

EMOTION MINING USING MACHINE LEARNING ALGORITHM

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

GURKAMALPREET KAUR

11507001

Supervisor

ARJINDER SINGH



School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

May 2017

PAC FORM



TOPIC APPROVAL PERFORMANCE

School of Computer Science and Engineering

Program : P172::M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE546

REGULAR/BACKLOG : Regular

GROUP NUMBER : CSERGD0243

Supervisor Name : Arjinder Singh

UID : 20858

Designation : Assistant Professor

Qualification : M.Tech

Research Experience : 6 months

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Gurkamalpreet Kaur	11507001	2015	K1518	8725916196

SPECIALIZATION AREA : Programming-II

Supervisor Signature: 

PROPOSED TOPIC : Emotion Mining using machine learning algorithm

Qualitative Assessment of Proposed Topic by PAC

Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	7.25
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	6.50
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.25
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	6.75
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.50

PAC Committee Members

PAC Member 1 Name: Janpreet Singh	UID: 11266	Recommended (Y/N): Yes
PAC Member 2 Name: Harjeet Kaur	UID: 12427	Recommended (Y/N): Yes
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): Yes
PAC Member 4 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: Emotion Mining using machine learning algorithm

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11011::Dr. Rajeev Sobti

Approval Date: 28 Oct 2016

ABSTRACT

Emotion mining is an area of research to study and analyze the opinions, sentiments, emotions from the written text and classify them into different classes. The main goal of emotion mining is to determine whether the text is subjective or objective in other words classifies the reviews into the positive or negative category. If the text is processed effectively then very useful knowledge can be discovered which will be beneficial for both the users as well as the researchers. The study shows that emotion mining can be used as an interactive way to enhance the academic quality by using the student's feedback. This work focuses on the various features of the academic institute such as teaching, examination, course content, practical work, library facilities and extracurricular activities in order to analyze and detect the positive, negative or neutral score of these factors based on the feedback collected from the students. By classifying the feedback reviews into different categories, one can easily detect and analyze the features where more focus is needed for the improvement. In this work, an approach has been proposed to extract the knowledge from the feedback given by the students to improve the effectiveness of academic activities. This approach concentrates on POS tagging for feature extraction and rule based supervised machine learning technique for classification. There are different types of machine learning algorithms such as Naïve Bayes, Support vector machine, K-nearest neighbor that are used for the classification. Proposed approach has been implemented in python.

Keywords: Emotion mining, opinion mining, sentiment analysis, student feedback, Part of speech tagging, machine learning techniques, natural language processing.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "EMOTION MINING USING MACHINE LEARNING ALGORITHM" in partial fulfillment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Arjinder Singh. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Name of the Candidate

R. No.....

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled “**EMOTION MINING USING MACHINE LEARNING ALGORITHM**”, submitted by **Gurkamalpreet Kaur** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Name of Supervisor)

Date:

Counter Signed by:

1) Concerned HOD:

HoD's Signature: _____

HoD Name: _____

Date: _____

2) Neutral Examiners:

External Examiner

Signature: _____

Name: _____

Affiliation: _____

Date: _____

Internal Examiner

Signature: _____

Name: _____

Date: _____

ACNOWLEDGEMENT

I would like to express my sincere gratitude to concerned people who helped me out to learn this technology. I sincerely thank Mr. Arjinder Singh for his exemplary guidance, monitoring and constant encouragement throughout the dissertation. Without his encouragement and guidance, this research work would not have materialized. I also take this opportunity to express a deep sense of gratitude to university Lovely professional university for their cordial support, valuable information and guidance, which helped me in completing this work through various stages. I am obliged to Faculty members of L.P.U, for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my dissertation 1.

I'm highly grateful to Mr. Dalwinder Singh, Head of Department, for his thorough guidance right from day 1 to end of dissertation 1. He actually laid the ground for conceptual understanding of research work.

My parents receive my deepest love for being the strength in me.

TABLE OF CONTENTS

CONTENTS	PAGE NO.
Title Page	i
PAC Form	ii
Abstract	iii
Declaration Statement.....	iv
Supervisor’s Certificate	v
Acnowledgement	vi
Table Of Contents	vii
list Of Figures	ix
list Of Tables.....	x
CHAPTER1 INTRODUCTION.....	1
1.1 EMOTION MINING	1
1.2 EMOTION MINING TECHNIQUES	2
1.2.1 Unsupervised learning.....	3
1.2.2 Supervised learning	4
1.3 TYPES OF EMOTION MINING.....	9
1.3.1 Document level emotion mining	9
1.3.2 Sentence level emotion mining	9
1.3.3 Feature level emotion mining	9
1.4 PHASES OF EMOTION MINING.....	10

1.4.1 Collection of Data	10
1.4.2 Preprocessing of Data.....	11
1.4.3 Feature extraction	11
1.4.3 Negation handling.....	12
1.4.4 Opinion classification	12
1.5 ORGANIZATION OF REPORT	13
CHAPTER2 LITERATURE REVIEW	14
CHAPTER3 PRESENT WORK.....	28
3.1 PROBLEM FORMULATION	28
3.2 OBJECTIVES	28
3.3 RESEARCH METHODOLOGY	29
3.3.1Collection of data.....	30
3.3.2 Preprocessing of data.....	33
3.3.3Tokenization.....	34
3.3.4Feature Identification.....	34
3.3.5 Feature Extraction.....	35
3.3.6 Opinion Classification.....	37
CHAPTER4 RESULTS AND DISCUSSION	41
4.1 EXPERIMENTAL RESULTS.....	41
4.2 COMPARISON WITH EXISTING TECHNIQUE.	45
CHAPTER5 CONCLUSION AND FUTURE SCOPE	48
5.1 CONCLUSION	48
5.2 FUTURE SCOPE.....	48
REFERENCES.....	49

LIST OF FIGURES

Figure 1.1 Semantic Relation in wordnet.....	3
Figure1.2 Polarity in SentiWordNet	4
Figure 1.3 Linear separable hyperplane	6
Figure 1.4 Non- Linear separable hyperplane.....	7
Figure 1.5 A Decision Tree.....	8
Figure 1.6 General architecture of emotion mining.....	10
Figure 1.7 Emotion mining Process.....	13
Figure 3.1 Work flow diagram.....	29
Figure 3.2 Google Survey form	31
Figure 3.3 Dataset	32
Figure 3.4 Stop word list.....	33
Figure 3.5 Emoticon list.....	35
Figure 3.6 list of sentiment words.....	37
Figure 3.7 Positive response file for training of the classifier	38
Figure 3.8 Negative response file for training of the classifier.....	39
Figure 3.9 Neutral response file for training of the classifier	39
Figure 4.1 Output of the proposed approach.....	41
Figure 4.2 Classification of different features	42
Figure 4.3 Performance comparison of different features using KNN classifier.....	44
Figure 4.4 Overall Opinion classification	44
Figure 4.5 Output of existing base paper technique.....	45
Figure 4.6 Performance comparison of proposed technique with existing technique	46
Figure 4.7 Overall performance Comparison	47

LIST OF TABLES

Table 3.1 Conversion of POS tags to SentiWordNet tags.....	34
Table 3.2 Sequence of 2 tags.....	36
Table 3.3 Sequence of 3 tags.....	36
Table 4.1 Classification of different features	42
Table 4.2 Performance comparison of different features using KNN classifier	43
Table 4.3 Performance comparison of the base paper technique	46

Chapter1

INTRODUCTION

This chapter deals with the terms and terminologies that are used in emotion mining. This chapter includes the background of emotion mining, its techniques and the process of the emotion mining.

1.1 EMOTION MINING: Emotion mining is the method of studying or detecting the viewpoints or perspective of the writer from any text and classifies the text into different categories. Emotion mining has achieved a huge amount of significance with the advancement in the web technology because the bulk of unstructured data also been produced from reviews, blogs, websites etc. If this data is processed effectively then very useful knowledge can be discovered which will be beneficial for both the users as well as the researchers.

In the recent past, emotion mining techniques have become very popular for information extraction and analysis of the data. The main goal of emotion mining is to determine whether the text is subjective or objective in other words classifies the reviews into the positive or negative category. Emotion mining has been utilized in many areas such as marketing, elections, e-commerce, education, movie reviews, hotel reviews etc. Emotion mining is a sub domain of text mining which focuses on extracting the useful knowledge from huge amount of content.

Opinion or Emotion mining allude to identify and analyze the opinion or viewpoint of the internet user expressed toward a particular topic. Analyses of viewpoint provide useful information regarding people interest to the business analyst and other interested parties. There are various machine learning techniques available for emotion mining for e.g. naïve Bayes, support vector machine (SVM), K-nearest neighbor (KNN).

Everyday huge amount of data is being generated over the web. By analyzing the text, interesting patterns of human behavior can be determined. As a result, emotion mining has become a necessary task. Recently, Social media is enormously used in exhibiting personal

opinion on any object that can be any product, service, news, issue or any celebrity. Social media provides a platform for expressing personal opinions where traditional data collecting techniques like surveys are more time-consuming. Opinion can be positive or negative. In order to understand the satisfaction of a person regarding any object, opinions play a major role. Accurate opinion mining can help in decision making.

Emotion mining can also be used in the education to extract the useful information from the student feedback. In order to escalate the overall performance of the academic institute, there are different parameters that are taken into consideration such as examination, course content, teaching, practical work, library facilities and extracurricular activities. Student's feedback or review on these parameters is very important to increase the academic quality. Positive feedback represents that students have good experience with the institution and negative feedback represents that students have a bad experience with the institution. Thus there is a need to conduct the surveys upon these parameters. It is very significant to interpret and analyze the patterns originated from the student feedback data to enhance the teaching, learning as well as overall performance of the institution. On the basis of student feedback, management can take steps to remove the issues specified by the students.

Questions in the survey form can be of two types: qualitative or quantitative[1]. Quantitative data is collected by closed ended questions such as multiple choice questions and qualitative data is collected by open-ended questions such as reviews or comments provided by students in text form. Qualitative data provides more detailed information with insights because students are free to give the feedback in textual format. Although qualitative data is rich in information but instructors usually needs to struggle to infer knowledge from it. Analyzing the qualitative data helps to understand the student feedback on academic curriculum more precisely.

1.2 EMOTION MINING TECHNIQUES: There are two main techniques for the emotion mining:

- i. Unsupervised machine learning
- ii. Supervised machine learning

1.2.1 Unsupervised learning: It is also called lexicon-based approach. It is a classic approach for opinion mining to classify the lexicons into positive, negative and neutral words. For example, pretty has positive polarity and horrible has negative polarity. There is various lexicon based methods to classify the emotions.

- i. WorldNet
- ii. Sentiwordnet
- iii. English subjectivity lexicons

1.2.1.1 WorldNet: It is a largest online lexical database which contains English words, nouns, pronouns, verbs, adverbs, adjectives along with their set of synonyms. WorldNet has more than 118,000 unique words with different word senses. Each category of word is organized in a semantic network of words i.e. synonymy, antonymy, hyponymy, meronymy, troponomy, entailment. [2]

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
<i>Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs</i>		

Figure 1.1 Semantic Relation in wordnet

1.2.1.2 SentiWordNet: - It is also a lexicon resource which is related with three polarity scores positive, negative, objective and classifies the data into PN-polarity or SO-polarity. It can use for all parts of speech i.e. adjectives, noun, pronoun, verb, adverb and associates the polarity to words according to the sense rather than terms. Polarity scores can be positive, negative. For example, the word “healthy” can have the scores of polarity as:

Positive=0.6, Negative= 0.0, Neutral= 0.4 (sense1 for healthy economy)

Positive=0.85, Negative= 0.0, Neutral= 0.15 (sense2 for a good health)

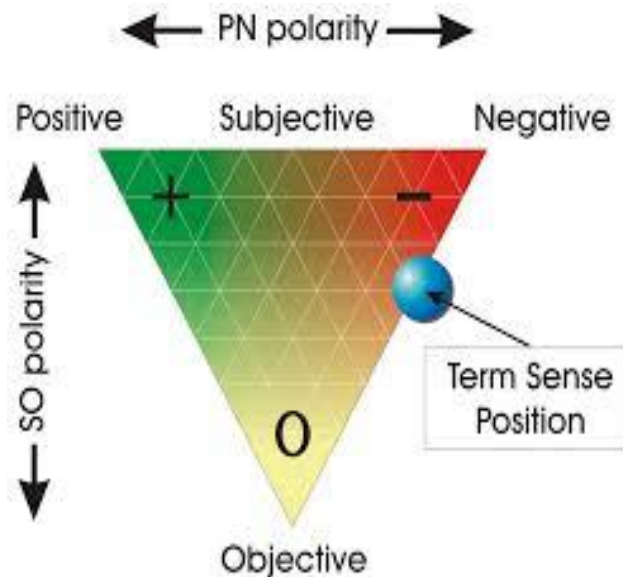


Figure1.2 Polarity in SentiWordNet

1.2.1.3 English Subjectivity lexicons: lexicons are a list of words that represents the subjectivity of the text. Subjectivity lexicons consist of 8221 words and each word has allocated the polarities along with 4 levels i.e. positive, negative, neutral and both.

1.2.2 Supervised learning: There are various machine learning classification techniques used for emotion Classification[3]. Following are the two main steps in machine learning techniques. The first step is to construct a crisp model to distribute the class labels of the training set (with known class labels) and the second step is to use the

resulting classifier to predict the class labels to the testing set (unknown class labels).

Supervised learning techniques are as follows:

- i. Naive Bayes
- ii. Support Vector Machine
- iii. K- nearest neighbor
- iv. Decision tree classifier

1.2.2.1 Naive Bayes: Naive Bayes classifiers are based on Bayesian networks. Bayesian networks are a graphical model to predict the probability of relationship among different variables and are comprise of directed acyclic graph with only one parent node and multiple child nodes. Assure that child node is independent of their parent nodes. Naive Bayes is defined by a formula:

$$R = \frac{P(m|Y)}{P(n|Y)} = \frac{P(m)P(Y|m)}{P(n)P(Y|n)} = \frac{P(m)\pi P(Yr|m)}{P(n)\pi P(Yr|n)}$$

After the comparison of two probabilities, the highest probability predicts the class level value of tuple Y belongs to m if and only if $R > 1$. If $R < 1$ then class label value of tuple Y belongs to n. Merits of using the Naive Bayes classifier is its less computational time for training. Its product form can also convert into a sum by using logarithm.

The major drawback of naive Bayes is the assumption that all the child nodes are independent because of this naive Bayes classifier has less accuracy than other machine learning classifiers. In order to solve this problem, extra edges are added to incorporate some dependencies between the variables.

