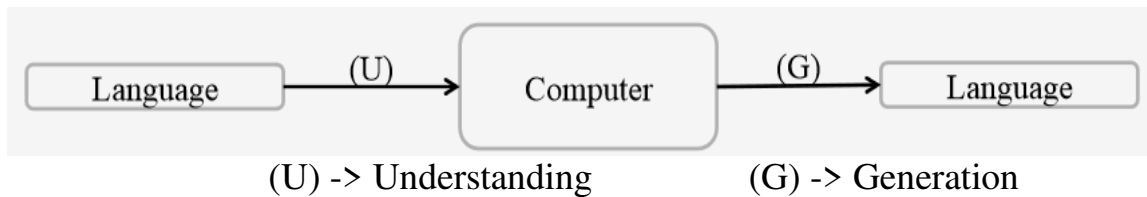


Figure 1: Language Encoding and Decoding

NLP is HARD. Human language is heavily ambiguous. There are types of ambiguity as:

- Morphological: Joe is quite impossible. Joe is quite important.
- Phonetic: Joe's finger got number.
- Part of speech: Joe won the first round.
- Syntactic: Call Joe a taxi.
- Pp Prepositional Phrase attachment: Joe ate pizza with a fork / with meatballs / with Samantha / with pleasure.
- Sense: Joe took the bar exam.
- Modality: Joe may win the lottery.
- Subjectivity: Joe believes that stocks will rise.
- Cc Conjunctive attachment: Joe likes ripe apples and pears.
- Negation: Joe likes his pizza with no cheese and tomatoes.
- Referential: Joe yelled at Mike. He had broken the bike. Joe yelled at Mike. He was angry at him.
- Reflexive: John bought him a present. John bought himself a present.
- Ellipsis and parallelism: Joe gave Mike a beer and Jeremy a glass of wine.
- Metonymy: Boston called and left a message for Joe.

There are a large variety of underlying tasks and machine learning models powering NLP applications. Recently, deep learning approaches have obtained very high performance across many different NLP tasks. These models can often

be trained with a single end-to-end model and do not require traditional, task-specific feature engineering.

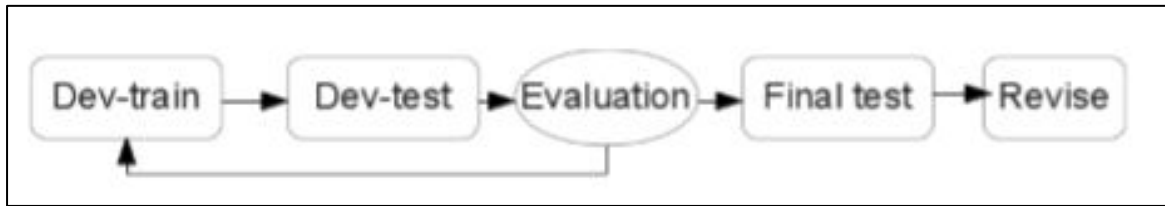


Figure 2: Train-Test-Evaluate cycle of machine learning

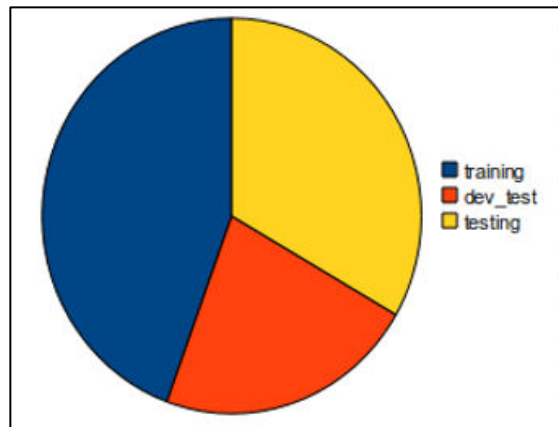


Figure 3: Typical Corpus Division

3.2.2 Text Analytics

Text Analytics, often called Text Mining, is the process of converting unstructured *text* data into meaningful data for analysis. Text Analytics tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management.

On a functional level, Text Analytics systems have four main areas:

- Pre-processing
- Core mining operations
- Presentation layer components
- Refinement techniques

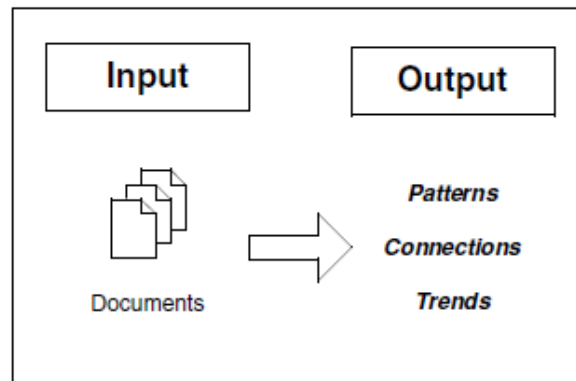


Figure 4: Simple Input-Output model

3.2.3 Computational Linguistics

This is an interdisciplinary field which involves looking at the nature of a language, its morphology, syntax, and dynamic use, and drawing any possible useful models from this observation in order to help machines to handle language.

3.2.4 Sentimental Analysis

Sentiment analysis is the computational study of people's opinions, appraisals, and emotions toward entities, events and their attributes. This refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjectivity in source documents.

Sentiment analysis is a common text categorization task. It is necessary to implement subjectivity analysis at the statement level. Subjective analysis is used to express private states in the context of a text or conversation. Private state is a general covering term for opinions, evaluations, emotions, and speculations. An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs.

Sentiment analysis can be applied on two different levels. Level 1 is the sentence level, which detects positive, negative and neutral sentiment for each sentence. Level 2 is the document level, which detects the whole document sentiment as one unit or one entity positive or negative or neutral [23].

3.2.5 Tokenization and Part of Speech (POS) Tagging

Tokenization and part-of-speech tagging are two fundamental NLP tasks. Tokenization aims at segmenting words from running text while POS tagging uses the recognized words and assigns each word its syntactical category. POS tagging is a process of assigning a part-of-speechmaker to each word in an input text.

Text: The child ate the cake with the fork

Tokens: ["the", "child", "ate", "the", "cake", "with", "the", "fork"]

POS Tags:

The/DT
child/NN
ate/VBD
the/DT
cake/NN
with/IN
the/DT
fork/NN

3.3 Rule Based Classifier

Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification [24]. An IF-THEN rule is an expression of the form

IF *condition* THEN *conclusion*

An example of rule,

IF *age=youth* AND *student =yes* THEN *buys_computer=yes*

The “IF” part (or left side) of a rule is known as the rule *antecedent* or precondition. The “THEN” part (or right side) is the rule *consequent*. In the rule antecedent, the condition consists of one or more attribute tests (e.g., *age=youth* AND *student=yes*) that are logically ANDed. The rule’s consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer). The above rule can also be written as,

$((age=youth) \wedge (student=yes)) \Rightarrow (buys_computer=yes)$

If the condition (i.e., all the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied and that the rule covers the tuple.

3.4 CBR

CBR classifiers treat every problem-solution pair as a case and each such case is stored in a base. An unsolved problem is supplemented with its correct solution which represents its class value. Often, a case base, besides a detailed statement of the problem and its solution, also houses the necessary meta-data required for the problem.

As mentioned in the work by U. Farhan et.al. [25], CBR brings some important advantages to the problem-solving strategy. It can reduce the processing time significantly and also be very useful when domain knowledge is not completely available or not easy to obtain, although extensive knowledge and expertise in the field always helps while modifying the similar solutions to produce a new solution.

Most importantly, potential errors can be avoided and past mistakes rectified in similar cases, while attending to problem at hand. Search time may be reduced by a fool-proof indexing technique.

Thus CBR is an artificial intelligence (AI) technique that considers old cases to take decision for new situations. The old cases constitute past experiences on which one can rely, rather than on rules, during the decision making process. CBR works by recalling similar cases to find solution to new problems [25].

To insert a problem-solution pair in the case-base, at first we search for that particular problem within the existing base with the help of some predefined indexing system. If an exact match is found for the present problem, there is no need for insertion. Otherwise, cases constituting the nearest matches are found, and their class information are collected to provide the new case with a suitable class value. The new case with its new solution is now ready for insertion into the case-base.

The CBR process includes four main steps [25]:

Retrieve

Given a target problem, *retrieve* from memory cases relevant to solving it. For fast retrieval the pre-requisite is an efficient indexing technique.

Reuse

The solution(s) from the retrieved case(s) need to be mapped to the target problem. This may involve adaptation or *reuse* of the solution(s) as needed to fit the new situation.

Revise

Having mapped the retrieved solution(s) to the target situation, the new solution generated has to be tested in the real world (or a simulation) and, if necessary, need to be *revised*.

Retain

After the solution has been successfully adapted and revised for the target problem, *retain* the resulting experience as a new case in memory.

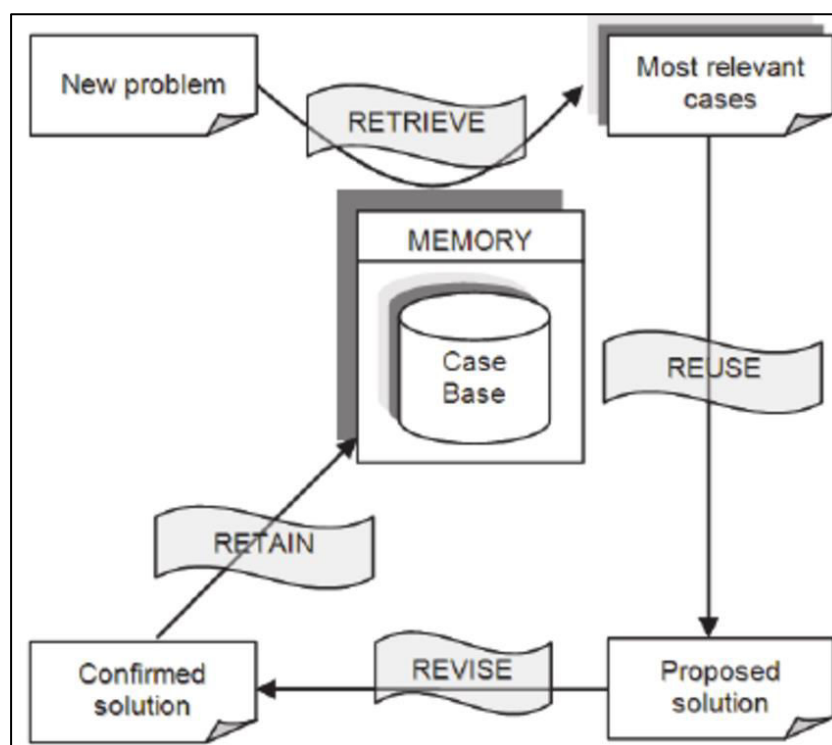


Figure 5: CBR life cycle [26]

3.5 Graphology

The term graphology is an ancient one – dating back to the 17th century. Camillo Baldi, an Italian philosopher, is often referred to as the father of graphology. It was he who covered the topic of graphology in his famous essay “*Trattato Come Da Una Lettera Missiva Si Conoscano La Natura E Qualità Dello Scrittore*” in 1622 [27]. The actual term *Graphology* was coined by Abb Jean - Hippolyte Michon in the year 1897. The term combined two Greek words *graphein* (to write) and *logos* (science) [28]. Over the years it was revealed that handwriting can indicate certain personal traits and behaviours. Scientific researches uncovered the extent of emotional spectrum an individual can let out through handwriting and signature. It is like a window through which one can get

to know the person's intellect, his emotional responsiveness and energy, his defences and fears, his motivation, integrity and imaginative power and his aptitude, even reveal his sex drive and his issues related to trust. Handwriting analysis can lead to the detection of diseases like Parkinson's and cancer.

Signature and handwriting have also been used for identification purpose [29] from time immemorial. But one must remember that age and mental and physical health affect one's writing hand continuously and provision for resulting changes must be kept in an automated system for such analysis. Another factor that should be noted here is that the absence of a particular characteristic in the handwriting does not necessarily reflect in a trait in that person.

An automated system utilising graphological techniques can be depicted as in figure 6 below:

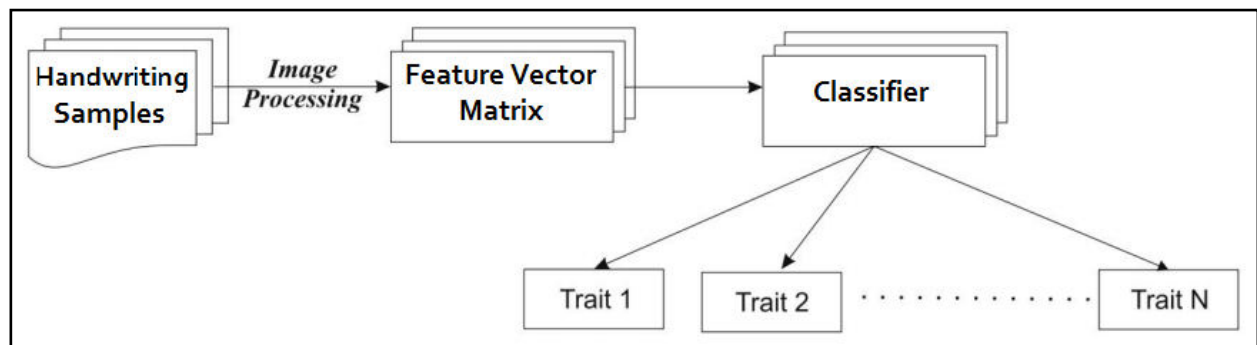


Figure 6: Automated system of graphological technique

Chapter 4

Methodology

4.1 Schematic of the System

In this chapter, we describe the technical details of the work done. In the process we also provide the descriptions of the type of data utilized as well as the probable output from the system. The impact of a project depends majorly on the outcome obtained from experimental results whereas the outcome can be considered as a function of inputs and proper implementation of technologies employed on those input.

We have already discussed in Chapter 2 various techniques available to predict emotional polarity based on the social media interaction messages. Here we use a Lexicon based approach combined with some techniques utilized in Graphology. In the next section we describe our Lexicon based approach.

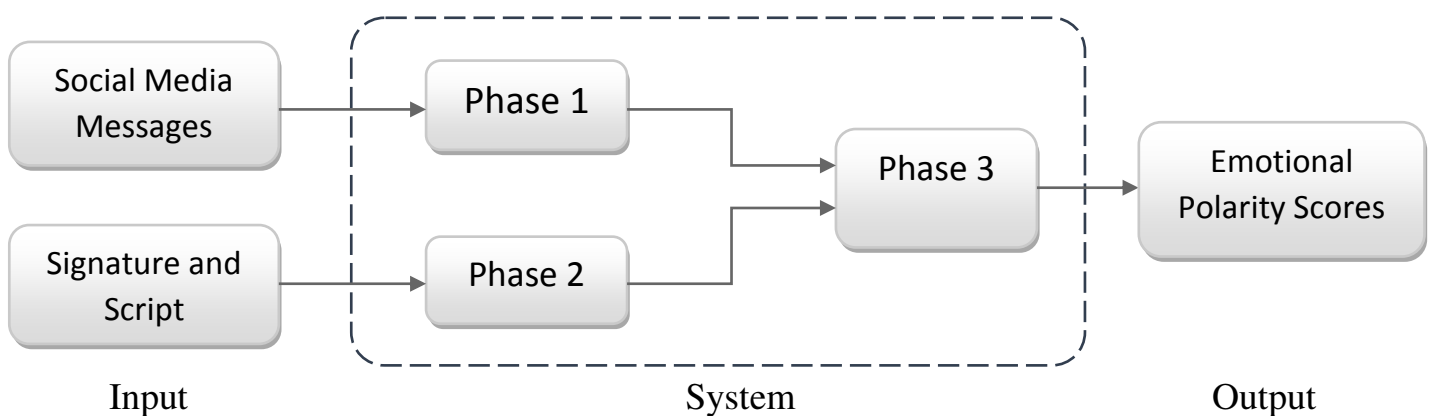


Figure 7: Schematic diagram of System

4.2 Phase-1

We mainly use here a Dictionary Based approach with different types of Lexical Resources on the social media messages.