1.2.2.2 K-nearest neighbor: KNN classifier is based on the concept of similarity between the test tuples and training tuples. Test tuples within the dataset assigned a class label based on the closeness with training tuples having common attributes. The similarity between the two instances are defined by using the Euclidean distance

$$\text{Dist. } (y_1 - y_2) = \sqrt{\sum_{i=1}^n (y_{1i} - y_{2i})^2}$$

Here y_1 and y_2 are two tuples having attributes $y_1 = (y_{11}, y_{12}, \dots, y_{1n})$ and $y_2 = (y_{21}, y_{22}, \dots, y_{2n})$

1.2.2.3 Support vector machine: Support vector machine is a machine learning algorithm used to classify both linear and nonlinear data. SVM includes the concept of margin on the either side of a hyperplane that is used to separate two data classes. Hyperplane with larger margin is more accurate for classifying the data than hyperplane with a smaller margin. Therefore, SVM always looks around the hyperplane with a maximum margin this is called maximum marginal hyperplane (MMH).

In the case of linearly separable data after finding the optimized separating hyperplane, data points that are on its margin called support vector points and consider the linear combination of these points for solution discards the remaining points. Let D be the data set given as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x_i is the set of training tuples with associated class labels y_i . Each class can have one of the two values either +1 or -1.

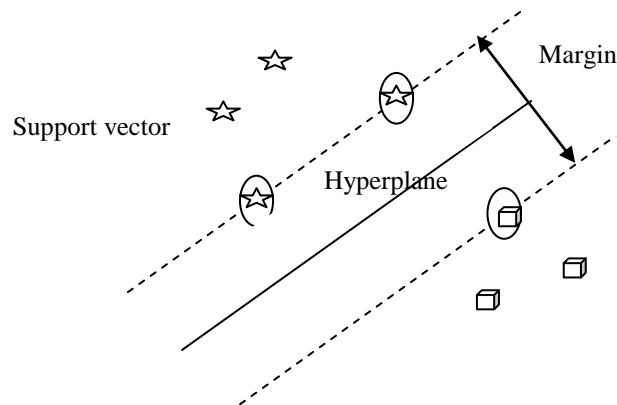


Figure 1.3 Linear separable hyperplane

In the case of nonlinearly separable data, there is no straight hyperplane exists to separate the classes. The benefit of using SVM is that it can also find the nonlinear decision boundaries. SVM uses nonlinear mapping to convert the input data to higher dimensional and then define a separating hyperplane. Training tuples depend on the dot product, $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ where $\Phi(x)$ for non-linear mapping of data to some other dimensions.

K is kernel function that allows dot product calculated directly in feature space. Once the maximal separating hyperplane gets created K will map new points to feature space for classification.

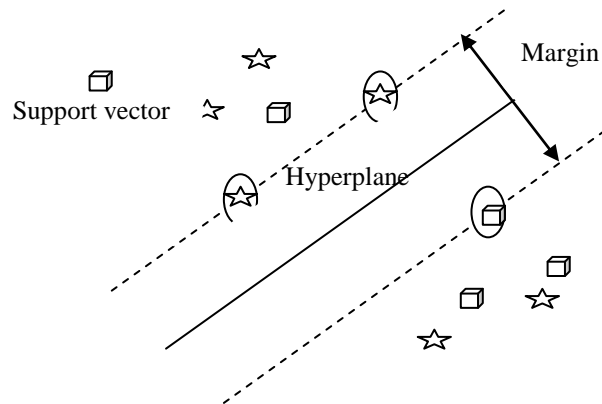


Figure 1.4 Non- Linear separable hyperplane

1.2.2.4 Decision tree classifier: Decision tree is one of the supervised learning classifiers. It is a tree-like structure in which internal nodes indicate the conditions on the attribute and arcs denote the results of the conditions and leaf node defines the class label of the tuple. A complete path from the root node to the leaf nodes predicts the class for that tuple. Decision tree classifier has applications in various areas like prediction analysis, commercial industry, medical and financial analysis.

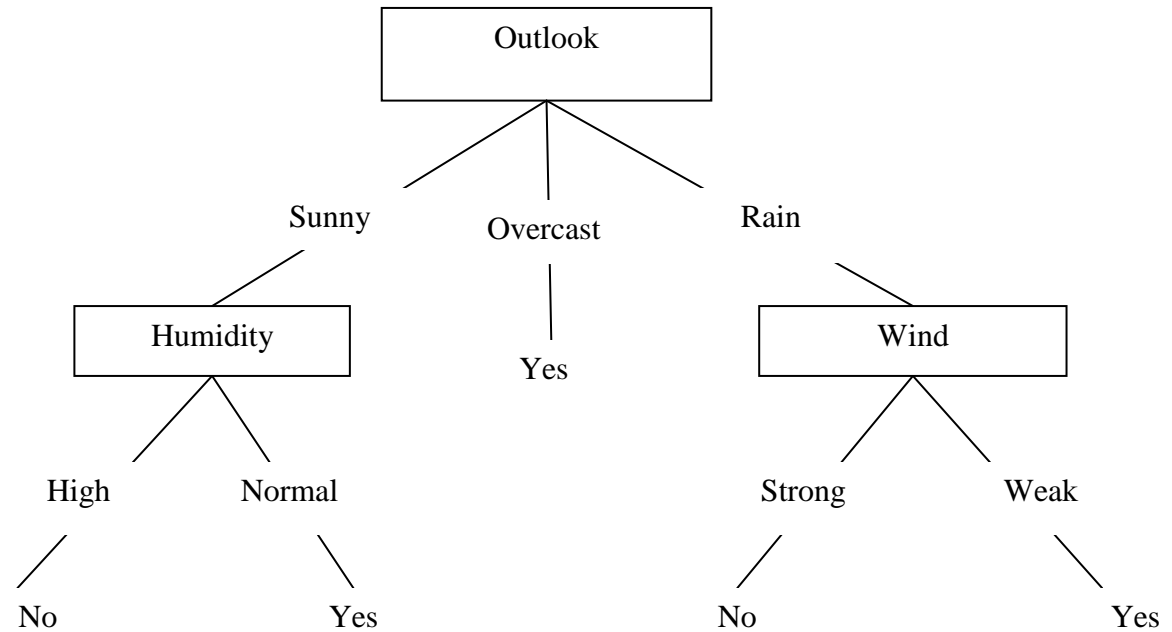


Figure 1.5 A Decision Tree

Attribute Selection measures: In a decision tree, the main focus is to split the training set recursively into further sub parts until there will be no further split. Attribute selection measure is a method of best splitting the data of training set into independent classes. There are three main attribute selection measures:

Info gain: In this measure, attribute having the maximum value of info gain is selected as splitting attribute.

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

Info (D) is the average value of information required to associate the class label to a tuple in Dataset D.

$$info_A(D) = \sum_{j=1}^v (|D_j|/|D|) * info(D_j)$$

Info_A(D) is the expected amount of information needed to identify the class label of a tuple in D.

$$Gain(A) = info(D) - info_A(D)$$

Finally, info gain is calculated by the difference between the actual information requirement and expected new requirement.

1.3 TYPES OF EMOTION MINING: There are three main types of emotion mining those are as follows:

- i. Document level emotion mining
- ii. Sentence level emotion mining
- iii. Feature level emotion mining

1.3.1 Document level emotion mining: It is one of the easiest types of emotion mining. In this type, it is assumed that whole document contains only single opinion[4]. Thus the whole document is used to analyze the subjectivity whether it is of positive, negative or neutral polarity. It provides brief information in terms of a total number of positive or negative documents. In document level, opinion words are extracted from the whole document and analyze the polarity of the document.

1.3.2 Sentence level emotion mining: A document may contain multiple opinionated sentences thus sentence level emotion mining is used to classify the complete sentence into positive, negative or neutral polarity. In this type, it is assumed that each and every sentence contains a single opinion. As single document can have several opinions regarding the same entity thus sentence level emotion mining is used to get more detailed information about the different views conveyed in a document.

1.3.3 Feature level emotion mining: Now a days, feature based emotion mining is becoming more popular because people are more focusing on the specific features of a particular entity than the complete entity. Every entity has multiple features like speed, quality, price, size etc. and different people have different opinions about these features. In this type, the sentence is further broken into nouns, pronouns, adjectives etc. and all the features of a sentence is analyzed and classification is performed based on relevant features.

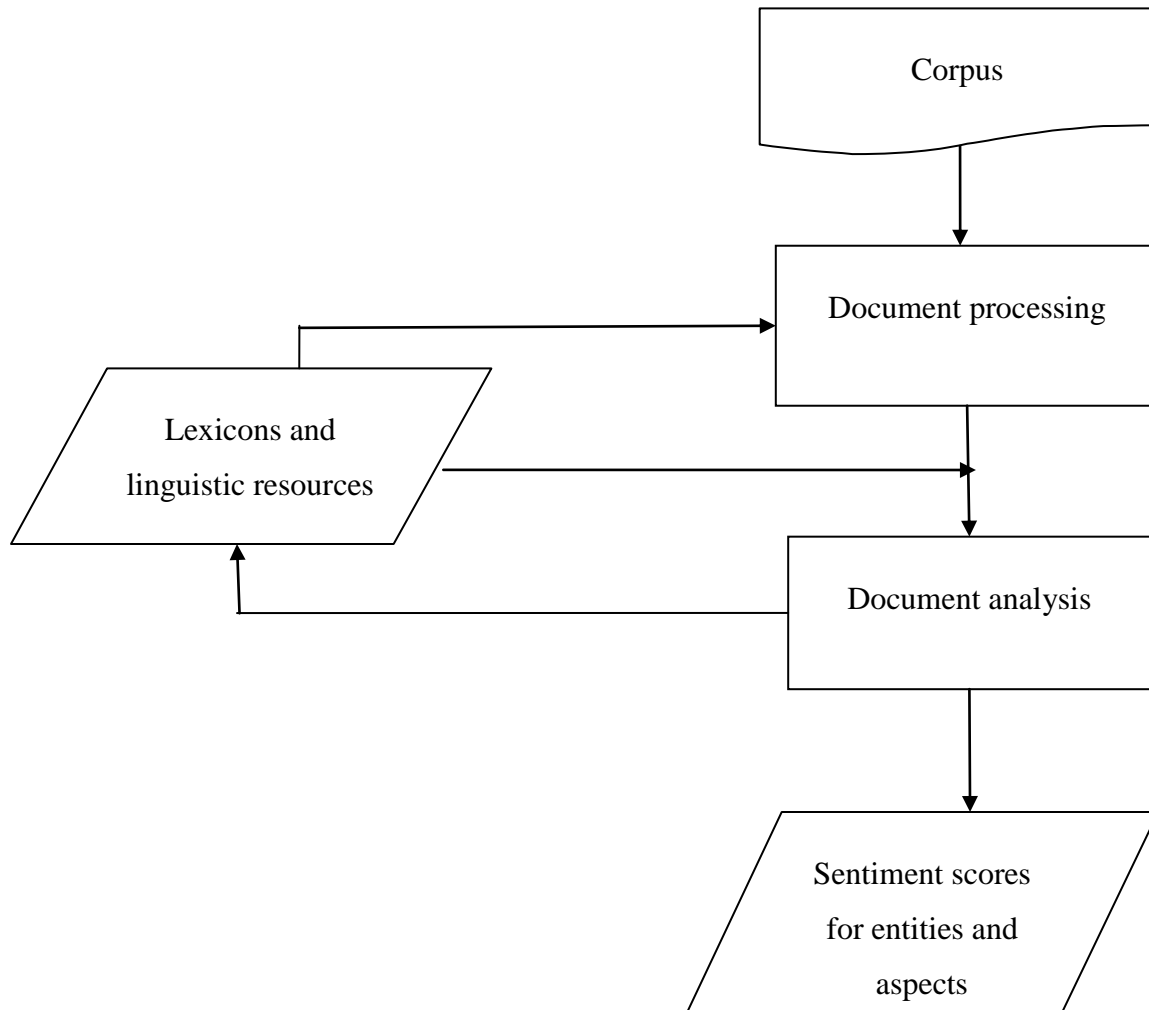


Figure 1.6 General architecture of emotion mining

1.4 PHASES OF EMOTION MINING: The process of emotion mining includes the following phases:

- i. Collection of data
- ii. Preprocessing of data
- iii. Feature Extraction
- iv. Negation Handling
- v. Opinion Classification

1.4.1 Collection of Data: In this phase, the first step is to collect the data from the different sources. These sources can be surveys or social networking sites such as twitter,

Facebook, blogs, any economic site or review sites etc. and organize the data in proper format.

1.4.1 Preprocessing of Data: Data collected from the various social networking sites are highly unstructured in nature because of written in the informal language. Thus it needs to clean the data before processing it. It includes:

- i. Stemming:** Stemming is performed on each word to get a root word i.e. common morphological endings of the related words. For Example, walking, walked, walks are stemmed to walk.
- ii. Removal of stop words:** Dataset collected contains a huge amount of data and while processing the data it would give inaccurate results. Thus it needs to remove the stop words such as “is, from, are, across, am etc.” that are not useful in emotion mining. For example, “this girl is beautiful” will be processed and gives the output “girl beautiful”.
- iii. Part of speech tagging:** POS tagging is a technique of applying a part of speech tag to every word present in the input data such as noun, verb, adjective. Thus with POS tagging, words of the input data are grouped into categories of noun, adjective, verb. For example, “the quality of the product is good” will be tagged as “the (determinant) quality (noun) of (preposition) the (determinant) product (noun) is (preposition) good (adjective)”.

1.4.2 Feature extraction: the purpose of this phase is to extract the features from the sentence which can be used for opinion classification. Features are basically tagged as nouns or noun phrases which are present as subject or object in the sentence. There are various approaches used for feature extraction such as n-gram, feature ranking algorithm, syntactic rules etc.

- i. Term frequency:** Term frequency aims to extract the terms which are occurring very frequently in the document.

$$\text{Term frequency} = F_{mn}/F_{dn}$$

F_{mn} = total number of times m term occurs in document d.

F_{dn} = total number of terms are there in document d.

- ii. **Term frequency-inverse document frequency:** This method defines that how significant a term to a document.

$$TF - IDF = (F_{mn}/F_{dn})\log(1/F_{tm})$$

Where F_{tm} is the total number of documents having term m.

- iii. **N-gram:** N-gram method defines the sequential list of n terms from a given order. It is used to find the probability that term will occur in the future and to extract the multiword features such as battery life, memory card etc. In n-gram, n defines the chunk of adjacent words of size 'n'. n-gram model has various types i.e. unigram, bigram, trigram etc.

For example, text: Sun rises in the east.

Unigram: “sun”, “rises”, “in”, “the”, “east”.

Bigram: “sun rises”, “rises in”, “in the”, “the east”.

Trigram: “sun rises in”, “rises in the”, “in the east”.

Unigram defines the individual words available in the text. Bigram defines the pair of consecutive words and trigram defines the three consecutive words.

1.4.3 Negation handling: In this step, the polarity of the negative words is swapped. If the word is identified with negation, then its polarity will be changed from positive to negative.