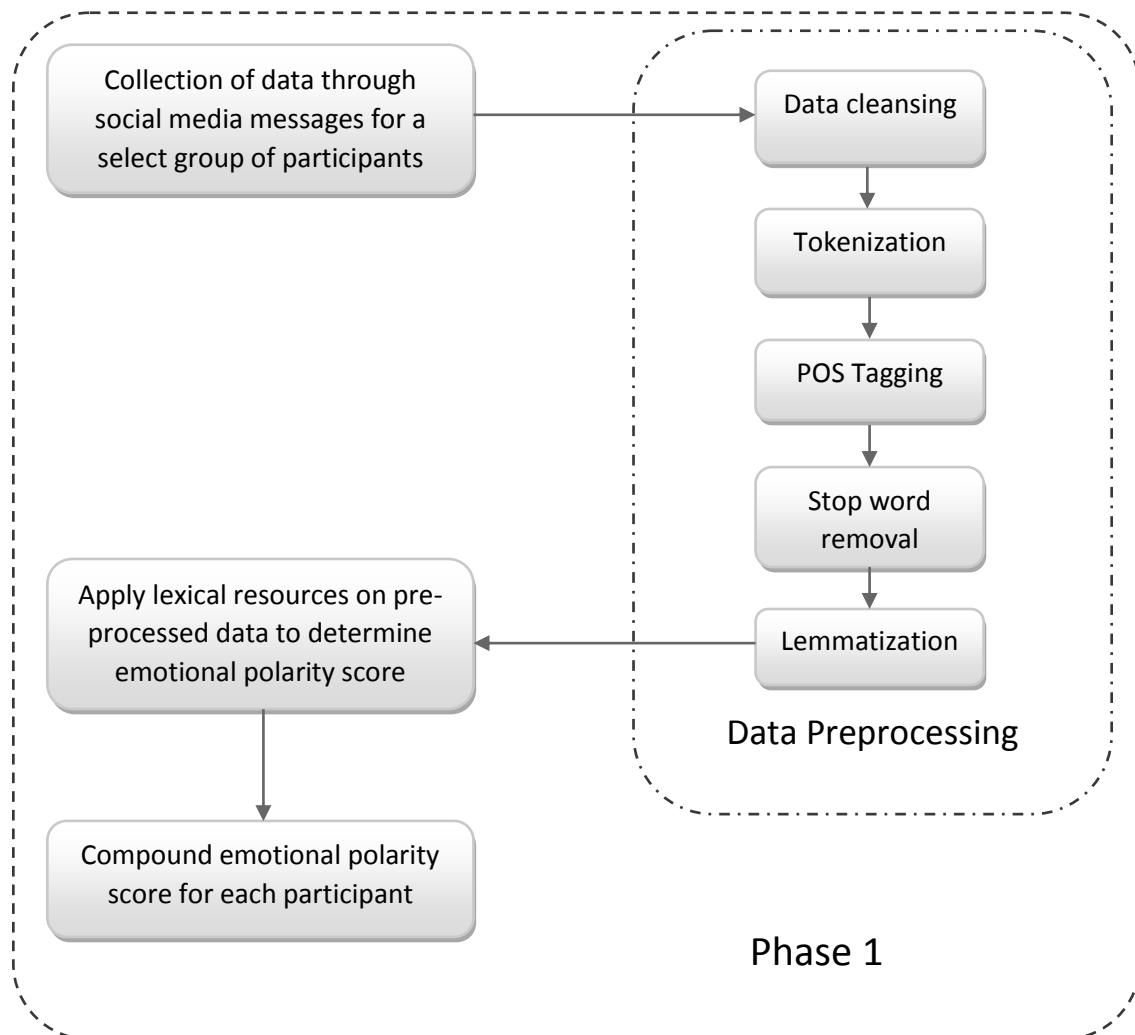


Figure 8: Schematic diagram of Phase-1

4.2.1 Data Collection

Given the private nature of the WhatsApp, this study's first challenge was to create a WhatsApp message dataset while still ensuring users' privacy. At first we created a WhatsApp group and added a few participants into the group. We also collected the participants' general demographic information including their age, gender, place of residence and educational background. The participants were given some particular topics altogether to share their views, opinions and were also allowed to interact with each other to observe emotional trend while discussing on the topic and interacting with fellow participant. The only restriction was that the

whole interaction had to be conducted in English only. However, there were few occasions where the restrictions were overlooked by the participants.

We found it challenging to recruit participants, as some people were quite reluctant to interact only in English, may be due to their inferiority or not having strong grip in English. Nonetheless we did make a concerted effort to find people through word-of-mouth. Through this process we recruited a total of 17 participants, all of them fellow students. We are however aware that the data collection process was biased for younger people, and hope to address this deficit in the future through a different collection process for other age groups too. We further removed 3 participants who did not participate at all in any topic.

Thus, the dataset in this study contains messages from 14 participants of which 4 were female and 10 were male, all being young adults between 20 and 28 years of age. The 14 participants sent a total of 350 messages over an average period of approximately 1 month.

To extract the messages from WhatsApp, we used a option called “Export chat” available in every group of WhatsApp. We extracted the messages omitting all kind of media and stored locally into a text file. This text file is our collected dataset and next it undergoes a process called Data Pre-processing.

4.2.2 Data Pre-processing

Here we perform the necessary data pre processing and cleaning on the collected dataset. It involves several steps, as described below:

- Data Cleansing
- Tokenization
- POS Tagging
- Stop word removal
- Lemmatization

4.2.2.1 Data Cleansing

Data Cleansing is the process of detecting and identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the irrelevant data. Here in our dataset, we removed all emojis and non-English words. Also we detected abbreviated words and tried to replace some commonly used abbreviated words like “etc”, ”sms” by their synonyms or removed otherwise.

4.2.2.2 Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens. The tokens may be words or numbers. Tokenization does this task by locating word boundaries. Tokenization is also known as word segmentation.

4.2.2.3 POS Tagging

In corpus linguistics, POS (Part-Of-Speech) Tagging, also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. All these have evolved from our primary school experience where we are taught to identify words as nouns, verbs, adjectives, adverbs, etc.

4.2.2.4 Stop word removal [30]

Stop words are words which are filtered out before or after processing of natural language data since they usually refers to the most common words in a language. We would not want these words taking up space in our dataset. Using NLTK and its readily available “Stop Word Dictionary”, we removed the stop words as they were not useful. Italicized words in the following context are examples of stop words:

Text:	Alexander was king of Greece and he was a great leader
Tokenization:	[('Alexander', 'was', 'king', 'of', 'Greece', 'and', 'he', 'was', 'a', 'great', 'leader']
POS Tagging:	[('Alexander', 'NNP'), ('was', 'VBD'), ('king', 'VBG'), ('of', 'IN'), ('Greece', 'NN'), ('and', 'CC'), ('he', 'PRP'), ('was', 'VBD'), ('a', 'DT'), ('great', 'JJ'), ('leader', 'NN')]
Stop word removal:	[('Alexander', 'NNP'), ('king', 'VBG'), ('Greece', 'NN'), ('great', 'JJ'), ('leader', 'NN')]

4.2.2.5 Lemmatization

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item. In computational linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Lemmatisation

depends on correctly identifying the intended part of speech and meaning of a word in a sentence.

For example-

am, are, is ⇒ *be*
car, cars, car's, cars' ⇒ *car*

4.2.3 Applying Lexical Analysis

After data pre-processing, we had a set of tokens for each participants. All the used lexical resources were available in text file format. At first, separate dictionary type data structure (as shown in table number 4, 5, 6, and 7) were created for each lexical technique. Given a word, these structures would provide polarity scores of that word according to corresponding techniques. If any word was not found in the lexical resource, then the word was considered to be neutral with polarity score of zero. Polarity scores for each tokenized word were added up and finally divided by the number of tokenized words to produce normalized polarity score, lying between -1 to +1, for each participant.

Table-1: *Data pre-processing & lexical analysis on sample message*

<i>Social media message</i>	<i>Pre-processed data</i>	<i>Lexical Analysis Techniques</i>			
		<i>Vader Lexicon</i>	<i>SentiWords</i>	<i>SentiWordNet</i>	<i>SenticNet</i>
Eligibility criteria for country leaders please ;) let that be the first step	[('Eligibility', 'NNP'), ('criteria', 'NNS'), ('country', 'NN'), ('leaders', 'NNS'), ('please', 'VBP'), ('let', 'VB'), ('first', 'JJ'), ('step', 'NN')]	0.325	0.3493	0.28125	0.47057

4.2.4 Algorithm: Phase-1

Input:

1. Social media messages
2. Lexical resources

Output:

Normalized polarity score for each participant

Main Method:

1. Begin
2. Call Method-1 //Data pre-processing
3. Call Method-2 //Applying LA tools
4. End

Method-1:

1. Begin
2. For each sentence s_i in Social media messages
3. $p_i :=$ sender of s_i
4. $a_i :=$ cleansed data from s_i
5. $b_i :=$ tokenized output from a_i
6. $c_i :=$ output of b_i after POS tagging
7. $d_i :=$ output after removal of stop words from c_i
8. $e_i :=$ lemmatized d_i
9. $case[p_i] := case[p_i] + e_i$ // Case data structure updated
10. End

Method-2:

1. Begin
2. For each lexical resource r
3. Open source text file of r
4. Create dictionary type data structure d_r
5. For each participant p_i
6. Set $len := 0$
7. For each lexical resource r
8. Set $s_r := 0$
9. For each token t_j of $case[p_i]$
10. For each lexical resource r
11. $x := d_r[t_j]$
12. $s_r := s_r + x$
13. $len := len+1$
14. For each lexical resource r
15. $score_r[p_i] := s_r / len$
16. Update $case[p_i]$ with $score[p_i]$ solution value

4.3 Phase-2

We mainly used here graphological methods to extract features from signature and script collected from the same group of participants. The features reflect textual descriptions of the signatures and scripts. A rule based classifier next accepted these textual descriptions as antecedents and produced the corresponding sentiment words as consequent. These sentiment words were processed under the same set of Lexical Resources as applied on Phase-1.

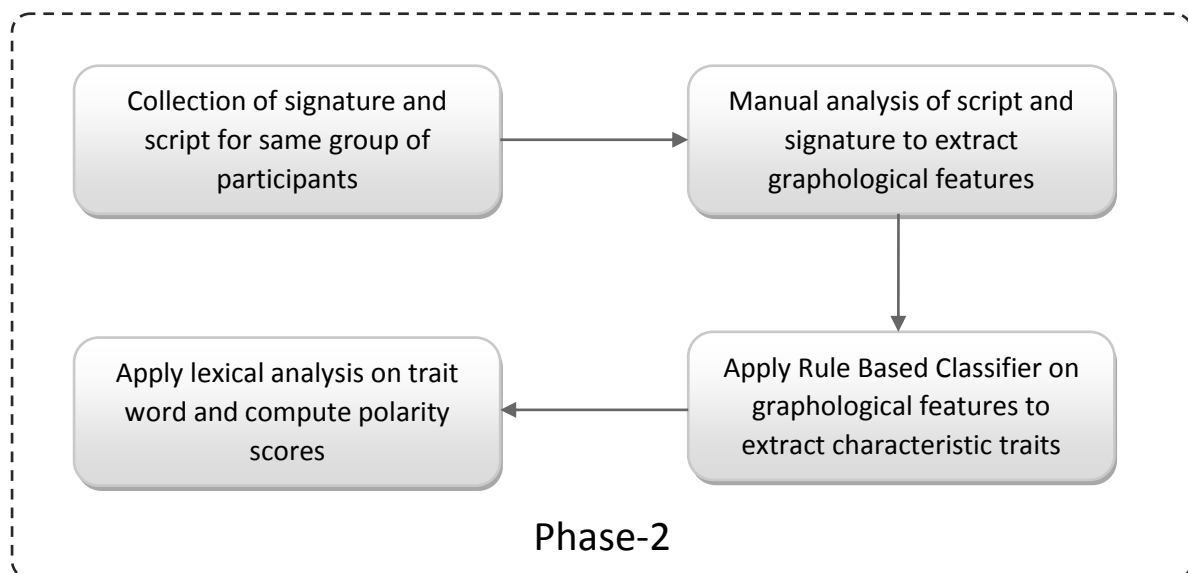


Figure 9: Schematic diagram of Phase-2

4.3.1 Signature and Script Collection

We collected signature and script from the same group of participants who participated in the WhatsApp group, to extract some graphological features like slant, baseline, size, margin, pressure etc. One example each from the collected signatures and scripts follow below:

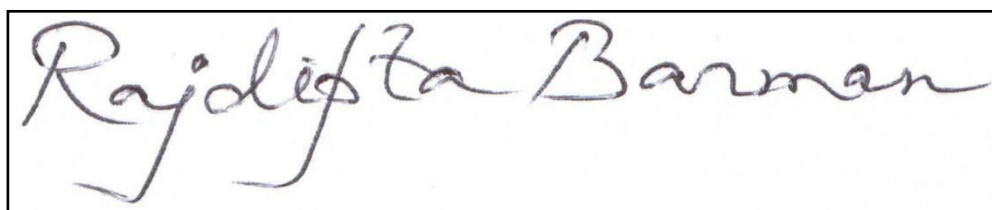
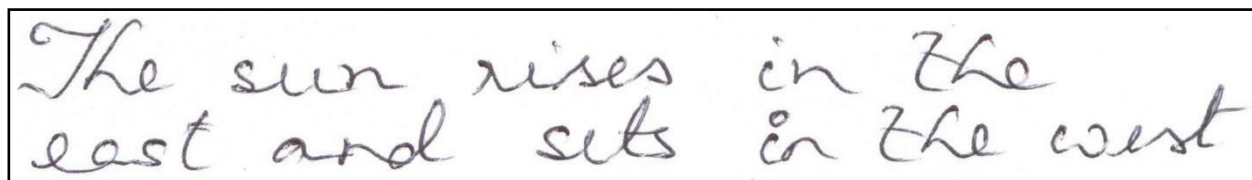
A rectangular box containing the handwritten signature "Rajdip Barman" in cursive script.A rectangular box containing the handwritten sentence "The sun rises in the east and sets in the west" in cursive script.

Figure 10: Signature and Script sample for a participant

4.3.2 Manual Analysis

We have already scanned all the scripts and the signatures, preserved those in the ultimate CBR, detail of which is discussed in Phase-3. But due to paucity of time, we could not apply image processing techniques to extract graphological features of the same. We have done the whole analysis process manually, guided by available support documents [35][34].

4.3.3 Applying Rule Based Classifier

As mentioned above, we have set up the manually extracted graphological features as the antecedents in a rule based classifier. The characteristic traits for each graphological feature obtained as output, form the corresponding consequent part of the classifier. In the subsequent stages these trait words would produce the sentiment polarity score for each participant. The screenshots below depict the rules in two column, the first representing the antecedent and the second consequent.

0.Last Ascending	=>	0.Optimist, 1.Ambitious, 2.Creative,
1.Overall Descending	=>	3.Weak Minded,
2.Overall Straight	=>	4.Balanced Minded,
3.Quite Lengthy	=>	5.Dependable, 6.ShowOff, 7.Adament, 8.Enterprising, 9.Mildly Tedious,
4.Quite Short	=>	10.Responsive, 11.Impatient, 12.Restless,
5.Large Lettered	=>	13.Erratic, 14.Fastidious,
6.Consistently Medium Lettered	=>	15.Submissive, 16.Gentle, 17.Shy,
7.Small Lettered	=>	18.Tightfisted, 19.Logical, 20.Orderly,
8.Tiny Lettered	=>	21.Proud,
9.Big & Few Lettered	=>	22.Dreamy, 23.Independant, 24.Straight Forward, 25.Trusting, 26.Kind,
10.Smoothly Curved End	=>	16.Gentle, 26.Kind,
11.Spikey Lettered	=>	27.Rough, 28.Stubborn, 23.Independant,
12.Letters Stuffy & Crowded	=>	19.Logical, 29.Synchronised, 30.Conservative,
13.Letters Stylized	=>	21.Proud, 31.Crooked, 6.ShowOff,
14.Too Compact Signature	=>	32.Skilled Organizer,
15.Wide Signature	=>	33.Strategist,
16.Extra Pressurized Letters	=>	34.Extrovert, 35.Spendthrift, 36.Overactive,
17.Uneven Impression Of Letters	=>	37.Uncertain, 38.Self-Critic,
18.Underlined Signature	=>	39.Sensitive,
19.Line Above Signature	=>	21.Proud, 40.Over Ambitious,
20.Line Through Signature	=>	14.Fastidious, 38.Self-Critic, 39.Sensitive, 41.Permanently Morose,
21.End Tail Line	=>	42.Skilled, 43.Intolerant Towards Criticism, 44.Intolerant Towards Orders,
22.End Loop	=>	23.Independant, 45.Strong Minded, 46.Recharged, 47.Strongly Determined,
23.End Dot	=>	20.Orderly,
24.Start Dot	=>	48.Organized,
25.Complex & Illegible	=>	49.Make Pretexts,
26.Wavy Signature	=>	15.Submissive, 50.Adjustable, 51.Diplomatic,
27.All Uniform Last Small	=>	52.Good Beginner, 53.Inconsistent,
28.All Uniform Last Big	=>	54.Ends Well, 55.Open Minded,
29.Oscillatory Stroke at start	=>	56.Unimaginative,
30.Oscillatory Stroke at middle	=>	57.Too Romantic,
31.Oscillatory Stroke at end	=>	52.Good Beginner, 53.Inconsistent,

Figure 11: Rule Based Classifier for Signature

0.Signature Larger than Script	=>	0.False Self Esteem, 1.Craving for Recognition,
1.Signature Medium and Equal to Script	=>	2.Sense of Value, 3.Modest, 4.Honest,
2.Signature Smaller than Script	=>	5.High Self Motivation, 6.Low Self-confidence,
3.Signature Totally Illegible	=>	7.Arrogant, 8.Self Importance,
4.Surname Legible	=>	9.Preliminarily Aloof,
5.Firstname Legible	=>	10.Approachable, 11.Direct, 12.Friendly,
6.All Legible	=>	11.Direct, 13.Straight Forward,
7.Both Script & Signature Illegible	=>	14.Enigmatic,
8.Only Script Legible	=>	26.Self Efacng,
9.Only Signature Legible	=>	15.False sense of Self importance,
10.Signature on Left	=>	16.Falsely Nostalgic,
11.Signature in middle	=>	17.Demanding Attention,
12.Signature on right	=>	18.Positive,
13.Signature rising more	=>	19.False Optimism, 20.Balanced, 21.Level Headed,
14.Signature falls more	=>	22.False Pessimism,
15.Right Slant	=>	23.Extrovert, 24.Effervescent,
16.Left Slant	=>	25.Introvert,

Figure 12: Rule Based Classifier for Script

4.3.4 Applying Lexical Analysis

After successfully applying rule based classifier, we extracted set of sentiment words from the traits for each participant, according to the norms of lexical analysis. Next we generated the sentiment polarity scores for each participant from the sentiment words in the same manner as discussed in the Phase-1.