1.4.4 Opinion classification: There are various machine learning classification techniques that are used for opinion Classification such as Naïve bayes, Support Vector Machine, K-nearest neighbor, Decision tree etc. which classifies the words into positive, negative or neutral category, happy or sad category, subjective or objective category. There are two main steps in machine learning techniques. The first step is to construct a crisp model to distribute the class labels of the training set. Class labels of the training set is provided to the classifier that defines whether the dataset belong to positive class or negative class and the second step is to use the resulting classifier to predict the class labels to the testing set having unknown class labels. Machine learning is also called supervised learning.

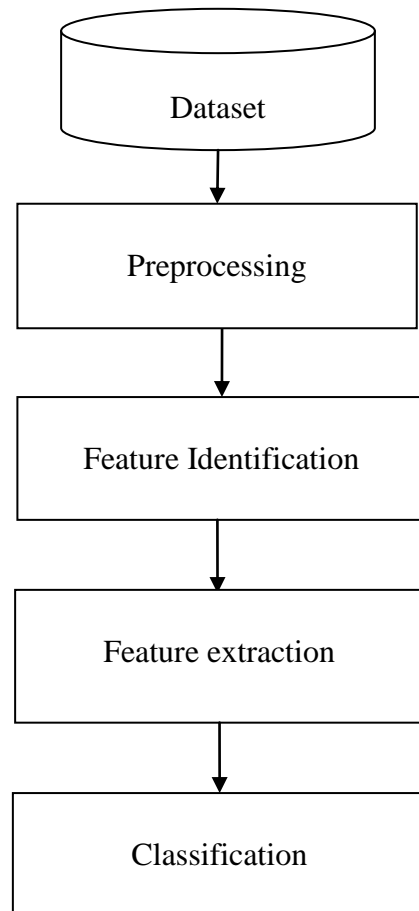


Figure 1.7 Emotion mining Process

1.5 Organization of Report: The rest of the chapters of the report are mentioned as below:

Chapter 2: Chapter 2 describes the literature survey that are studied to get the precise knowledge of the topic as well as various methods and technologies that are available for the emotion mining.

Chapter 3: Chapter 3 deals with the present work that contains the description about the problem formulation, objectives of the study and methodology of the proposed approach.

Chapter 4: Chapter 4 contains the experimental results, discussion and the comparison of the propose results with the existing results.

Chapter 5: Chapter 5 discusses about the conclusion and the future of the proposed work.

Chapter 2

LITERATURE REVIEW

Literature review contains the literature of all the research papers that are studied to get the complete knowledge about the research topic and to find the research gap. Literature review includes the research papers that are related to various techniques and methods that are available for the classification and feature extraction.

Mohamed Yassine et al.[5] In this paper, emotion mining performed on the text shared online in the form of comments and wall post of online social networking sites in order to identify whether the text is subjective or objective. Because most of the data shared online are written in informal language thus lexicons are developed for social acronyms, emotions and for foreign languages. After collecting the data from social networking sites, subjectivity features are extracted from the text based on correlation measure to avoid redundant attributes i.e. number of affective words, punctuation words, repeated words, social acronyms, emotions etc. continuous attributes are mapped to discrete values by using K-Mean clustering with K=3 for subjective, objective and moderate subjective text. Min-max normalization is performed to map the attribute values between 0 and 1. In order to predict the relationship strength between two users SVM is used.

Alaa El-Halees et al.[6] In this paper, a technique had proposed to extract the course related features and perform opinion mining at the document level to improve the teaching quality. First of all, data was collected from the discussion forums related to five courses then preprocessing was performed to remove the stop words and HTML tags and then stemming and tokenization was performed on the data. TF-IDF weight technique was performed on the text to obtain the vector representation. After that course related features were identified, extracted and grouped by considering the semantic orientation expressed in feature and then opinion classification was performed by using the supervised learning methods to classify the document into positive or negative polarity and then results of various supervised learning methods i.e. SVM, K-nearest neighbor, naïve Bayes were compared.

HAN-XIAO SHI et al.[7] In this paper, a supervised machine learning and TF-IDF approach was used to classify the hotel reviews from the web into positive or negative polarities. First of all, the reviews have been collected from the discussion forum and then TF-IDF technique is used to extract the important unigram features from the text. In TF-IDF, weight is assigned to every feature and evaluates how significant the words are in the corpus then classification is performed by using the SVM classifier which classifies the reviews into positive or negative reviews and then performance is evaluated by using the recall, precision, and f-score. Results are compared with the frequency and it has been analyzed that TF-IDF is more accurate than frequency.

R. R. Kabra et al.[8] In this paper, an approach had proposed to predict the performance of the students based on their past performance records by using the decision tree classifier. This analysis will help the teachers and administration to take necessary actions on weak students and help them to improve their performance and thus it will improve the overall performance of the institution. In this approach data regarding the past performance of the students are collected from the enrollment forms such as higher class percentage, secondary class percentage, entrance exam marks etc. After identifying the necessary attributes, data is classified by using the decision tree classifier. In decision tree classifier, internal nodes define the value of the attribute and arcs defines the condition on the attribute and leaf nodes defines the class label associated with the tuple and path from the root to a leaf node defines whether a student will pass or fail. Thus in this way, decision tree classifier is used to predict the future performance of the students.

Ayesha Rashid et al.[9] In this paper, two sequential pattern mining techniques such as generalised sequential pattern and apriori algorithm had applied in order to extract the opinion words and frequent features from the Student feedback data and the results of two techniques was compared. In this methodology, first of all 5 data files of online student feedback were crawled from the web and then preprocessing was performed to remove the irrelevant text like fake comments, HTML tags, spelling mistakes. After preprocessing Go tagger is used to tag the features (nouns) and opinion words (adjectives) and then tri gram modelling was applied to convert the text into structure form and to extract the valid rules. GSP and apriori algorithm was used to extract the best features and adjective rule

combination. In apriori algorithm two parameter i.e. support and confidence are used to extract the best rules and in GSP, candidate sequence is used to extract the frequent features. The value less than minimum support was discarded. In this way best rules were extracted and then these rules were applied on testing files in order to test the accuracy of all the rule combinations and their Precision, recall and accuracy of both the techniques are compared and it analyzed that GSP is more accurate than the apriori algorithm.

Nabeela Altrabsheh et al.[10] In this paper, a sentiment analysis approach was used to analyze the real-time feedback of students on teaching by using the four aspects i.e. preprocessing, features, different classification techniques and usage of neutral class. In this work, dataset was collected from the graduate and postgraduate students of Portsmouth University and the other Institutes about the lectures. After collecting the data, it was divided into three class labels that is positive negative and neutral. Preprocessing is performed at various levels to remove the irrelevant information such as numbers, punctuations, repeated letters, blanks and stop words. For feature selection, n-gram technique was used for combining the unigram, bigram and trigram. Different machine learning techniques were used for classification that is naive Bayes, SVM, maximum entropy and complement naïve Bayes. In this work, it has been analyzed that accuracy has been increased by 20 percent by using the preprocessing methods and the best feature selection method was the combination of unigram and bigram and the best classification models are SVM and complement naïve Bayes for neutral as well as non-neutral classes.

Amandeep Kaur et al.[11] In this paper, opinion mining was performed on Punjabi text by using the hybrid approach of Naive Bayes and N-gram. Dataset has been collected from various Punjabi newspapers then stemming was performed on each Punjabi word to get a root word. For example, walking, walked, walks was stemmed to walk. Data may contain some negative words thus negative handling was performed to convert the positive polarity to negative polarity and negative polarity to positive polarity. Features were extracted by using N-gram model then trigrams of testing set matches with the trigrams of the training dataset. If matches found, then trimatched value gets incremented else repeat the steps for bigrams and unigrams and the value of respective match gets incremented. All the

information related to unigrams, bigrams, and trigrams is contained in feature vector for testing and training of naive Bayes.

Vijay B. Raut et al.[12] In this paper, both supervised machine learning as well as SentiWordnet approach was used to classify the opinion of the users upon hotels that were posted on the websites. The proposed approach has three phases: text retrieval, classification and then summarization. First of all, reviews that were posted on hotel websites are retrieved such as TripAdvisor. Com by using the web crawling and then classification was performed to classify the reviews into positive or negative reviews. Two approaches were used for this i.e. machine learning and SentiWordNet. In machine learning method, algorithms were trained and tested to classify the reviews. Three classifiers were used for this i.e. naïve Bayes, decision tree and support vector machine. Sentiwordnet repository contains all the positive or negative words along with their polarity score and those scores were calculated to get the polarity of the reviews. The last phase was opinion summarization, in this most relevant sentences are extracted from the document and sentences are sorted according to the relevance score. Thus this approach is very useful to the users to make decisions about the hotels based on reviews.

Mondher Bouazizi et al.[13] Sarcasm is defined as opposite meaning of the sentence that any people speak. In this paper, tweets on any topic were classified into the polarity of positive and negative. Sarcastic tweets were also detected to improve the accuracy of sentiment analysis. After collecting the tweets, textual and non-textual features were extracted from the tweets. Negative handling was performed on the text to convert the positive words into negative and vice versa. After extracting all the features, classification is performed on the test set by using naïve Bayes, SVM and maximum entropy. It has been analyzed that some of the tweets are misclassified due to the presence of sarcasm. Sentiment related, punctuation related, syntactic, pattern related features are extracted for sarcasm detection. After comparing the results, it has been analyzed that results are more accurate after taking sarcasm into consideration.

Eman M.G. Younis et al.[14] In this research paper, text mining and sentiment mining was applied to analyze the twitter data about the products and services of two different UK stores (Tesco and Asda). This will help the business to check the views of customers about

their products and services. Twitter messages are accessed using `twitterR` package. Next, the data is cleaned from stop words, spaces and perform stemming. In this `tm` packages are used. It produces a structure representation of tweets. Next, the data mining techniques are used such as association rules, finding most frequent words and sentiment mining (lexicon-based approach). In lexicon-based approach, there is a set of positive and negative words, which are combined with a scoring function to determine the polarity of sentiments. Finally, `wordcloud` package and bar charts are used to show the frequency of words and sentiment score in the customer tweets.

Vandana Jha et al.[15] In this paper, opinion mining was performed on the movie reviews of Hindi language and classifies the whole document into positive, negative and neutral class. There are various challenges while performing opinion mining on the Hindi language because there is no particular order of subject. Some words may have multiple meanings and lack of resources. There are two methods used in this work. The first method is using supervised learning in which naïve Bayes classification is used to classify the whole document. In the second approach unsupervised learning is used in which part of speech (POS) tagging is taken into consideration. POS tagger extracts the adjectives from the whole document and makes a word list of positive and negative seed list according to the occurrence of the adjectives and then all the adjectives in the document are matched with the seed list. If the positive adjectives are more than negative then the review is classified as positive else classified as negative and if both are equal then review belongs to a neutral class.

Gautami Tripathi et al.[16] In this paper, both natural language processing and machine learning techniques were used to extract the sentiment from the movie reviews. First of all, preprocessing is performed on the data collected from the web because of its unstructured nature. Preprocessing includes various steps such as tokenization, pruning, filtering tokens and stemming. Next step is to extract the relevant features from the data. For this term occurrence, TF-IDF, term frequency and binary term occurrences are used and then n-gram method is used to extract the adjacent words i.e. unigram, bigram, trigram, four-gram and for classification both SVM and naïve Bayes algorithms are used and their accuracy is measured

and it has been analyzed that TF-IDF along with SVM classifier gives the maximum accuracy and term occurrence along with naïve Bayes gives the maximum accuracy.

Monika Arora et al.[17] Most of the information on the internet written by the user is in unstructured form i.e. in abbreviation form or syntactically incorrect. In this paper, a framework had proposed to deal with the informal language and classify the opinions. Input has been given from the web document. It could be from blog, forum or twitter then the documents are processed. First of all, tokenization is performed to separate each and every sentence from the document and then filtering of the stop words are performed to extract the root words and next step is to perform the POS tagging by using the wordnet in which each word is tagged with the noun, adjective, adverb etc. but it is a big challenge for the tagger to tag the slang words as they are syntactically incorrect. In order to overcome this problem, a slang dictionary is made which contains the slangs as well as their corresponding English words. Thus if the word is not found in the wordnet dictionary then it is checked from the slang dictionary to obtain the corresponding English word and then words are retagged by the tagger and tagging classification is performed to classify the opinions.

Ms. Ashwini Rao et al.[18] In this paper, an algorithm had proposed to select the relevant features in order to reduce the complexity of classification. An unsupervised and domain independent technique is used to extract the features from the data. As data collected from the web is highly unstructured thus preprocessing is required to remove the stop words, special words, tags. The boundary of the sentence is also determined. For feature generation, first of all, Stanford parser is used to tag POS to identify the common features and then proposed algorithm is used to extract the more frequent features and then threshold value is used to filter out the rules which provide more relevant features and then some rules are defined to further refine the features. At the end, it has been analyzed that there is a reduction in the feature space.

Chandrika Chatterjee et al.[19] In this paper, an approach had used to predict the best features in order to extract the relevant information from the responses to the student feedback questions and the sentence level and token level sentiment classification is also compared in this paper. In the sentence level, data has been collected from the online survey

and then preprocessing is performed for tokenization and stemming then TF-IDF method is used to extract the most frequent features and n-gram method is used to extract the consecutive tokens from the text and then naïve Bayes and SVM is used to classify the opinions into positive or negative polarities. In the token level, the same dataset has been used then tokenization is performed to break the text into words. After that each word is tagged into corresponding POS using the Stanford POS tagger and then stemming is performed to get the root word and n-gram technique is used to extract the consecutive words or characters and then classification is performed by using the SVM, naïve Bayes, j48, decision tree and it has been analyzed that decision tree has more accuracy than naïve Bayes and SVM.

Alok Kumar et al.[20] In this paper, an approach was proposed in which system has itself extract the relevant features of teachers from the student feedback rather than working on a predefined set of questions. In this approach, three datasets have been collected. Two from the websites and the third one is manual feedback from students. Data cleaning is performed to remove the unwanted text, tags, links etc. and then preprocessing is performed such as fragmentation, removal of stop words, root word conversion, tokenization, eliminating the named entities etc. after preprocessing, essential features from the student feedback are identified by using TF-IDF and maximum entropy & mutual information gain. Supervised and unsupervised techniques are applied to extract the subjective sentences and positive and negative sentences are stored in separate files then by using the PLA (probabilistic latent semantic analysis) and LDA (Latent Dirichlet Allocation) sentiment scores are allocated to every feature. After feature evaluation, feature aggregation is applied to aggregate the score of every feature and then features of different teachers are compared by using different report generation methods.

Gokarn Ila Nitin et al.[21] In this paper, feature based sentiment analysis as well as natural language processing techniques are used to extract and analyze the topics as well as a sentiment from the student feedback on particular course based on different features like teaching, learning, course content. First of all, the feedbacks from the students on the different courses are collected then preprocessing is performed on the data to remove the stop words and to perform the stemming and tokenization. After preprocessing, the frequency of

each term is calculated to evaluate the importance of the term in the document. For this, each term in the comment is transformed into the numerical matrix using the TF-IDF method and then agglomerative clustering is used to grouping the similar topics in the comments based on their similarities and the topics with high frequency are extracted from the clusters. After feature extraction, logistic regression classifier is used to classify the comments into a positive or negative category and the results are presented using the barcharts.