Table-2: Graphological features & lexical analysis of sample signature

<i>Name/Signature</i>	<i>Antecedent</i>	<i>Consequent</i>
Ankit Padia .	Large Lettered	Erratic
		Fastidious
	Extra Pressurized Letter	Extrovert
		Overactive
		Spendthrift
	End Dot	Orderly
	Signature larger than Script	False Self Esteem
		Craving for Recognition
	Signature and Script both legible	Direct
		Straight Forward
Signature in middle	Demanding Attention	

<i>Name/Signature</i>	<i>Lexical Analysis Resources</i>			
	<i>Vader Lexicon</i>	<i>SentiWords</i>	<i>SentiWordNet</i>	<i>SenticNet</i>
Ankit Padia .	0.01	0.041244	0.023065	0.02251

4.4 Phase-3

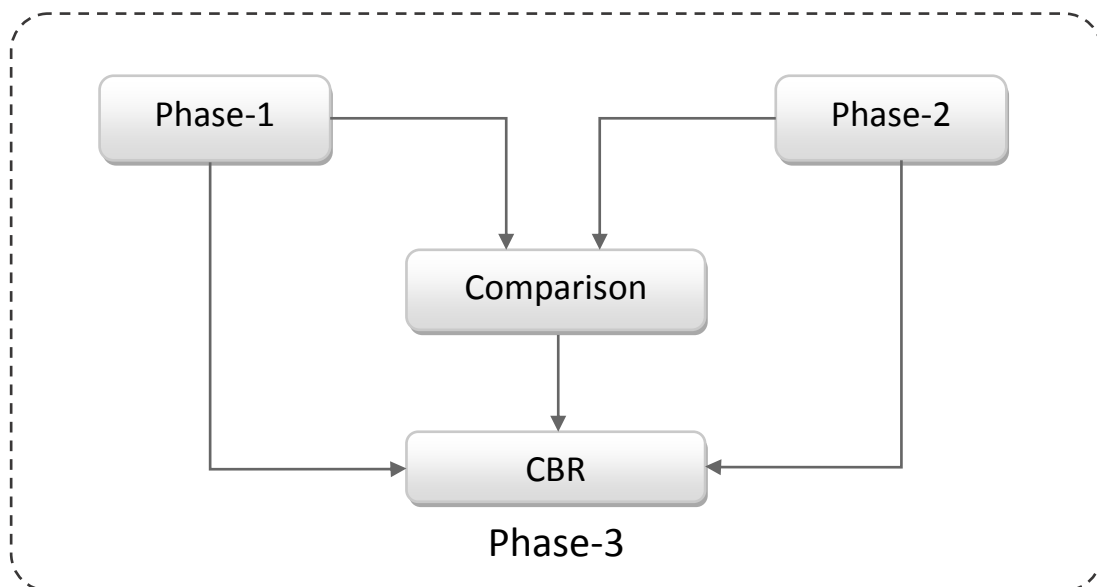


Figure 13: Schematic diagram of Phase-3

4.4.1 Comparison

We analysed the two sentiment polarity scores received from Phase-1 and Phase-2. First we computed the dissimilarity between two scores for each participant. Then the dissimilarity scores (table number 11) were added up to find total dissimilarity for each lexical technique.

We further computed mean value of the two polarity scores for each participant, which constitute the solution part in the CBR system, discussed in our next section.

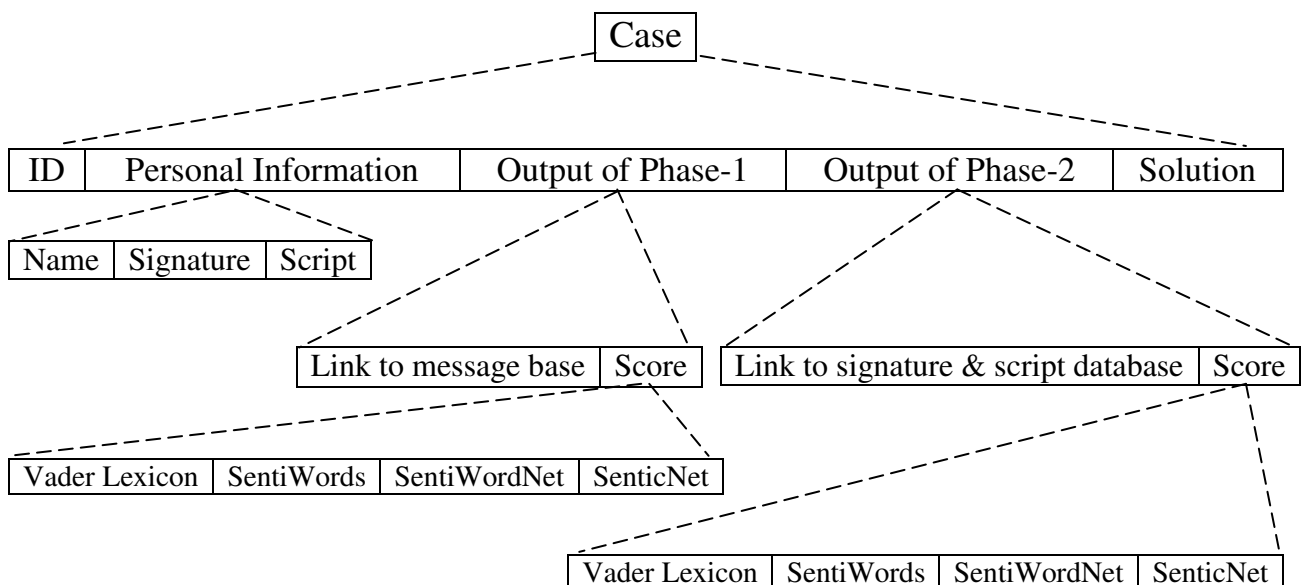
Sample tuples for an individual participant are depicted below in table-3.

Table-3: Dissimilarity and mean measurement for participant ID #1

<i>Measurement</i>	<i>Lexical Techniques</i>			
	<i>Vader Lexicon</i>	<i>SentiWords</i>	<i>SentiWordNet</i>	<i>SenticNet</i>
Social media messages	0.106845	0.0777629	0.052256	0.1352392
Signature & Script	0.065625	0.075922	0.04514	0.0113
<i>Dissimilarity</i>	0.04122	0.001841	0.007116	0.123939
<i>Mean</i>	0.086235	0.076842	0.048698	0.07327

4.4.2 CBR System

We preserved all details of each person as a case in the case-base, the outline structure of which is shown below:



ID 1	Personal Information	Output of Phase-1	Output of Phase-2	Solution
ID 2	Personal Information	Output of Phase-1	Output of Phase-2	Solution
ID 3	Personal Information	Output of Phase-1	Output of Phase-2	Solution
:	:	:	:	:
:	:	:	:	:
ID n	Personal Information	Output of Phase-1	Output of Phase-2	Solution

Figure 14: Schematic diagram of CBR System

The ID of each participant were stored in ID field. The field Personal Information was used to store relevant information about the participants such as name, current signature, and script. The next two fields consisted of two parts each. The first part was used as link to either the message or the signature-script database. The second part in each case corresponded to the polarity scores from the four lexical techniques involved. We had stored all the messages and signatures of a participants to observe the changes in emotion flow during a substantial period of time. The mean value scores, computed in the Comparison section, were used as solution because these reflected the average emotion of a person over the considered time period.

For the time being, we were storing it in case-base so that in future the output of the system can be generated directly from the collective solution parts of the individual cases.

Although the case-base was being used here as a database of collective information and results, we have kept provision for applying reasoning techniques based on similar solution parts clustered together to correlate background information with emotion.

Chapter 5

Experimental Configuration

5.1 Lexical Techniques

We used 4 lexical techniques to compute emotional polarity scores on pre-processed text. The lexical techniques are briefly discussed below:

1. Vader Lexicon:

Vader is a lexicon with both polarity and intensity information attached to each entry [31]. The basic structure is shown below:

Table-4: Sample from Vader Lexicon

<i>Word</i>	<i>Polarity Score</i>	<i>Standard Deviation</i>	<i>Human Evaluation Vector</i>
Accomplish	1.8	0.6	[1, 2, 3, 2, 2, 2, 1, 1, 2, 2]
Danger	-2.2	0.87178	[-1, -1, -2, -4, -2, -3, -3, -2, -2, -2]

The intensity of each word is calculated by averaging human evaluation vector gathered from ten experts' annotation. The experts gave score between -4 to $+4$ to every word where -4 stands for most negative and $+4$ for most positive word.

2. SentiWords:

SentiWords is a high coverage resource containing roughly 1,55,000 English words each associated with a sentiment score between -1 and 1 [32]. Words in this resource are in the form *lemma#POS* and are aligned with *WordNet* lists (that include adjectives, nouns, verbs and adverbs). Scores are learned from SentiWordNet and represent state-of-the-art computation of words' prior polarities (i.e. polarity for non-disambiguated words). Sample entries of SentiWordNet is shown below:

Table-5: Sample from SentiWords

<i>Word</i>	<i>POS</i>	<i>Intensity</i>
Accomplish	V	0.66303
Danger	N	-0.54552

3. SentiWordNet:

SentiWordNet provides users with clusters of synonymous words ready to be used in sentiment analysis tasks [31]. Sample entries of SentiWordNet are shown below:

Table-6: Sample from SentiWordNet

<i>POS</i>	<i>ID</i>	<i>PosScore</i>	<i>NegScore</i>	<i>SysnSet Terms</i>	<i>Gloss</i>
v	02526085	0.125	0.125	accomplish	to gain with effort; "she achieved her goal despite setbacks"
n	14541044	0	0.75	Danger	a cause of pain or injury or loss; "he feared the dangers of traveling by air"

SentiWordNet provides real-value positive score and negative score (+ve Score and -ve Score) for each entry. It also contains not only unigram words, but also multi words expressions (n-grams). SentiWordNet clusters words with similar sentiment orientation together into different sets. For example, “run dry” and “dry out” are in the same set.

4. SenticNet:

SenticNet performs concept-level sentiment analysis that is performing tasks such as polarity detection and emotion recognition by leveraging on semantics and linguistics instead of solely relying on word co-occurrence frequencies [33].

SenticNet provides a set of semantics, sentics, and polarity associated with 100,000 natural language concepts. In particular, semantics are concepts that are most semantically-related to the input concept (i.e., the five concepts that share more semantic features with the input concept), sentics are emotion categorization values expressed in terms of four affective dimensions (Pleasantness, Attention, Sensitivity, and Aptitude) and polarity is floating number between -1 and +1 (where -1 is extreme negativity and +1 is extreme positivity). Sample entries of SenticNet polarity file is shown below:

Table-7: Sample from SenticNet

<i>Concept</i>	<i>Polarity</i>	<i>Intensity</i>
Accomplish	Positive	0.849
Danger	Negative	-0.87

5.2 Machine Configuration:

System: hpTM 15-AY542TU

Processor: Intel® CoreTM i3-6006U CPU @ 2.00GHz

RAM: 4.00 GB

System type: Ubuntu 16.04, 64-bit Operating System, x64-based processor

Chapter 6

Results and Performance Analysis

5.1 Phase-1

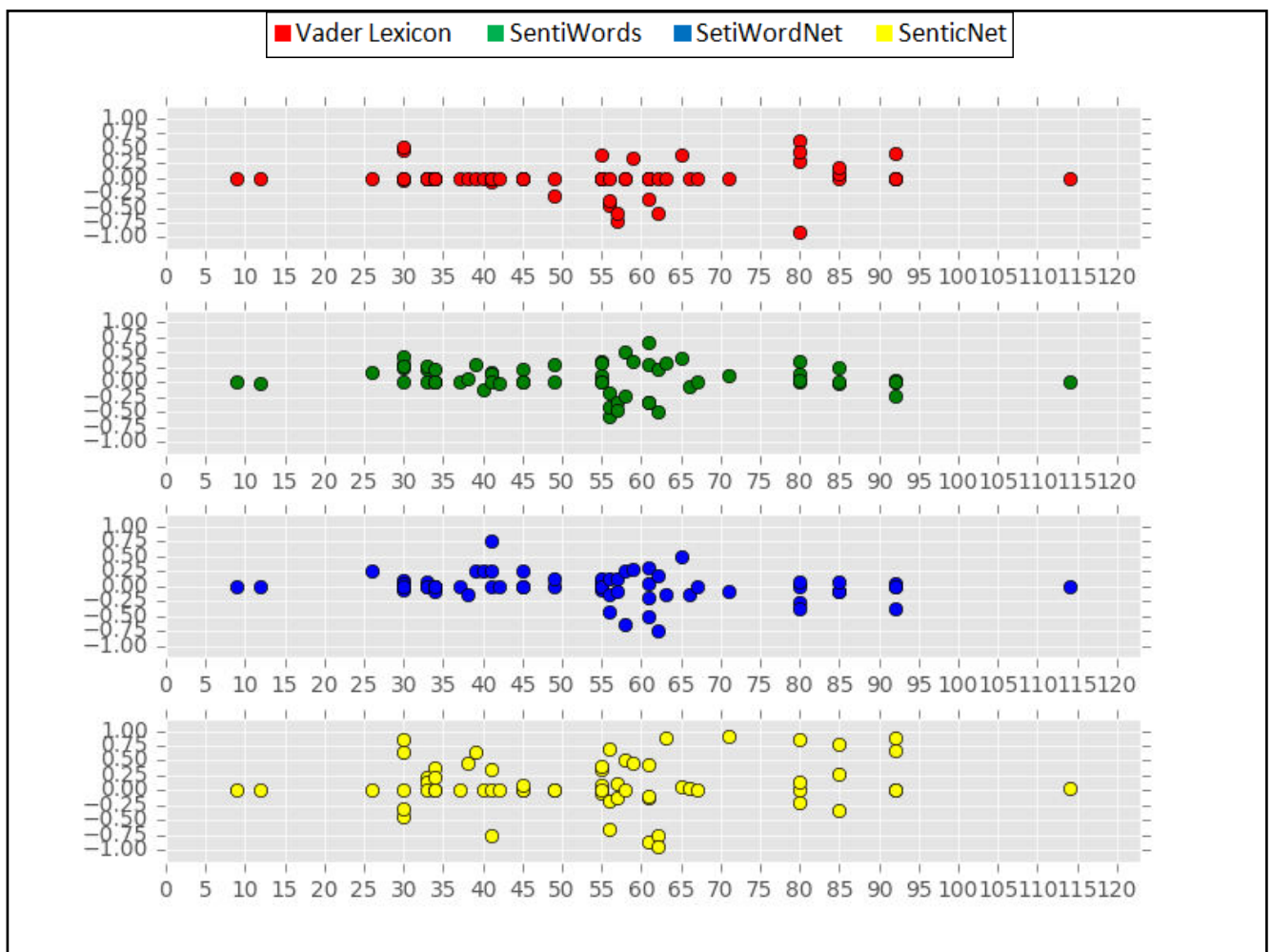


Figure 15: Sentimental polarity score over whole period of time for participant-4

Inference from the above figure 15 is that participant-4 has been most active during a period starting from 25th minute to 95th minute of the total duration.

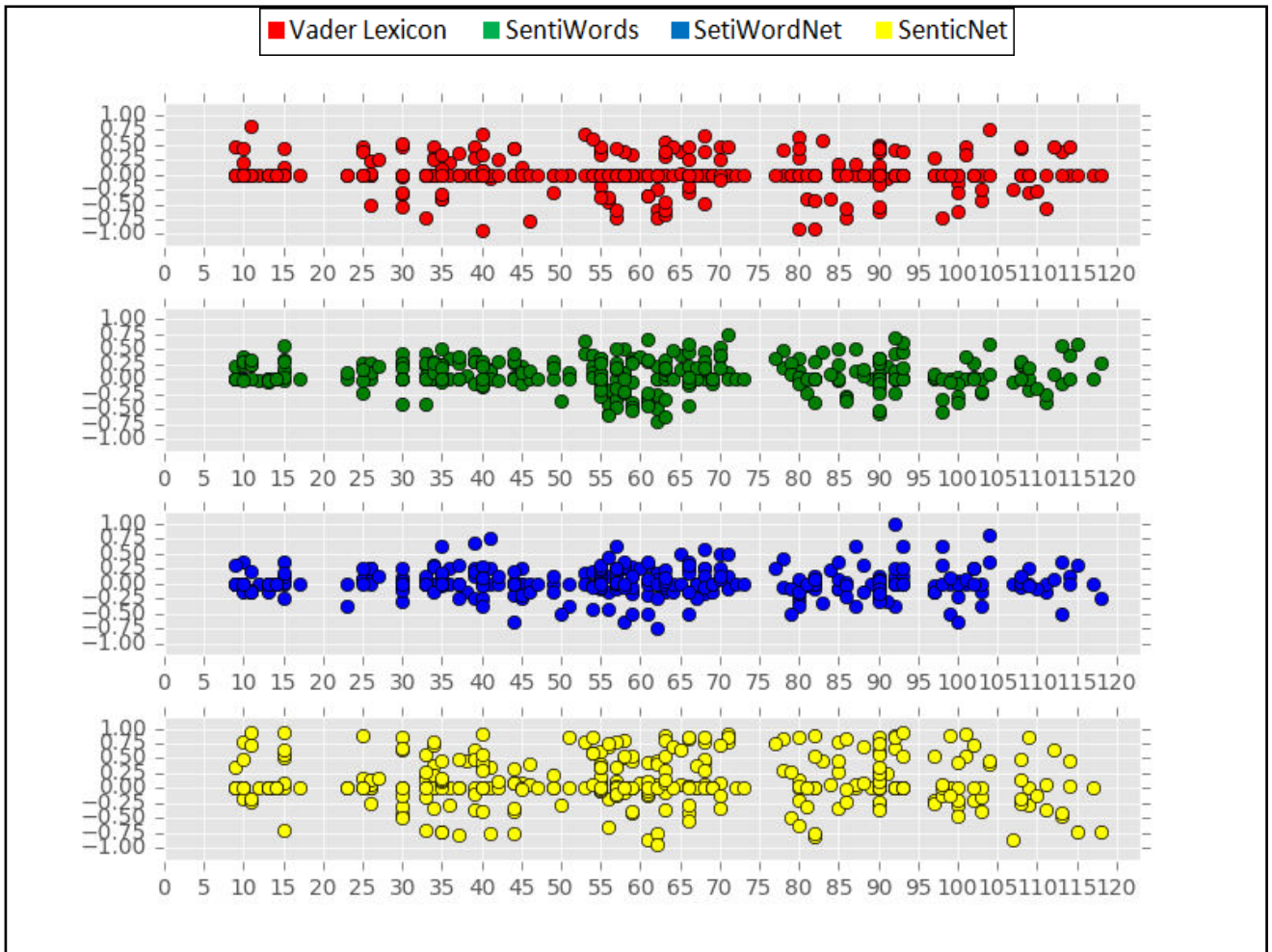


Figure 16: Sentimental polarity score over whole period of time for all participants

Inference from the above figure 16 is that most of the participants have been most active during a period starting from 20th minute to 115th minute of the total duration. Activeness of the participants has increased in the first half and it has reached highest during 50th minute to 75th minute. As the time reaches to end, the activeness of participants decreases.