Ang Yang et al.[22] This paper aims to classify the tweets into a positive or negative category by using different classifiers. In this paper, a technique has been proposed to increase the accuracy and performance using the sentimental lexicons and unigram with high frequency. First of all, tweets from the Twitter are extracted using the API and then positive and negative tweets are separated manually. Preprocessing is performed to reduce the numbers of irrelevant features as it is very difficult to process all the features. In preprocessing special characters, repeated letter words, URLs, stop words are removed. After preprocessing frequent features are extracted by counting the occurrence of positive and negative words and Chi-square of high information gain is also calculated and then sentiment lexicons are used to remove other little sentimental words. After extracting the features, these features are trained with six classifiers such as Bayesian multinomial naive Bayes (MNB), SVM, KNN, decision tree, Logistic regression model and it has been analyzed that MNB and SVM are more accurate than other classifiers. The false positive, f-measure recall rate and accuracy of this proposed work are also compared with previous works.

Shoiab Ahmed et al.[23] In this paper, a technique has been proposed to classify the online reviews of movies and mobiles into seven different categories such as strongly positive, weak positive, positive, negative, neutral, strong negative and weak negative be based on the SentiWordNet. SentiWordNet is a lexical repository available for research purpose which provides a semi-supervised technique for opinion classification. first of all, online reviews are collected by using the web crawler and are organized in a text document and then preprocessing is performed to remove the stop words and to remove the suffixes to get the root word such as 'ing', 'ed', 'full', 'ness' etc. by using the Porter stemming algorithm. After that POS tagging is performed to tag the string of words with its parts of speech such as verb, noun, adjective etc. by using the Stanford parser and then SentiwordNet is used to score

each and every word and to convert the score in order to classify them into seven different categories. Different classifiers like SVM, naïve bayes, multilayer perception are used to evaluate the Precision value, accuracy, correctly or incorrectly classified instances.

Fajri Koto et al.[24] In this paper, an approach has been proposed to detect and analyze the sentence pattern of the tweets in two areas i.e. polarity and subjectivity by utilizing the POS sequence and information gain. Tweets can be positive, negative, subjective or objective. People mostly use adjective or adverb instead of noun to express their emotions or opinions. In this work five datasets have been used i.e. Stanford Twitter sentiment, Health Care Reform, Obama McCain debate, Sanders and international workshop sem-eval. All the tweets in these datasets contain a positive negative or neutral label. After extracting the tweets various sequence of n-tags are used such as n=2, n=3 and n=5. By using the information gain, only top hundred sequences of n-tags are selected and then SVM weights are assigned to get the top 10 sequences that are included in two or more datasets. Subjectivity of the tweets are evaluated by calculating the subjective, objective positive and negative frequencies of the sequences and it has been analyzed that adverb and adjective are used more for subjective tweets and nouns are used in objective tweets. The performance of the POS sequence is increased by using the AFINN lexicon which contains the score for positive and negative words.

Ashwani Rao et al.[25] In this paper, unsupervised and domain independent technique had proposed to extract the relevant features. First of all, domain-independent data set has been collected. This data is in the unstructured form due to which preprocessing is performed on the data. Preprocessing includes removal of noise and overused punctuations, converting repeating symbols to single occurrence etc. After cleaning the data, Stanford parser is used to extract the relevant features from the data. Every word in the sentence tagged by the parser according to the parts of speech but with this many irrelevant features also got extracted. In order to overcome this, n-gram rule is used to extract the multiword features or consecutive occurrence of words. The output of this rule can be bigram, trigram tagset with NN/NNS or NNP. The last step is to find the frequency of occurrence of rules. In order to extract the relevant features optimum threshold value is decided. This value is obtained when a maximum number of features from the tagged dataset is matched with the features extracted

from the set of rules. This approach makes the feature set lesser in size, more compact and relevant by removing the irrelevant features.

MinaraP anto et al.[26] This paper proposed a technique which provides an automatic feedback of a product by collecting the data from the twitter. For this, first of all, data from the twitter has been extracted using the twitter 4J API. After extracting the data, preprocessing is performed in which stop words like “is, am, but” are eliminated and the sentences are divided into smaller sentences and then POS tagging is performed on the text in which each word is assigned part of speech tag such as noun, verb, adjective etc. then SVM classifier is used to classify into positive, negative or neutral words. After analyzing the efficiency, it has been analyzed that SVM classifier is more accurate among other classifiers i.e. Naïve Bayes, Maximum Entropy etc. In order to classify the data more accurately dual prediction technique has been used which evaluates the sentence from both directions which gives the output of two distinct values and the mean among them has used to find the sentiments from it and provides feature based rating to a product and get the overall rating, unigram approach is used in which the frequency of the words is evaluated.

Jumayel Islam et al.[27] With the advancement in the opinion mining, the users are not satisfied with the overall opinion mining of the document. Users are focusing on the feature based opinion mining of an entity. Thus it becomes very important to extract the relevant features from the dataset. In this paper, two datasets are taken one is domain dependent and other is domain independent then Stanford parser is utilized to analyze the syntactic composition of every sentence and syntactic rules are applied to extract the candidate features from every document. Candidate features include the subject or object found in a sentence. There are six types of syntactic rules are applied to extract the candidate features i.e. nominal subject relationship, preposition object relationship, conjunct relationship etc. After that inverse document frequency (IDF) method is used to assign the weights to the candidate features then average weight of term in all the documents, average weight of all the terms in the document, standard variance, dispersion, deviation, domain relevance of the term is calculated which signifies the frequency as well as the significance of the term over the dataset. Domain relevance of domain dependent dataset is called intrinsic domain relevance

and the domain relevance of domain independent dataset is called extrinsic domain relevance and the threshold value for both IDR and EDR are chosen in order to extract the opinions.

Tirath Prasad Sahu et al. [28] In this paper, movie reviews are classified on the scale of 0 to 4 i.e. highly disliked to highly liked. First of all, more than 50,000 reviews are collected from IMDB because of the collected data is highly unstructured, preprocessing is performed on it i.e. stemming (extracting the root word), stopping (removing the most commonly used stop words), part of speech tagging (words are tagged as noun, verb, adverb etc.). After preprocessing, features that affect the polarity of the document are analyzed by using the Sentiwordnet i.e. positive sentiment words, positive sentiment bigrams, negative sentiment words, negative sentiment bigrams etc. after analyzing the features, information gain and feature ranking algorithm are used to scale all the features and provide sentiment score. After providing the sentiment score, class labels for the document is determined i.e. strong negative, weak negative, neutral, weak positive, strong positive. After classification, this technique is compared with the other classifiers i.e. naïve Bayes, decision tree, KNN etc. and it is found that this technique is more accurate than other classifiers.

Dhanalakshmi V. et al.[29] This paper performs the opinion mining on the student feedback data and classifies the data into positive or negative polarity by using the various machine learning classifiers such as SVM, K-Nearest Neighbor, neural network, naïve Bayes and also compares the performance of these classifiers. First of all, the survey is conducted to collect the student's feedback data regarding the various features of learning and teaching then preprocessing is performed on the data which includes tokenization, removal of stop words and stemming. In order to train the supervised learning algorithms, positive and negative files are assembled into separate folders. TFIDF (term frequency-inverse document frequency) technique is used to generate the word vector representation and then various features of the data such as module, exam, teachers, lab resources are extracted to understand the polarity of the opinions and then machine learning algorithms are used to classify the opinions into positive or negative polarities and also the performance of the machine learning algorithms are compared based on accuracy, precision and recall. It has been analyzed that naïve Bayes classifier is more accurate than others.

Rushlenekaurbakshi et al.[30] In this paper, an algorithm has been proposed which provides more accurate results to understand and predict the ups and downs of stock prices. In this approach, first of all, tweets are extracted from the twitter using the twitter 4J library and store them into the database. Next preprocessing is performed to remove the irrelevant characters from each and every tweet. The unsupervised technique is used to assign the polarities to every word. For this dictionary is made which contains positive, negative and neutral words and their corresponding polarities. In an algorithm, string tokenizer is used to divide the tweets into words and then array is used to store all the relevant words and then each and every word in the array is matched with the dictionary and then polarity is assigned to every word according to the sentiment score and at the end, error and accuracy percentages for both the human prediction and software calculated is evaluated.

E Deepak et al.[31] In this paper, eight parameters of the faculty such as faculty profile, R & D, mentoring, organizational qualities etc. are considered to evaluate to the performance of the faculty. First of all, data has been collected from the faculties on the basis of 8 parameters and 31 attributes and then preprocessing is performed to find the missing value from the data and then feature selection is performed to extract the relevant features and to prune the irrelevant features. Data collected from the faculties are compared with the data taken from the UCI datasets such as iris, blogger, wine etc. and then classification is performed by using the different kernel methods of SVM like polynomial, Radial basis and PUKF (Pearson VII function based universal kernel) and it have been analyzed that PUKF has more accuracy while evaluating the performance of faculty dataset as compare to UCI datasets.

R.Nithya et al. [32] In this paper, a combination of lexicon based and syntactical based approach has been proposed in order to detect the overall sentiment score, feature score and also the most relevant features of the product. In order to find the overall sentiment score, first of all, data has been collected from the web and then preprocessing is performed to remove the tags, stop words and the text is divided into words and then POS tagger is used to tag every word. After the data cleaning, chunking of data is performed by the regular expression parser in order to extract and filter the adjective phrase accompanied by a noun phrase and then lexicon based dictionary is used to perform the sentiment classification. Thus

in this way, both syntactical rule and lexicon approach i.e. LPSA (Lexicon pattern sentiment analysis) is used to determine the overall sentiment score. In the second part to extract the most promising features from the text, term frequency method is used to avoid the irrelevant features. FBSC (Feature based sentiment score) algorithm is used to score the features of the product. In this way, opinion classification is performed based on overall scores as well as feature scores.

Pankaj Kumar et al.[33] In this paper, opinion mining concept has been used to extract and analyze the tweets from the twitter and evaluate the attitude of the users towards the two main enterprises and their products. The process starts with by extracting the tweets from the twitter by using the packages of R language in order to connect with twitter API and then tokenization is performed by using the POS tagger. In order to remove the URLs, hashtags, non-English words, numbers, nouns, prepositions and to replace the informal words, preprocessing is performed. Next step is to extract the polarity score i.e. number of positive or negative words, hashtags, emoticons and then sentiment score is evaluated on the basis of the higher value of the positive or negative score of words stored in the repository in the backend. For the better classification of the opinions, naïve Bayes classifier is used. After performing the sentiment classification, histogram and pie chart are used to visualize the results of positive, negative and neutral tweets. In this way market reputation of two enterprises (google and Microsoft) and their products are computed.

Neethu Akkarapatty et al.[34] In this work, a machine learning technique is applied to classify the reviews of the tripadvisor.com to classify the reviews into the positive and negative category. In the training phase, each input word is converted into feature set and corresponding class label using the feature extractor. First of all, data is collected from the TripAdvisor and then preprocessing is performed to remove the inconsistent and incomplete raw data. Preprocessing include removal of stop words, punctuations and perform tokenization. After preprocessing, positive and negative files are made for both training as well as testing. After that, different feature extraction methods are applied to extract the relevant features such as character based features, bag of word based features, POS tag based features, aspect based features. To further reduce the features space Chi-square method is used. After extracting the most relevant features different classifiers such as MNB, SVM

with the kernel (linear, polynomial, radial, sigmoid) are applied to classify the reviews into positive or negative class and the results show that POS tag, bag of words is the optimal feature extractor.

Mohammad aman ullah at al.[35] In this paper, a model has been proposed to analyze the feedback of students on teaching activities of social networking sites especially Facebook by using the machine learning techniques. In this approach, corpus has been collected from the university of Portsmouth and dataset has been assigned a class label such as positive, negative or neutral according to the intensity of the comment. Next step is to find the best preprocessing tool to make the data error free. For this different levels of preprocessing are used like preprocessing w/o in this no preprocessing is performed, preprocessing 1 in this unimportant characters, numbers and stop words are removed so on up to processing 5 to remove the repeated words, emoticons, URLs and to perform stemming. Each level of preprocessing uses the data that are being preprocessed by the previous level. After preprocessing, next step is to extract the features which enhances the accuracy for this n-gram method has been used and it has been analyzed that trigram has more accuracy than unigram and bigram. Last step is to classify the sentence by using the machine learning classifiers such as SVM, CNB (complement naïve bayes), NB, ME and it has been analyzed that SVM has highest accuracy.

R.Menaha et al.[36] In this paper, a student feedback system is made in which clustering and classification techniques are used to extract the topic and classify the responses into the positive or negative category. First of all, student feedbacks are collected for a particular course then preprocessing is performed to breaking the data into sentences and to remove the stop words. After preprocessing, topics such as faculty interaction, delivery style, punctuality etc. are extracted from sentences by using the highest frequency count and then clustering is performed to make a cluster of similar comments in every topic for example in faculty interaction topic there can be comments on teaching, entertainment, friendliness, help etc. thus clustering is used to make a group of these comments belonging to the same topic then classification is used to classify comments into positive or negative category by comparing with positive and negative words

Chapter 3 deals with the present work in the field of emotion mining that includes problem formulation and the objectives of the proposed work.

3.1 PROBLEM FORMULATION

From the literature survey, it is concluded that there are various feature extraction and machine learning classifiers that are used for the emotion mining. In the existing approach of student feedback mining there is not any feature extraction technique is applied to extract the relevant sentiment words. Students can also use special characters like emoticons to express their sentiments toward the university academic feedback so it is also required to consider the polarity of the emoticons. The accuracy of the output can be improved by using suitable feature extraction technique to remove the irrelevant features and to extract the more compact features of the input data.

3.2 OBJECTIVES

The main objectives of the research work are as follow:

- i. To propose an approach to analyze and classify the feedback of students on various parameters of university such as examination, course content, teaching, practical work, library facilities and extracurricular activities by using the n-gram rule generation feature extraction methods that works on the parts of speech of the input text in order to extract the opinionated words and to assign the sentiment scores to the opinioned words from the list of sentiment words.
- ii. To make improvements in the existing technique and compare the results in terms of accuracy, precision, recall and f1-measure and that will provide more detailed analysis of qualitative feedback of the students and helps the university management to improve the effectiveness of the various academic activities.

3.3 RESEARCH METHODOLOGY

Research methodology is designed to show the functionality of the research work before it is actually implemented.