Table-8: Polarity score for each participant in Phase-1

ID	Name	Lexical Analysis Techniques			
		Vader Lexicon	SentiWords	SentiWordNet	SenticNet
1	Ankan Biswas	0.106845	0.0777629	0.052256	0.1352392
2	Ankit Kumar Mandal	0.1575795	0.0809389	0.0158717	-0.194644
3	Ankit Padia	0.03890	0.091881354	-0.013020837	0.03680896
4	Anupurba Chatterjee	0.24498697	0.20358186	0.10960673	0.2147752
5	Aparna Pradhan	0.20729	0.15845	0.038541	0.000875
6	Kriti Purkait	-0.0875	-0.045610	0.142857	0.1575795
7	Pritam Sharma	-0.1724	-0.1373	-0.00173	-0.0588576
8	Rajdipta Barman	0.0506249	0.0918375	0.0316437	0.281756
9	Richik Chatterjee	0.18675	0.08901	-0.01313	0.08813
10	Rishi Dey	0.04358552	0.12859	0.089037	0.201917
11	Ritam Mondal	0.284	0.10263	0.0553	0.18352
12	Sk Hojayfa Rahaman	0.60625	0.27849	0.141666	0.476875
13	Tanmoy Kumar Das	0.1775	0.1545	0.137083	0.325228
14	Tanusree Das	0.2770833	0.258043	0.161458	0.284

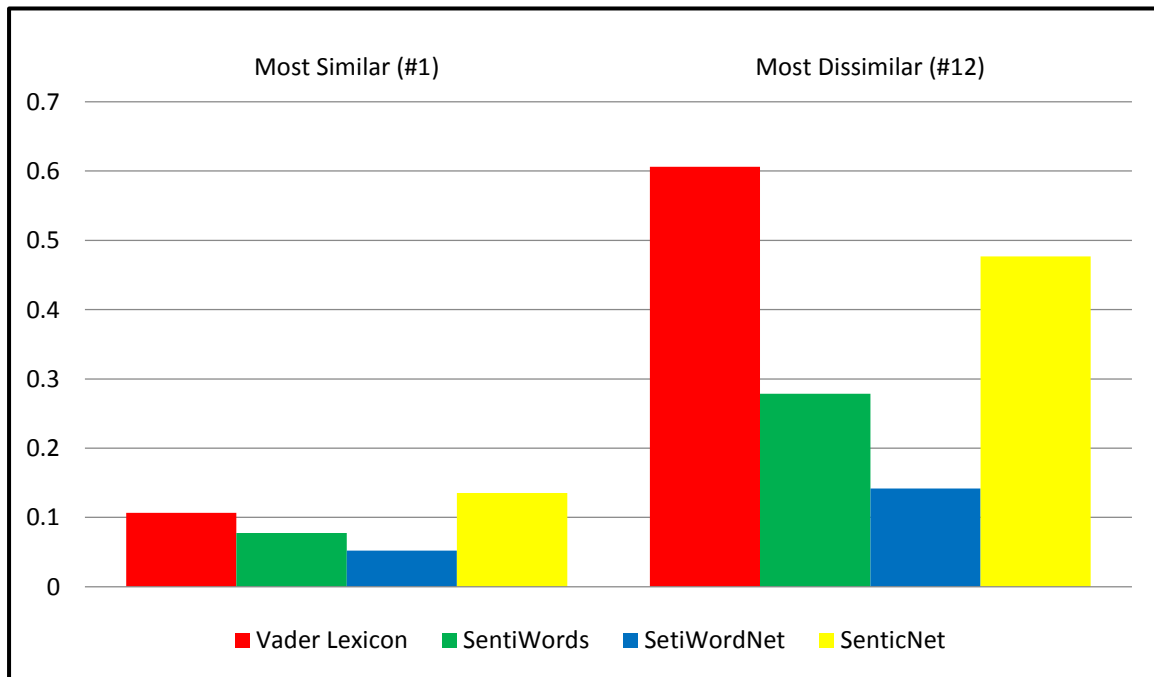


Figure 17: Most similar & dissimilar of polarity scores in Phase-1

5.2 Phase-2

Table-9: Polarity score for each participant in Phase-2

ID	Name	Lexical Analysis Techniques			
		Vader Lexicon	SentiWords	SentiWordNet	SenticNet
1	Ankan Biswas	0.065625	0.075922	0.04514	0.0113
2	Ankit Kumar Mandal	0.0355769	0.0494973	0.03846	0.057938
3	Ankit Padia	0.01	0.041244	0.023065	0.02251
4	Anupurba Chatterjee	0.04583	0.11158	0.040277	0.102091
5	Aparna Pradhan	0.07916	0.0899983	0.023611	0.06254
6	Kriti Purkait	0.07708	0.099253	0.0787	0.11693
7	Pritam Sharma	0.067307	0.070868	0.0272435	0.13191
8	Rajdipta Barman	0.09485	0.07806	0.07965686	0.2249
9	Richik Chatterjee	0.009935	0.0139179	0.01778846	0.059365
10	Rishi Dey	0.06726	0.1051633	0.05952	0.175492
11	Ritam Mondal	0.0604166	0.0626	0.03472	0.137266
12	Sk Hojayfa Rahaman	0.064583	0.06667	-0.02777777	0.02499
13	Tanmoy Kumar das	-0.0666	-0.004145	0.058333	-0.06781
14	Tanusree Das	0.09375	0.12909	0.059895	0.1848