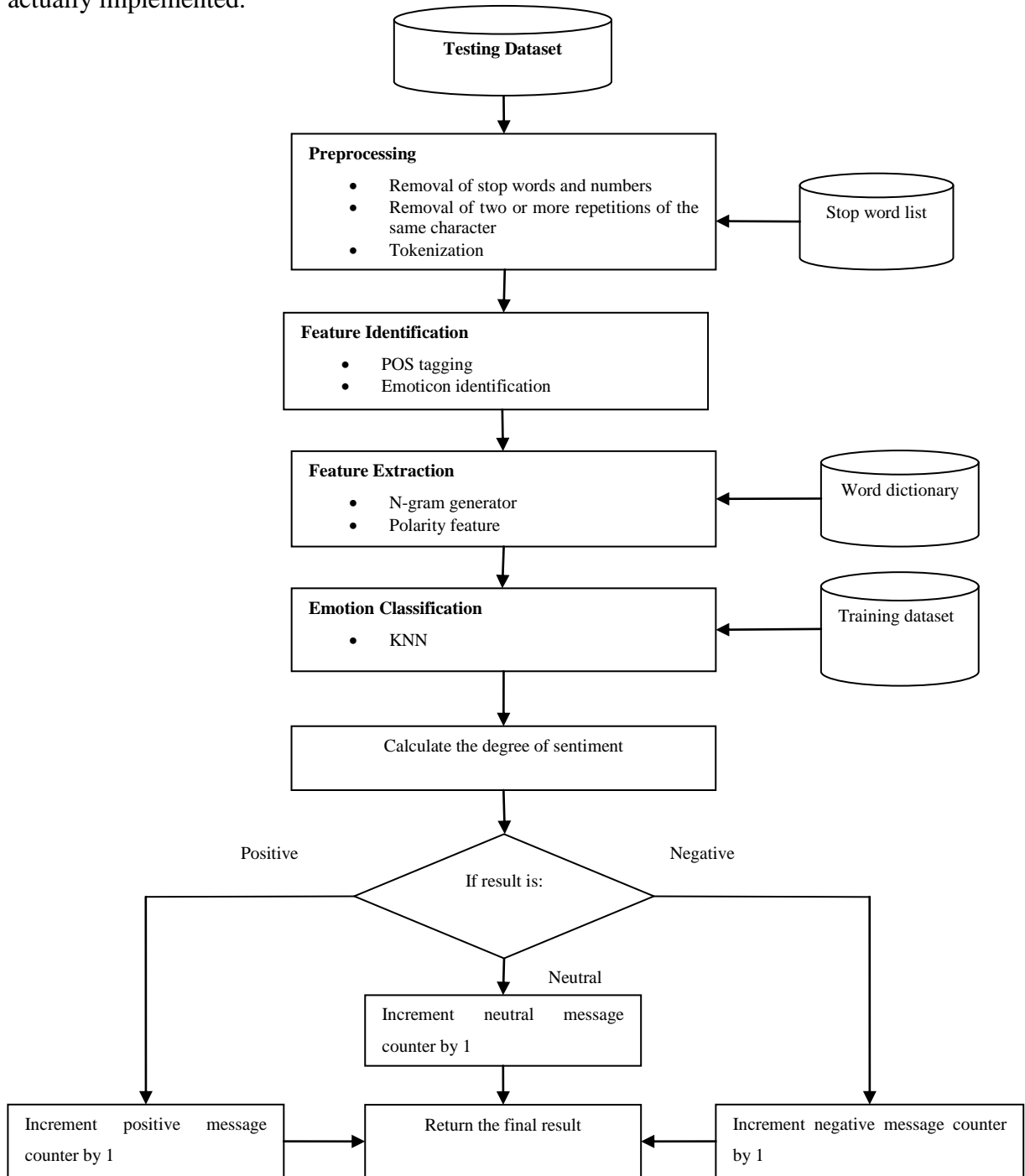


Figure 3.1 Work flow diagram

This phase elaborates how the formulated problem can be solved effectively before actually solving it. It describes various steps to accomplish the given problem are as follows:

- i. Online or manual collection of the reviews of the students on different parameters that effects the performance of the university like examination, teaching, course content, practical work, library facilities, extracurricular activities and organize the dataset in the Microsoft excel file.
- ii. Perform the preprocessing on the dataset to remove the irrelevant data which does not hold any sentimental value such as removal of numbers, stopwords, special characters and tokenize the text.
- iii. Identify the features from the text by using the Part of speech (POS) tagging. POS tagging is a technique of applying a part of speech tag to every words present in the input data such as noun, verb, adjective, adverb etc.
- iv. Extracting the compact and relevant features from the text by using the sequence of n-tags that is the combination of the two tags such as adverb-adjective, verb-adverb etc. or three tags such as verb-adverb-adjective etc.
- v. Provide score to the features by comparing them with the file containing list of words having weight assigned to each and every opinionated word in order to get the polarity of the word and then aggregate the scores of every review to get the overall polarity of the review.
- vi. Perform classification on the data by using the machine learning classifier such as K-nearest neighbor (KNN) classifier in order to classify the reviews of the students into positive, negative and neutral category. There are two main steps in machine learning classification. The first step is to construct a crisp model to distribute the class labels of the training set (with known class labels) and the second step is to use the resulting classifier to predict the class labels to the testing set (unknown class labels).

3.3.1 Collection of data: The first step of our methodology is to collect the dataset from the students. Data can be collected online as well as manually. In order to collect the dataset online first of all Google survey form is designed shown in figure 3.2 which contains the open ended question upon six parameters such as examinations, course content, teaching, extracurricular activities, practical work and library facilities and

the link of the form is provided to the students and ask them to fill the form in the free text format.

Student Feedback Form

Please enter the information regarding the following parameters

***Required**

Registration number *

Your answer

Name *

Your answer

Give us your reviews regarding our university teaching activities in terms of interaction, lecture delivery, punctuality etc *

Your answer

Give us your reviews regarding the content of the courses including knowledgeable, depth of course, proper course material *

Your answer

Give us your reviews upon examination including exam pattern, marks distribution, paper checking etc *

Your answer

Give us your reviews upon lab or practical work including evaluations. *

Your answer

what do you think about library facilities *

Your answer

What do you think about the extracurricular activities held in university *

Your answer

SUBMIT

Figure 3.2 Google Survey form

The advantage of using Google survey is that it takes less time to collect the data and it automatically generate the data in the MS Excel sheet. Data has also been collected manually by distributing the questionnaires to the students of the university. After collecting the data manually, data has been organized in MS Excel sheet. In this way feedback of 185 students on the predefined parameters are collected and total number of reviews collected on different parameters are 1110 as shown in figure 3.3

	B	C	D	E	F
1	Course content	Examination	Practical work	library facilities	extracurricular activities
2	content of courses are average	examination pattern is good	not satisfactory, lab work must include labes	library facilities are good but number of	extracurricular activities are excellent and g
3	Not good	Good	Good	Not good	Good
4	All courses material provide very good kno	Exam pattern is up to the mark and the Cg	Lab work is properly covered in the labs by tl	Library facilities are excellent in terms of	Extra curricular activities also help students
5	Content of course is perfectly in line with tl	Again the university tests students of the	Good	Its the best thing i have seen in this univ	Complete wastage of time. Again this opini
6	content of courses improves my knowledge	examination pattern is good	practical work provides detail knowledge of	library has huge collection of books from	extracurricular activities increases mental a
7	Yes	Yes	Yes	Yes	Yes
8	Good	Good	Good	Hardworking	No views
9	This semester university has provide best t	I like the question pattern	Everything is going fine in lab . learning new	I am satisfied with the facilities but few	No idea about the extracurricular activities.
10	Needs some improvement	Good	Good	Its required some libral in rules	Good it provides a great platform
11	Needs some improvement	Good	Good	Its required some libral in rules	Good it provides a great platform
12	Knowledge and depth of the course is goo	This university is far better than other uni	Lab and practical are not upto the mark. The	Good	While coming to extracurriculum activities i
13	Knowledge,depth of course is good but cou	Great	Not upto the mark	Good	Awesome
14	course content is knowledgeable	exam pattern is good and good paper che	lab improves practical knowledge	good library facilities	ok
15	Content of the courses is good but some co	Pattern and procedure of the examination	Sometimes systems at labs are screwed up .	Library facilities are very good. No issues	This university is no1 in terms of extracurric
16	Good	Good	Good	Hardworking	Good
17	Very good	Excellent	Good	Good	Good
18	Material of the course is not good enough	Good	Practical work is fair	Excellent	Great job
19	sometimes proper course material is not re	Good	Good	Good	excellent
20	Easily available course material with suffici	Good marks distribution	Students can eas	Practical and easy learning	Good
21	The materials or links provided for study pl	Exam pattern and how it is conducted is r	The evaluation is again bad for our CA, as on	Again some of the faculties don't know h	This is really a good thing that events happ
22	The course notes are very knowledgeable	Descent.	Good.	Descent.	Good.
23	Good	Excellect	Good	Good	Good
24	Good	Excellect	Good	Good	Good
25	content of the courses are update and mos	exam pattern and marks distribution is ve	well and good infrastructure and have trend	library is very well managed and provide	extracurricular activities held in university

Figure 3.3 Dataset

3.3.2 Preprocessing of data: Collected dataset may contain some noisy data such as special characters, numbers, and emoticons. Thus preprocessing is used to clean the noise from the data and to reduce the errors. By removing these unnecessary data, accuracy of the output will increase. Preprocessing includes various steps:

- i. Remove numbers:** Dataset may contain some numbers which are irrelevant for the opinion mining like people may write the word ‘too’ as 2 and ‘for’ as 4. Thus it is required remove these kinds of numbers.
- ii. Remove special characters:** Dataset may contain some special characters such as question marks, underscore, blanks, hashtags etc. Most of these types of characters do not provide any subjective information. Thus, it is important to remove these kinds of special characters.
- iii. Ignore case and remove two or more repetitions of the same character:** Ignoring the case of the letters is necessary to match the pattern in the training data and the removal of the two or more repetitions of the same character and replace with the character itself. For example, the word ‘loooove’ with be replaced with ‘love’.
- iv. Removal of stopwords:** Dataset may contain various stop words such as ‘which’, ‘whom’, ‘yourself’, ‘ourselves’, ‘until’. It is very important to remove stop words to improve the outcome of the analysis. List of stopwords shown in figure 3.4

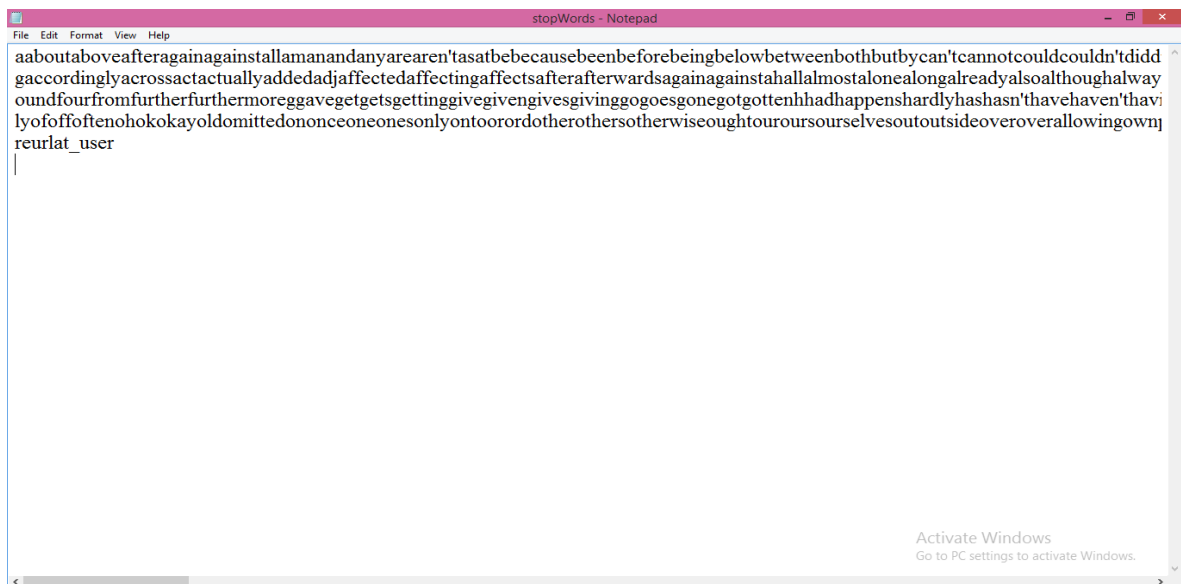


Figure 3.4 Stop word list

3.3.3 Tokenization: Tokenization is the procedure of breaking the text into smaller units called tokens. Text can be tokenized into phrases or single words by locating the word limits. Terminating of previous word and starting of the next word describes the word limit. In this module, based on the word list composed on the basis of most common words, the program is designed such that it reads the known words from the user comment. Then a word list is acquired by matching the words in the user comments besides a valid word list file already saved as a text file. Then the text file obtained is loaded into the memory and go through to the tokenization process for the further computations. The process of tokenization then extracts all of the words from the user comment and filters them on the basis of a list of common words containing no emotion. Then the filtered list is acquired after deleting the matching words in the comments of user. These common words are not provided in the word weight file that is containing the rank/weight of every word that is being used in the common English language, which contains neutral, positive or negative emotion.

3.3.4 Feature Identification: Feature identification includes the detecting the features from the text. Feature identification includes:

- i. Part of speech tagging:** POS (Part of speech) tagging is a technique of applying a part of speech tag to every words present in the input data such as noun, verb, adjective. Thus with POS tagging, words of the input data are grouped into categories of noun, adjective, verb and the tags collectively used for this process is known as a tagset. For example, “the quality of the product is good” will be tagged as “the (determinant) quality (noun) of (preposition) the (determinant) product (noun) is (preposition) good (adjective)”.Program that is used for the POS tagging is called POS tagger and it uses dictionaries, rules and lexicons to assign the tag to the text. After the POS tagging, some rules will be defined to use the verb, adverb, adjectives and their combinations to extract the subjective words.

Table 3.1 Conversion of POS tags to SentiWordNet tags

Sentiwordnet tag	POS tag
------------------	---------

a (adjective)	'JJ','JJR','JJS'
n (noun)	'NN','NNS','NNP','NNPS'
v (verb)	'VB','VBD','VBG','VBN','VBP','VBZ','IN'
r (adjective)	'RB','RBR','RBS'

ii. Emoticon identification: Emoticons depicts the facial expression of human. There is various type of emoticons used in the text to express the feeling of happiness or sadness. Thus it is very important to understand and interpret the meaning of emoticons for the accurate outcome of opinion mining. There are different categories of emoticons that are used for positive, negative and neutral opinions thus it is very important to distinguish the positive, negative and neutral emoticons. For example:

Positive emoticon::-) :) :o) :] :3 :c) :>

Negative emoticon::-(:(:c :< :[:{

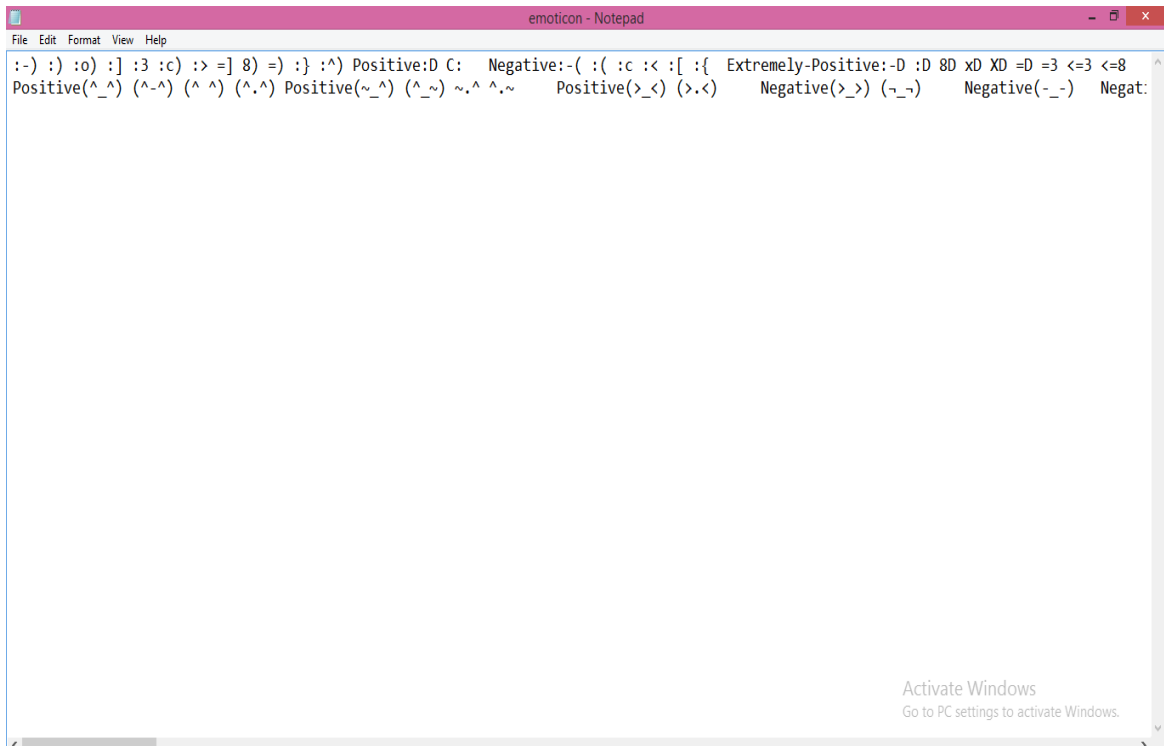


Figure 3.5 Emoticon list

3.3.5 Feature extraction: Feature extraction phase includes the process of finding the more compact and relevant features from the text. Feature extraction phase is as follows:

- i. **Sequence of n-tags:** After tagging the part of speech, various non-overlapping sequences of n-tags are applied. There are 2 forms of POS sequence used in this work i.e. n=2 and n=3 where n=2 is the combination of 2 tags and n=3 is the combination of 3 tags.

Table 3.2 Sequence of 2 tags

Sequence of 2 tags	Description
RB-VBG	Adverb-Verb
RB-VB	Adverb-Verb
RB-JJ	Adverb-Adjective
VBZ-VBG	Verb-Verb
NN-NNS	Noun-Noun
RB-NN	Adverb-Noun
VBP-JJ	Verb-Adjective

Table 3.3 Sequence of 3 tags

Sequence of 3 tags	Description
RB-JJ-NN	Adverb-Adjective-Noun
NN-NN-NN	Noun-Noun-Noun
NN-VBZ-RB	Noun-Verb-Adverb

- ii. **Providing scores to the features:** After extracting the sequence of the tags, sentiment scores are provided to the opinion word based on the sentiment scores of the SentiWordNet. Sentiwordnet is a lexicon resource that is related with three polarity scores i.e. positive, negative or objective. It is used for all parts of speech i.e. adjectives, noun, pronoun, verb, adverb and associates the polarity to words according to the sense rather than terms. Polarity scores can be positive or negative. Each

sentiment-bearing word in the dataset is given a score based on a logarithmic scale that ranges between -5 and 5 depending on its features. -5 is for the extremely negative terms and 5 is for the extremely positive terms as shown in figure 3.6. Score provided to every term is aggregated to get the overall polarity. For example, if in a review there are four positive and three negative scores then the overall polarity is calculated by aggregating four positive and three negative scores separately because the positive score is high than the negative score this review is classified as positive.



Figure 3.6 list of sentiment words

3.3.6 Opinion classification: The student comments are polarized in three major categories under this step. The three major categories are positive, negative and neutral. The tokenized comments are compared with a list of words. The file contains the ranking for each of the word listed on the list. The rank or weight or strength of the words has been listed in the document, which ranges between -5 to +5. The words are classified on the basis of their use and its impact in the natural English language. For opinion classification, K-nearest neighbor algorithm is used. KNN classifier is based

on the concept of similarity between the test tuples and training tuples. Test tuples within the dataset assigned a class label based on the closeness with training tuples having common attributes. The similarity between the two instances is defined by using the Euclidean distance.

There are two main steps in machine learning classification. The first step is to construct a crisp model to distribute the class labels of the training set (with known class labels) and the second step is to use the resulting classifier to predict the class labels to the testing set(unknown class labels). For this first of all, KNN-classifier is trained by using the training dataset which includes separate positive, negative and neutral response files shown in figure 3.7, 3.8, 3.9.

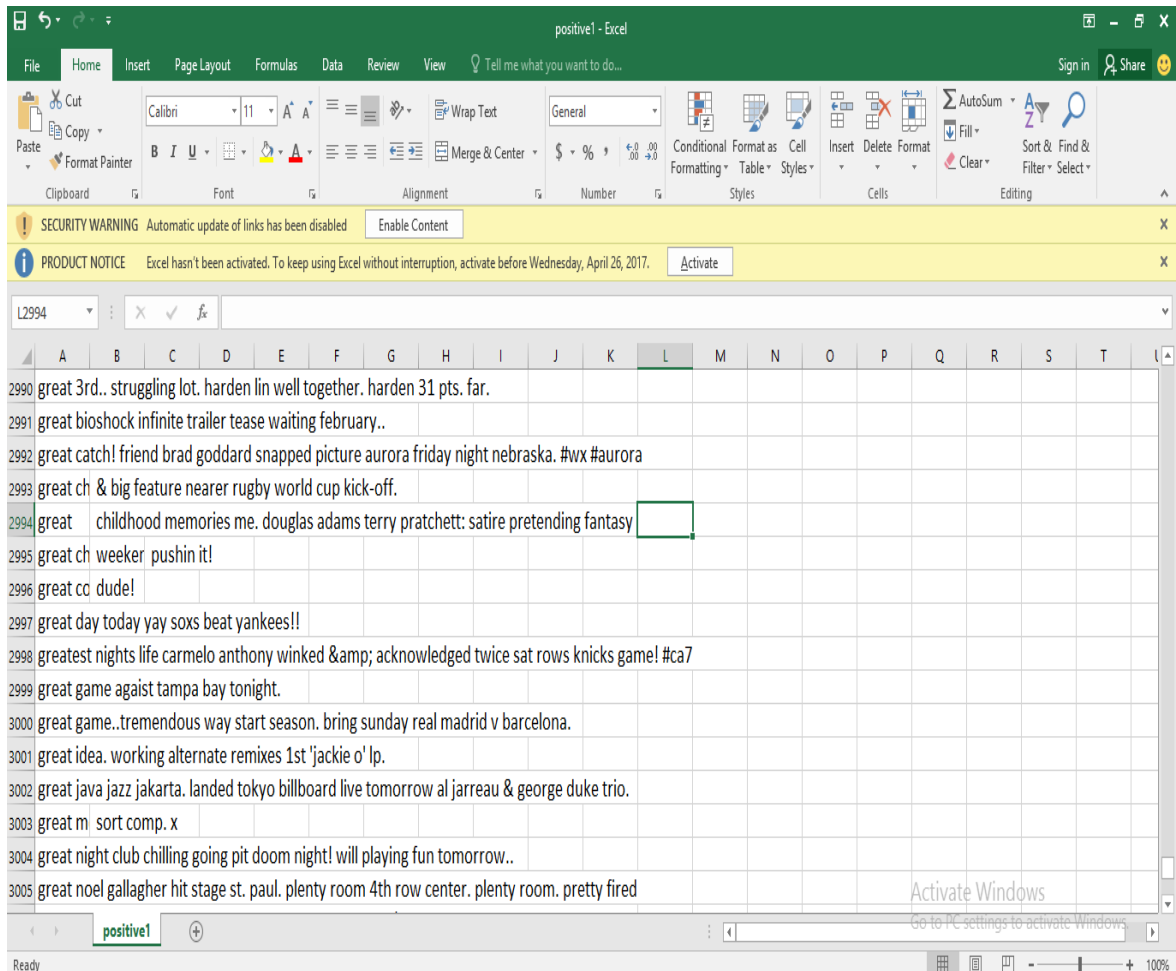


Figure 3.7 Positive response file for training of the classifier

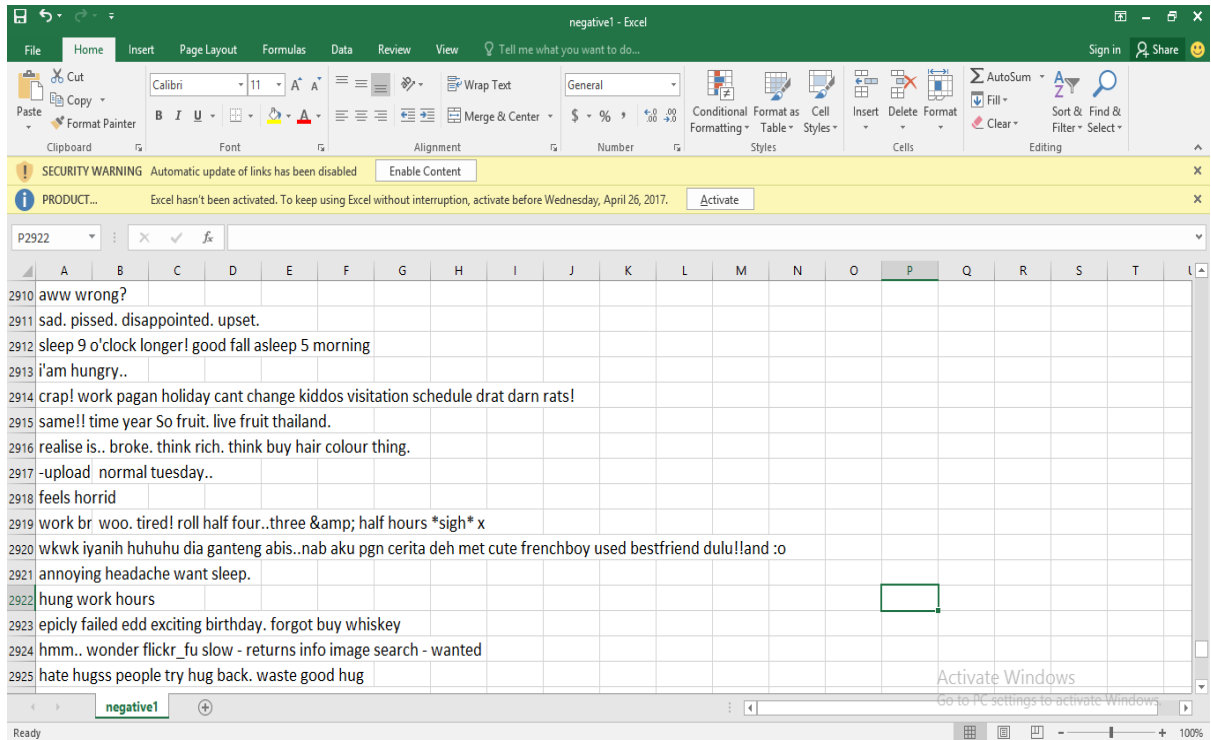


Figure 3.8 Negative response file for training of the classifier

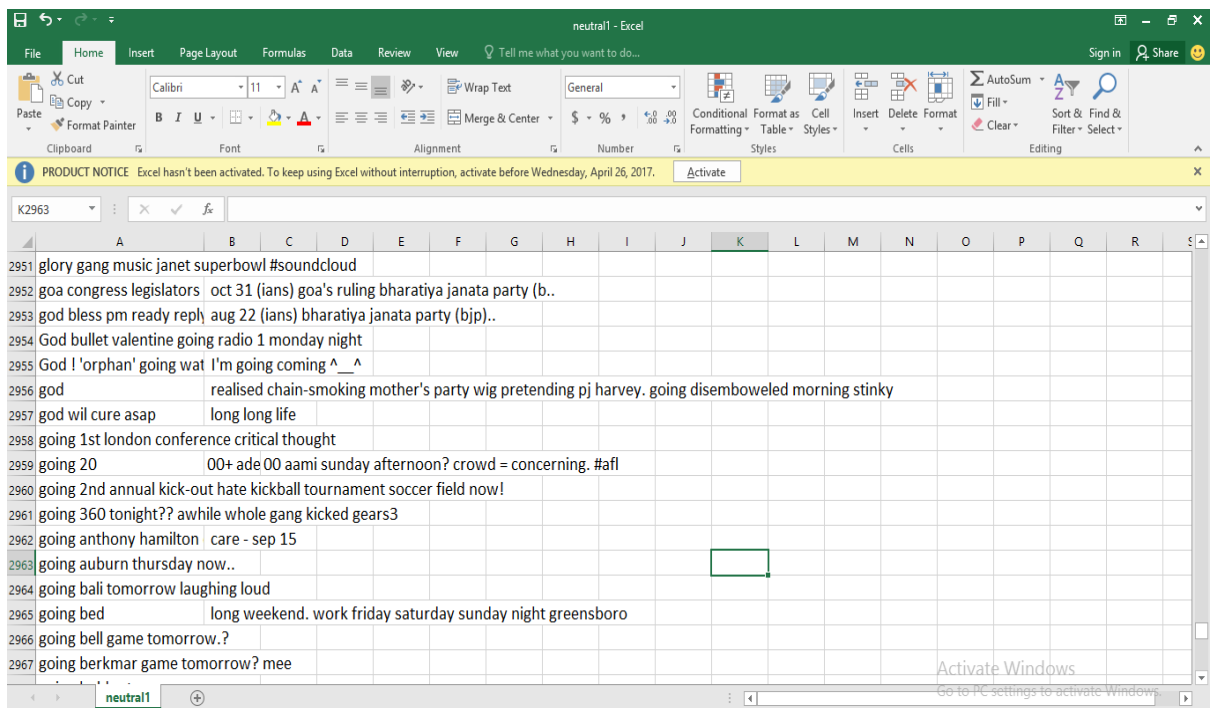


Figure 3.9 Neutral response file for training of the classifier

After training the classifier, class labels of the testing dataset are predicted whether the comments are positive, negative or neutral.