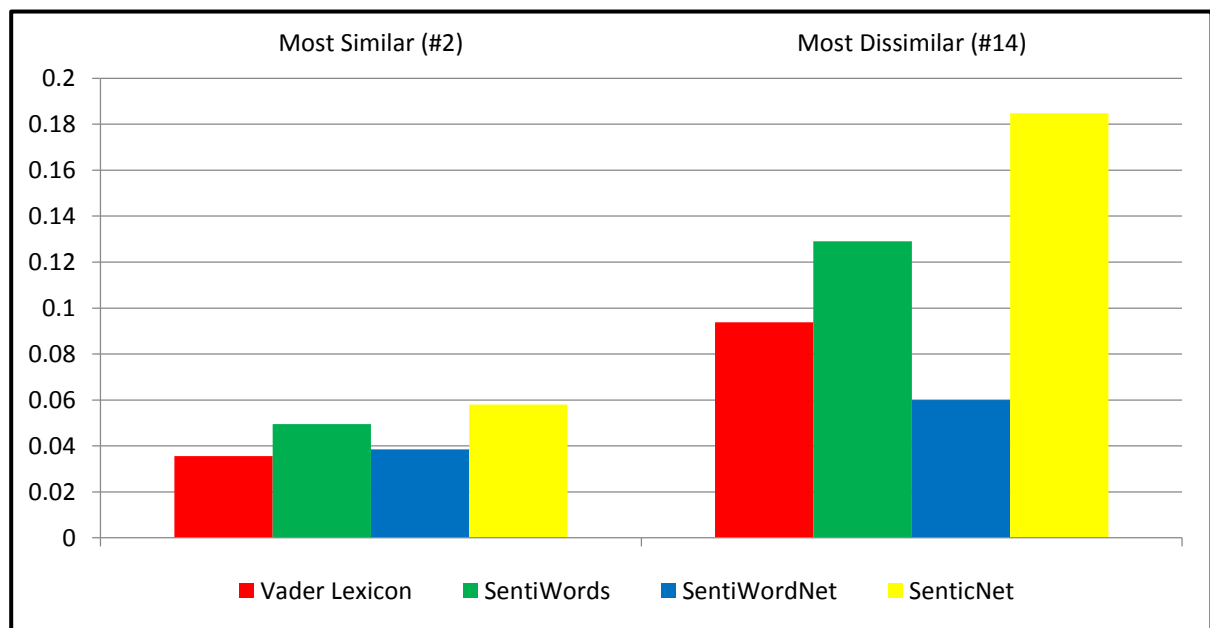


Figure 18: Most similar & dissimilar of polarity scores in Phase-2

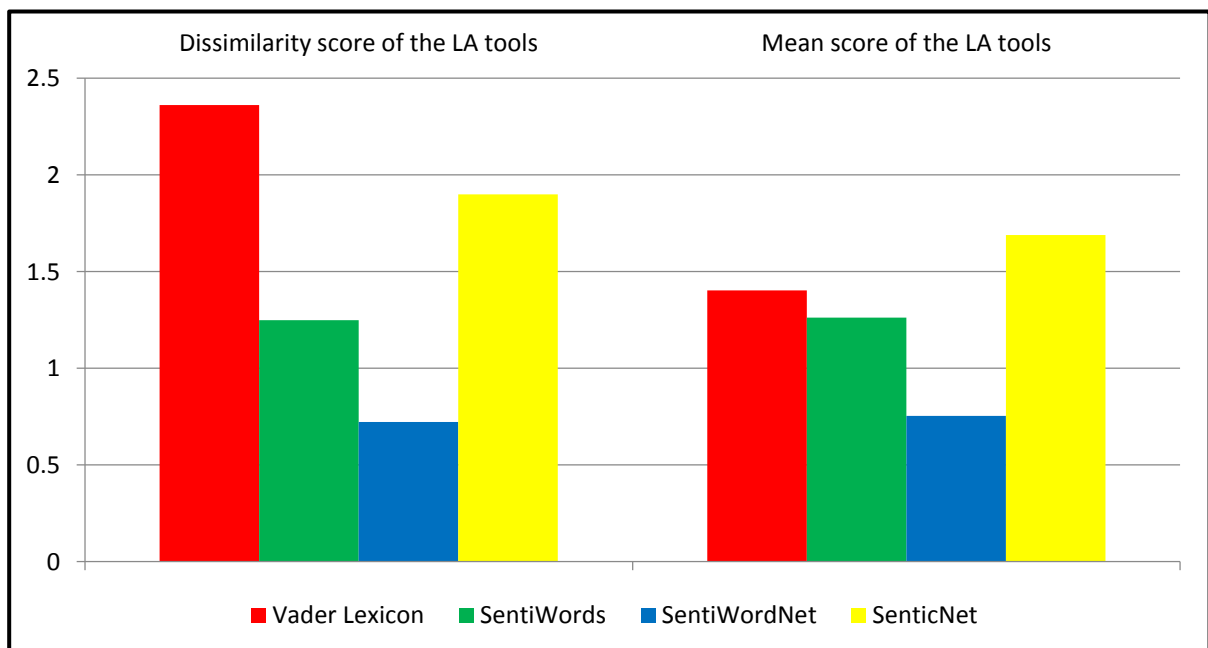
5.3 Phase-3

Table-10: Mean score for each participant in Phase-3

<i>ID</i>	<i>Name</i>	<i>Lexical Analysis Techniques</i>			
		<i>Vader Lexicon</i>	<i>SentiWords</i>	<i>SentiWordNet</i>	<i>SenticNet</i>
1	Ankan Biswas	0.086235	0.076842	0.048698	0.07327
2	Ankit Kumar Mandal	0.096578	0.065218	0.027166	-0.06835
3	Ankit Padia	0.01445	0.066563	0.005022	0.029659
4	Anupurba Chatterjee	0.145408	0.157581	0.074942	0.158433
5	Aparna Pradhan	0.143225	0.124224	0.031076	0.031708
6	Kriti Purkait	-0.00521	0.026822	0.110779	0.137255
7	Pritam Sharma	-0.05255	-0.03322	0.012757	0.036526
8	Rajdipta Barman	0.072737	0.084949	0.05565	0.253328
9	Richik Chatterjee	0.098343	0.051464	0.002329	0.073748
10	Rishi Dey	0.055423	0.116877	0.074279	0.188705
11	Ritam Mondal	0.172208	0.082615	0.04501	0.160393
12	Sk Hojayfa Rahaman	0.335417	0.17258	0.056944	0.250933
13	Tanmoy Kumar das	0.05545	0.075178	0.097708	0.128709
14	Tanusree Das	0.185417	0.193567	0.110677	0.2344
<i>Sum</i>		1.403134	1.261262	0.753036	1.688712

Table-11: Dissimilarity score for each participant in Phase-3

<i>ID</i>	<i>Name</i>	<i>Lexical Analysis Techniques</i>			
		<i>Vader Lexicon</i>	<i>SentiWords</i>	<i>SentiWordNet</i>	<i>SenticNet</i>
1	Ankan Biswas	0.04122	0.001841	0.007116	0.123939
2	Ankit Kumar Mandal	0.122003	0.031442	0.022588	0.252582
3	Ankit Padia	0.0289	0.050637	0.036086	0.014299
4	Anupurba Chatterjee	0.199157	0.092002	0.06933	0.112684
5	Aparna Pradhan	0.12813	0.068452	0.01493	0.061665
6	Kriti Purkait	0.16458	0.144863	0.064157	0.04065
7	Pritam Sharma	0.239707	0.208168	0.028974	0.190768
8	Rajdipta Barman	0.044225	0.013778	0.048013	0.056856
9	Richik Chatterjee	0.223583	0.04003	0.02058	0.046254
10	Rishi Dey	0.023674	0.023427	0.029517	0.026425
11	Ritam Mondal	0.223583	0.04003	0.02058	0.046254
12	Sk Hojayfa Rahaman	0.541667	0.21182	0.169444	0.451885
13	Tanmoy Kumar das	0.2441	0.158645	0.07875	0.393038
14	Tanusree Das	0.183333	0.128953	0.101563	0.0992
<i>Sum</i>		<i>2.361095</i>	<i>1.249149</i>	<i>0.721966</i>	<i>1.899009</i>

**Figure 19:** Dissimilarity and Mean score of LA tools

Chapter 7

Conclusion and Future Scope

According to our study results, among the 4 techniques used here, SentiWordNet has performed the best with least dissimilarity score of *0.721966*. In the runner up position we found SentiWords with dissimilarity score of *1.24914*. Vader Lexicon showed the worst performance with highest dissimilarity of *2.361095*. So we relied mostly on the emotional polarity scores predicted by SentiWordNet and elected it as the solution of our CBR system.

One thing we must mention that SenticNet's semantic analysis was not applied here due to paucity of time, only its bag of words with emotions were used. So comments on performance of SenticNet were slightly biased.

The emotion spectrum used here, was limited to only 3 kind of broad emotion category *Positive, Negative* and *Neutral*. The finer grains of emotion were unexplored in the present work.

Below are listed some of the strategies that we want to apply in future:

1. Predict polarity scores for various kind of emojis
2. Usage of colloquial words
3. Effect of punctuation marks like ! ?
4. Larger datasets with more participants and a variety of subject as well as free flowing conversation
5. Utilize machine learning techniques to extract information from signatures and scripts
6. Evaluation of media files available in the conversation (e.g., images, audio-video clips etc.)
7. Appending CBR with reasoning techniques

In future, the CBR system can be further upgraded to accommodate a psychoanalyst reading on the personality of a person which can be compared with already recorded machine scores, to assess the performance of the system.

Bibliography

- [1] Palsson A, S. D., "An Analysis of Methods and the Impact of Emoticon Removal", 2016
- [2] Seema Kedar, Vaishnavi Nair, Shweta Kulkarni, "Personality Identification through Handwriting Analysis: A Review", January 2015
- [3] Yelena A. Mejova, "Sentiment analysis within and across social media Streams", Spring 2012
- [4] Abdul Rahiman M, Diana Varghese, Manoj Kumar G, "HABIT- Handwriting Analysis Based Individualistic Traits Prediction"
- [5] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", 2002
- [6] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques", 2002.
- [7] Franky, O. Bojar and K. Veselovská, "Resources for Indonesian Sentiment Analysis", 2015
- [8] S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web", 2004
- [9] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter", July 2012
- [10] B. Pang and L. Lee, "Opinion mining and sentiment analysis", 2008
- [11] R. Parikh and M. Movassate, "Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", 2009
- [12] B. Supriyono, "Web Data Mining For Customer's Sentiment Classification For Telkom Speedy Using Twitter In Indonesian" August 2015
- [13] L. Jia, C. Yu and W. Meng, "The effect of negation on sentiment analysis and retrieval Effectiveness", 2009
- [14] A. Hogenboom, P. Van Iterson, B. Heerschop, F. Frasinca and U. Kaymak, "Determining negation scope and strength in sentiment analysis", 2011
- [15] M. Dadvar, C. Hauff and F. D. Jong, "Scope of Negation Detection in Sentiment Analysis", 2011
- [16] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification", 2011
- [17] Calvin and Johan Setiawan, "Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter)", February 2014
- [18] AizhanBizhanova, Osamu Uchida, "Product ReputationTrend Extraction from Twitter", 2014

- [19] Maks Isa, Vossen Piek. “A lexicon model for deep sentiment analysis and opinion mining applications”, 2012
- [20] Dipak Gaikar ,Bijith Marakarkandy“Product Sales Prediction Based on Sentiment Analysis Using Twitter Data”, 2015
- [21] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets”, 2014
- [22] <http://www.expertsystem.com/machine-learning-definition/>
- [23] Khaled Ahmed, Neamat El Tazi, Ahmad Hany Hossny, “Sentiment Analysis Over Social Networks: An Overview”, 2015
- [24] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques”, Third edition
- [25] Uday Farhan, Majid Tolouei-Rad, Adam Osseiran “Indexing and retrieval using case-based reasoning in special purpose machine designs”, 2017
- [26] https://www.researchgate.net/figure/Typical-CBR-life-cycle-comprising-four-stages-Each-of-the-steps-comprising-the-CBR-life_fig1_221916036
- [27] https://en.wikipedia.org/wiki/Camillo_Baldi#cite_note-3
- [28] D. John Antony, O. F. M. Cap,“Personality profile through handwriting analysis”, 2012
- [29] Vikram Kamath, Nikhil Ramaswamy, P. Navin Karanth, Vijay Desai And S. M. Kulkarni, “Development Of An Automated Handwriting Analysis System”, Vol. 6, No. 9, September 2011
- [30] https://en.wikipedia.org/wiki/Stop_words
- [31] Bo Yuan, “Sentiment Analysis Of Twitter Data”, 2016
- [32] <https://hlt-nlp.fbk.eu/technologies/sentiwords>
- [33] <https://sentic.net/about/>
- [34] <http://atozhandwriting.com/signature-analysis/>
- [35] <https://www.anandabazar.com/horoscope/articles/what-does-your-signature-tell-about-you-part-one-dgtl-1.955621>