- a. If the computed weight is lesser than 0

Mark the message as negative

- b. If the weight is higher than 0

Mark the message as positive

- c. If the weight equals zero

Mark the message as neutral

Tool used: The programming language that is used for the implementation of the research work is python 2.7. Python is a high level programming language that is widely used for general-purpose programming. Python provides a Natural language toolkit (NLTK) to work upon the natural language text. In python, there is a build-in corpora and operations on the natural language can be easily performed just by importing the NLTK packages. Spyder is an integrated development environment (IDE) that provides development environment to develop the programs in python. It is open source, user friendly and easily installable software. Spyder is easily available for windows, Mac as well as Linux operating system. There are different types of the packages that are present by default in the python library and just need to import in the code. Some of the packages are as follows:

- i. **NLTK:** NLTK is one of the basic packages that is required to perform the operations on the natural language processing text.
- ii. **Sklearn:** It is also called scikit-learn that includes all the machine learning algorithms for the classification, clustering, regression etc.
- iii. **Numpy:** This package provide support for the array handling.
- iv. **Pandas:** This package provides the support for the data analysis and to read the Microsoft excel file.
- v. **Re:** re is the acronym of regular expression and this package provides the help for the pattern matching and for the text cleaning.

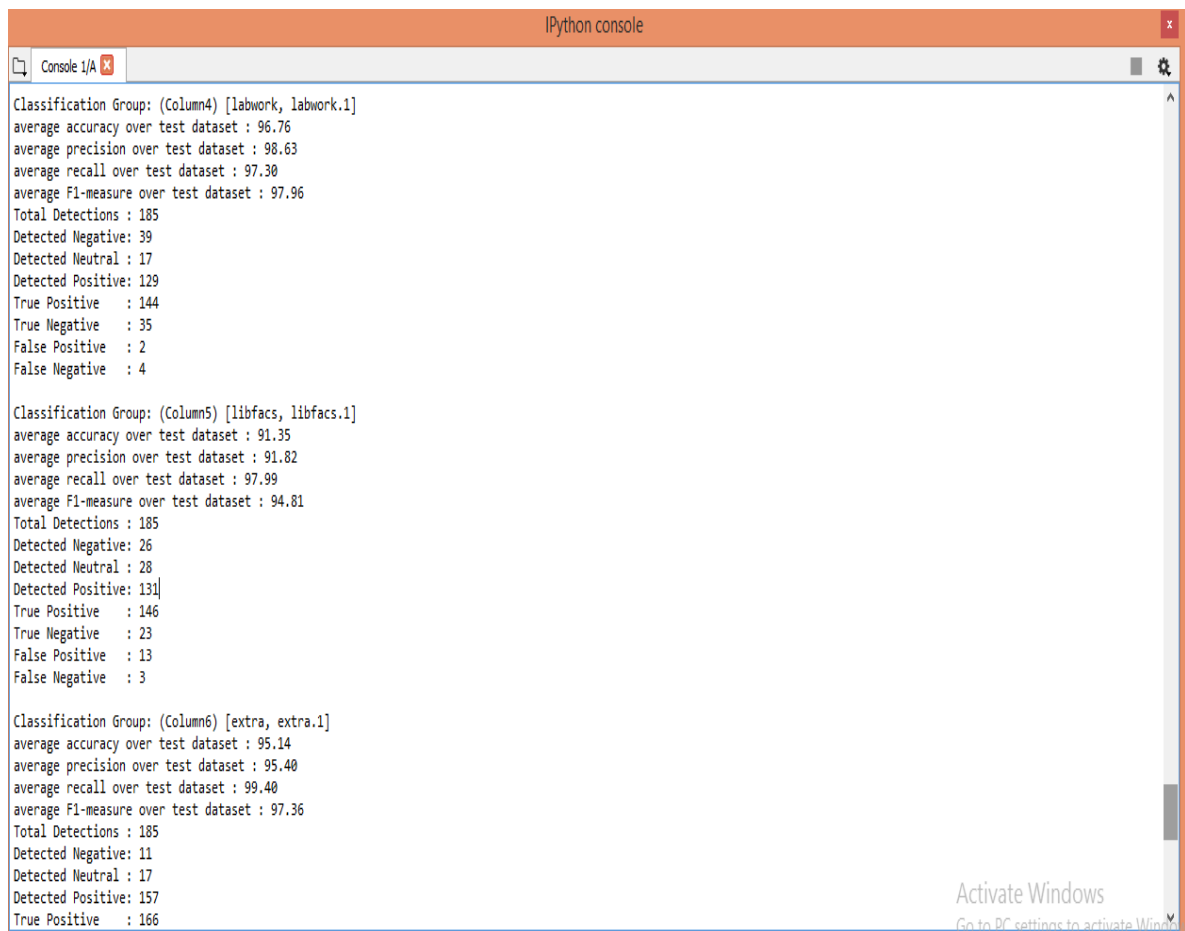
Chapter 3

RESULTS AND DISCUSSION

This Chapter deals with the final results of the proposed technique and comparison between proposed and existing technique

4.1 EXPERIMENTAL RESULTS

Figure 4.1 shows the output of the proposed approach in terms of performance parameters such as precision, recall, accuracy, F1 measure, positive, negative and neutral comments. Table 4.1 depicts the results of the emotion classification of the different features i.e. teaching, course content, examination, practical work, library facilities and extracurricular activities.



```
IPython console
Console 1/A
Classification Group: (Column4) [labwork, labwork.1]
average accuracy over test dataset : 96.76
average precision over test dataset : 98.63
average recall over test dataset : 97.30
average F1-measure over test dataset : 97.96
Total Detections : 185
Detected Negative: 39
Detected Neutral : 17
Detected Positive: 129
True Positive : 144
True Negative : 35
False Positive : 2
False Negative : 4

Classification Group: (Column5) [libfacs, libfacs.1]
average accuracy over test dataset : 91.35
average precision over test dataset : 91.82
average recall over test dataset : 97.99
average F1-measure over test dataset : 94.81
Total Detections : 185
Detected Negative: 26
Detected Neutral : 28
Detected Positive: 131
True Positive : 146
True Negative : 23
False Positive : 13
False Negative : 3

Classification Group: (Column6) [extra, extra.1]
average accuracy over test dataset : 95.14
average precision over test dataset : 95.40
average recall over test dataset : 99.40
average F1-measure over test dataset : 97.36
Total Detections : 185
Detected Negative: 11
Detected Neutral : 17
Detected Positive: 157
True Positive : 166
```

Figure 4.1 Output of the proposed approach

Table 4.1 Classification of different features

Feature	Positive	Negative	Neutral
Teaching	129	21	35
Course content	126	33	26
Examination	134	19	32
Practical work	129	39	17
Library facilities	131	26	28
Extracurricular activities	157	11	17

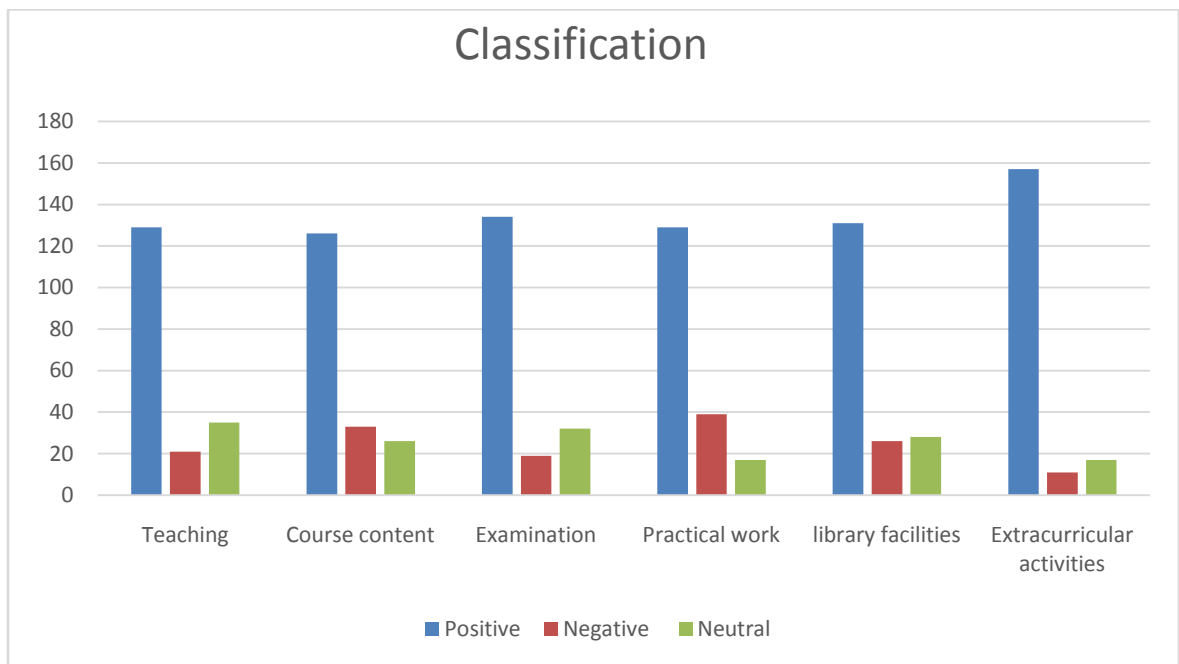


Figure 4.2 Classification of different features

Figure 4.2 shows the graph of positive, negative and neutral comments of the students on the different features of the university those are teaching, course content, examination, practical work, library facilities and extracurricular activities and it has been analyzed that students are also encouraged to give the negative comments on different features. From the results, it has been interpreted that number of the positive comments are more for the extracurricular activities held in the university and number of negative comments are more

for the practical work. Thus, we inferred that students are satisfied with the extracurricular activities and the university management needs to improve the practical work as it fetched most negative comments.

Table 4.2 Performance comparison of different features using KNN classifier

Feature	Accuracy	Precision	Recall	F1-measure
Teaching	87.57	92.68	93.25	92.97
Course content	91.89	94.74	95.36	95.05
Examination	87.03	89.76	95.51	92.55
Practical work	96.76	98.63	97.30	97.96
Library facilities	91.35	91.82	97.99	94.81
Extracurricular activities	95.14	95.40	99.40	97.36

Precision: It is determined by equation:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall: It is determined by equation:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Accuracy: Accuracy is determined by the equation:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

F1-measure: It is determined by the equation:

$$\text{F1-measure} = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN})$$

Where TP is true positive, FP is false positive, TN is true negative and FN is false negative.

True positive: Value is positive and predicted by the classifier is also positive.

False positive: Value is negative but predicted by the classifier is positive.

True Negative: Value is negative and predicted by the classifier is also negative.

False Negative: Value is positive but predicted by the classifier is negative.

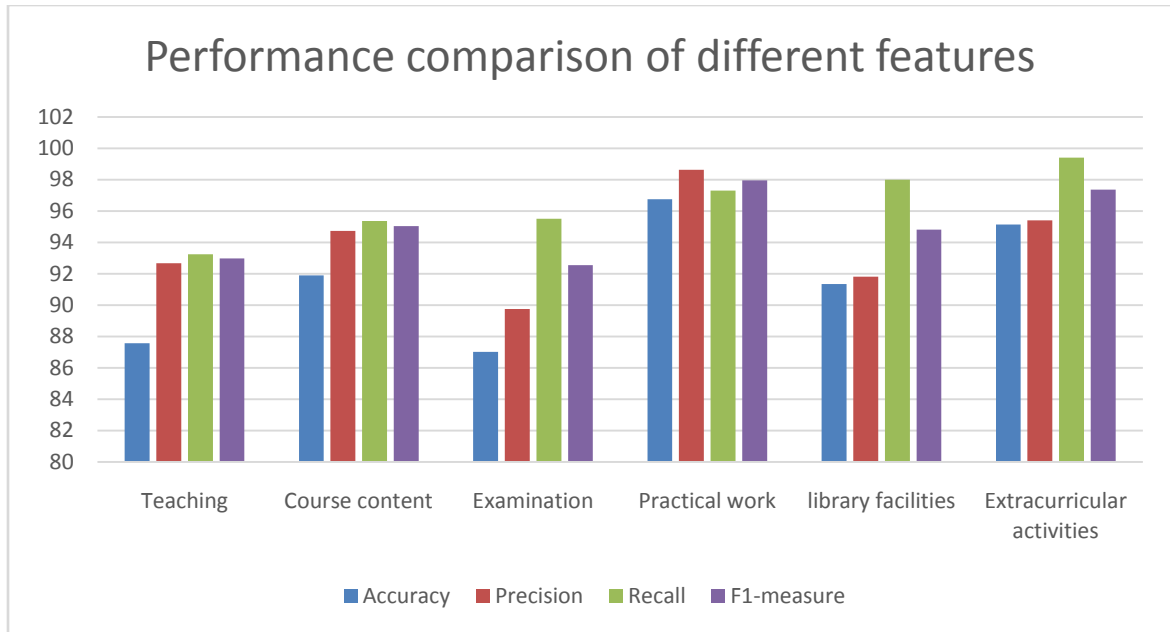


Figure 4.3 Performance comparison of different features using KNN classifier

On the basis of the value shown in table 4.2, figure 4.3 depicts the performance of KNN classifier on different features i.e. teaching, course content, examination, practical work, library facilities and extracurricular activities.

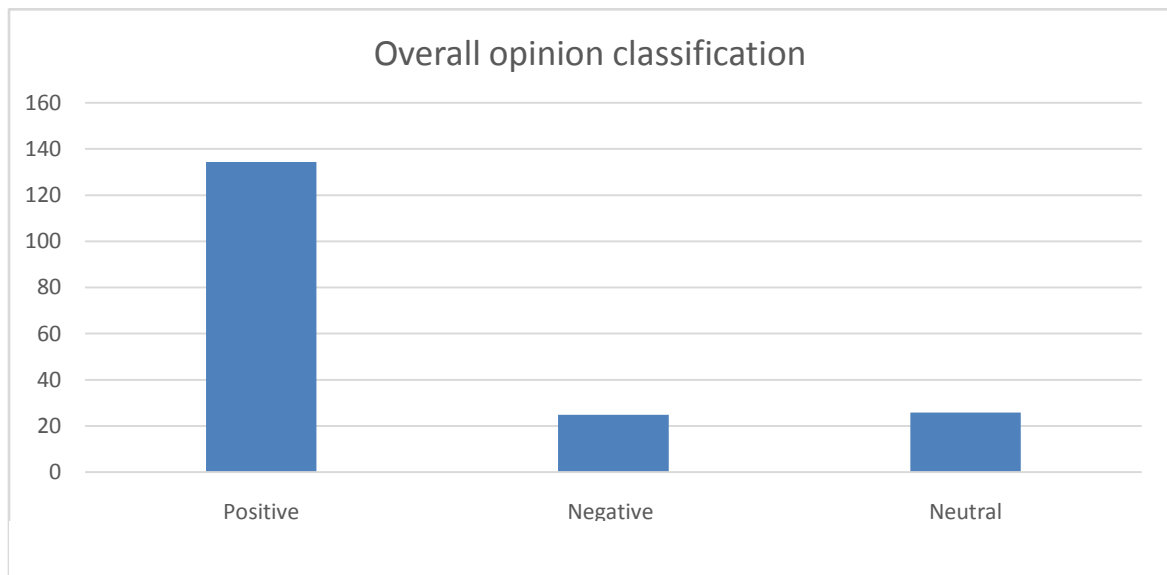
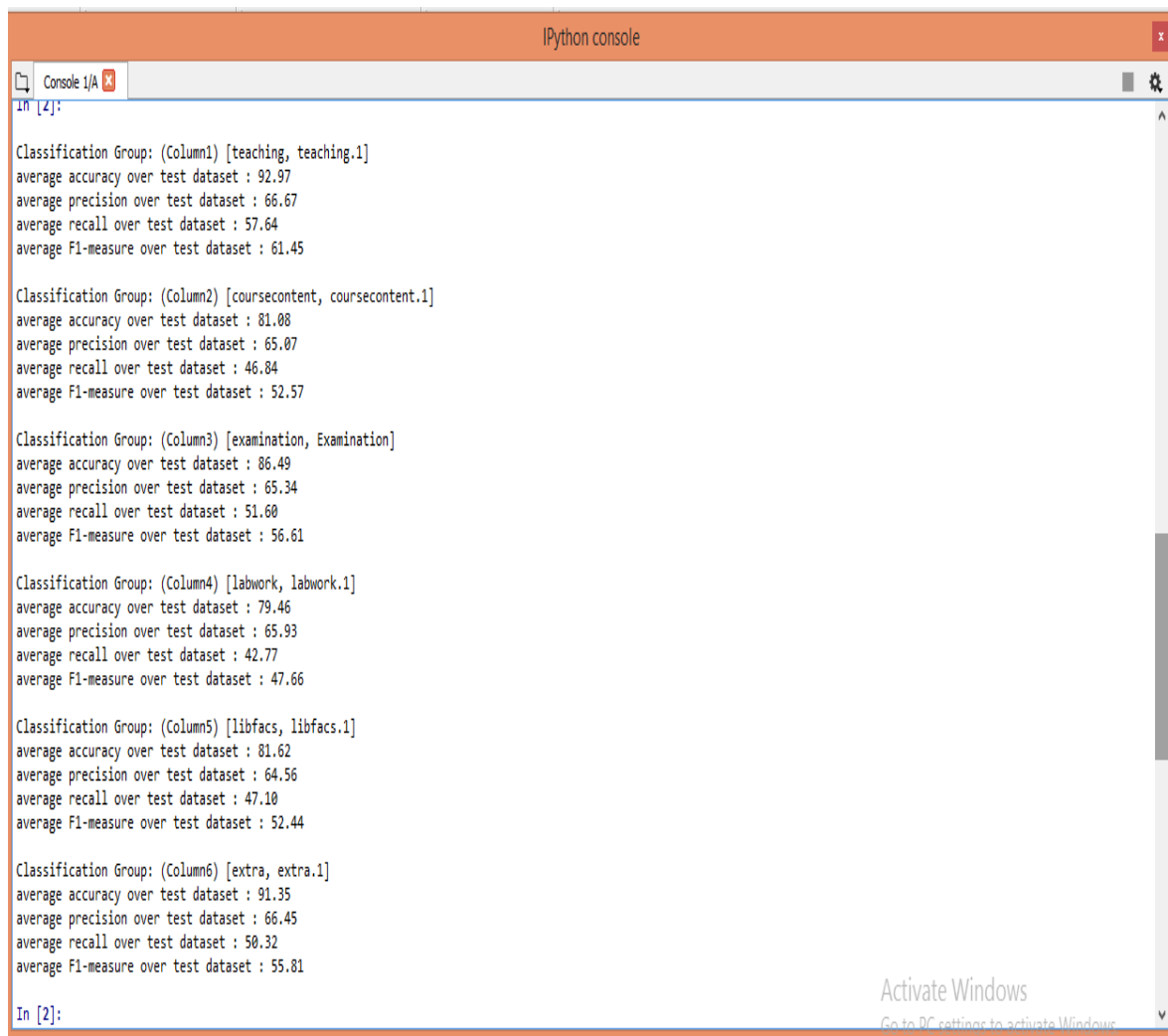


Figure 4.4 Overall Opinion classification

Figure 4.4 depicts the overall classification of the student comments into positive, negative and neutral category. This result shows that most of the students are having a good experience with the academic activities of the university.

4.2 COMPARISON WITH EXISTING TECHNIQUE.

This section compares the results of the proposed technique with the existing technique in terms of accuracy, precision, recall and f1-measure. Figure 4.5 and table 4.3 depicts the output of the existing approach in terms of performance parameters such as precision, recall, accuracy, F1 measure. It has been analyzed performance of the output has been increased by using the proposed approach.



```
Python console
Console 1/A
In [2]:
Classification Group: (Column1) [teaching, teaching.1]
average accuracy over test dataset : 92.97
average precision over test dataset : 66.67
average recall over test dataset : 57.64
average F1-measure over test dataset : 61.45

Classification Group: (Column2) [coursecontent, coursecontent.1]
average accuracy over test dataset : 81.08
average precision over test dataset : 65.07
average recall over test dataset : 46.84
average F1-measure over test dataset : 52.57

Classification Group: (Column3) [examination, Examination]
average accuracy over test dataset : 86.49
average precision over test dataset : 65.34
average recall over test dataset : 51.60
average F1-measure over test dataset : 56.61

Classification Group: (Column4) [labwork, labwork.1]
average accuracy over test dataset : 79.46
average precision over test dataset : 65.93
average recall over test dataset : 42.77
average F1-measure over test dataset : 47.66

Classification Group: (Column5) [libfacs, libfacs.1]
average accuracy over test dataset : 81.62
average precision over test dataset : 64.56
average recall over test dataset : 47.10
average F1-measure over test dataset : 52.44

Classification Group: (Column6) [extra, extra.1]
average accuracy over test dataset : 91.35
average precision over test dataset : 66.45
average recall over test dataset : 50.32
average F1-measure over test dataset : 55.81

In [2]:
Activate Windows
Go to PC settings to activate Windows
```

Figure 4.5 Output of existing base paper technique

Table 4.3 Performance comparison of the base paper technique

Feature	Accuracy	Precision	Recall	F1-measure
Teaching	92.97	66.67	57.64	61.45
Course content	81.08	65.07	46.84	52.57
Examination	86.49	65.34	51.60	56.61
Practical work	79.46	65.93	42.77	47.66
Library facilities	81.62	64.56	47.10	52.44
Extracurricular activities	91.35	66.45	50.32	55.81

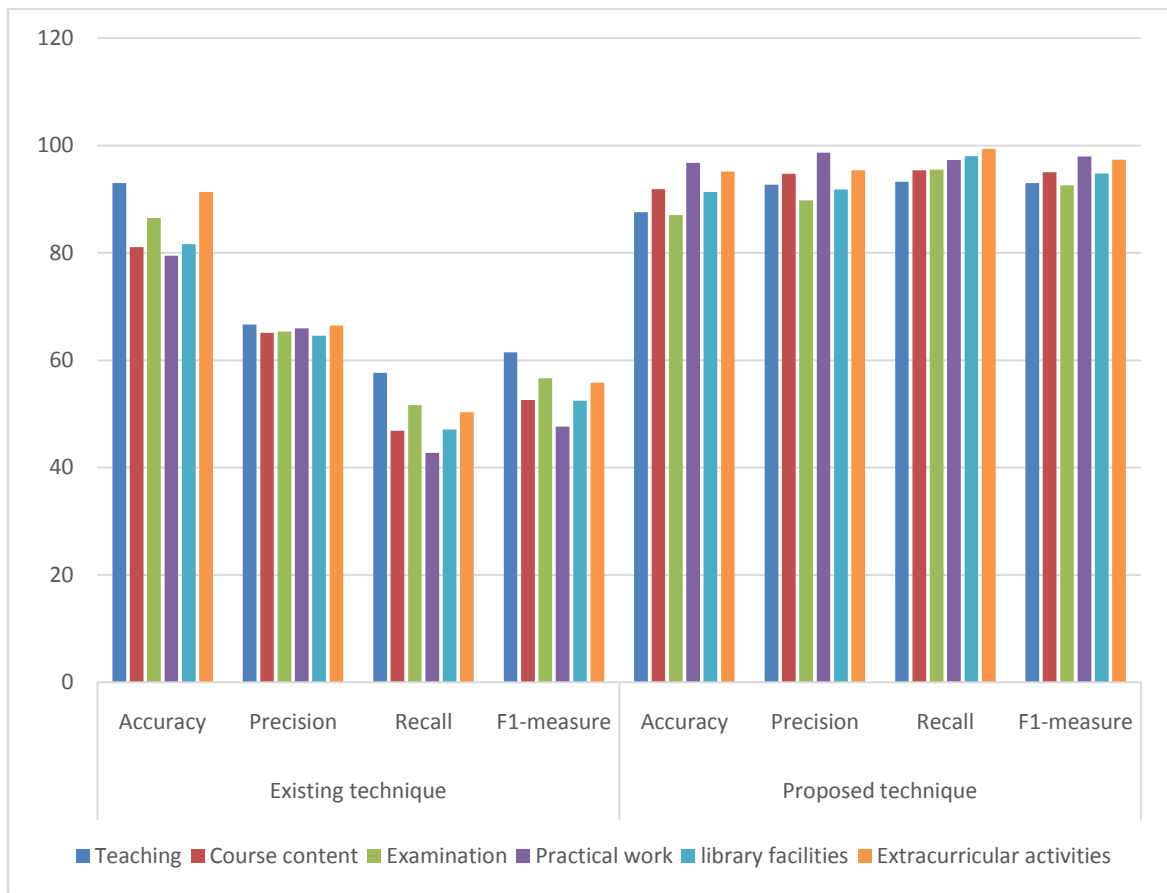


Figure 4.6 Performance comparison of proposed technique with existing technique

Figure 4.6 shows the performance comparison of different features of proposed technique with the different features of existing technique in terms of accuracy, precision, recall and f1-measure and figure 4.7 depicts the overall performance comparison of proposed technique with the existing technique and it has been analyzed that accuracy has been increased by 5 percent, precision increased by the 26 percent, recall increased by 47 percent and f1-measure increased by 41 percent. By increasing the performance means this technique will more accurately classifies the data into different categories using the KNN-classifier. Thus by enhancing some of the parameters of the existing technique, the performance of the technique has been improved successfully in terms of precision, recall, accuracy and f1 measure.

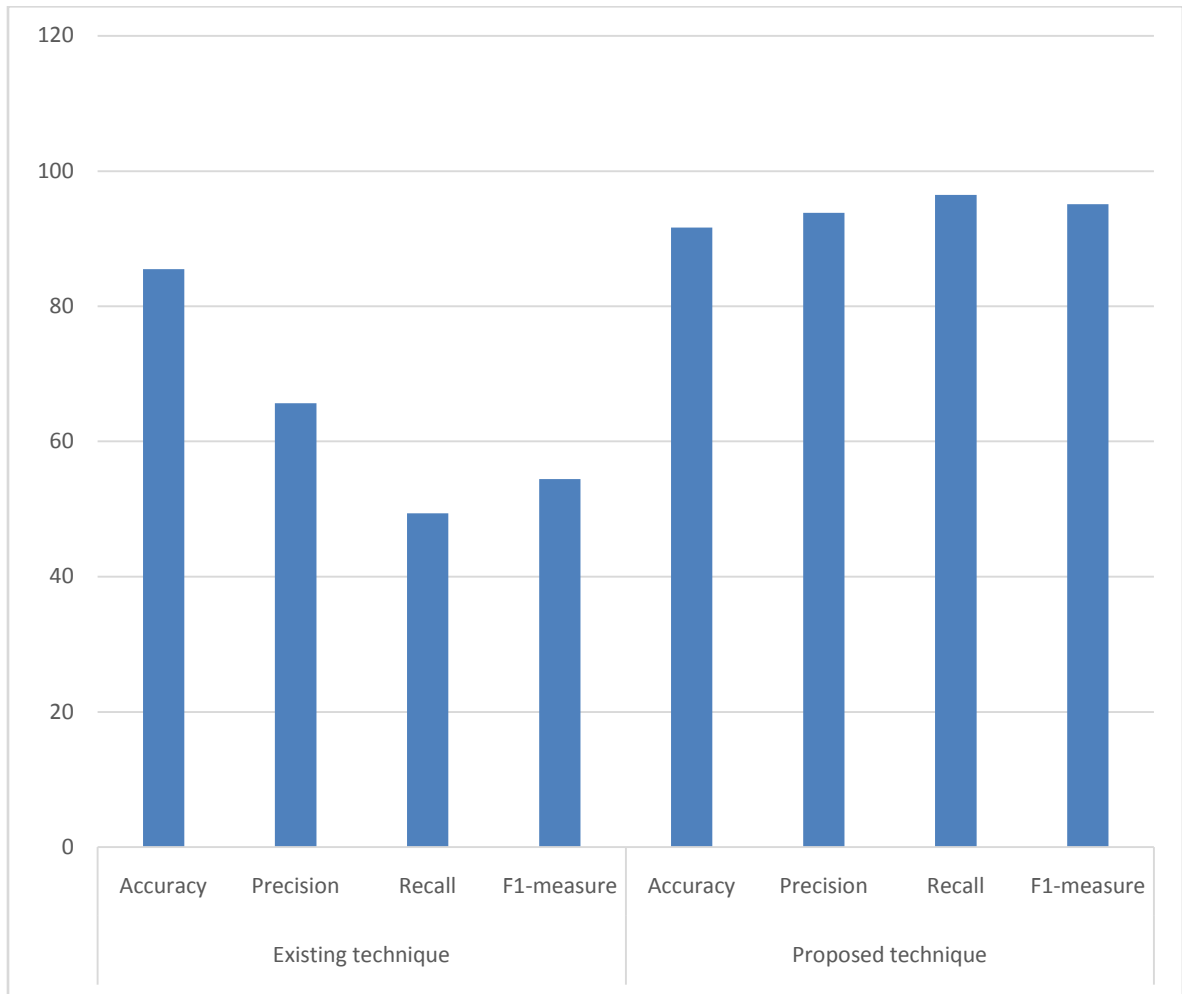


Figure 4.7 Overall performance Comparison

Chapter 4

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Emotion mining is the method of studying or detecting the viewpoints or perspective of the writer from any text and classifies the text into different categories according to the polarity of the text. Emotion mining has becoming an enthralling research area with the advancement of the web technology. Emotion mining has been utilized in many areas such as marketing, elections, e-commerce, education, movie reviews, hotel reviews etc. In this report, we have studied about the basic terms, terminologies, methods and the process of the emotion mining that how to extract the subjective knowledge from the text that will further help for the analysis of the data.

In this report, it has been shown how emotion mining is performed on the student feedback reviews using the machine learning algorithm to extract the subjective information from the text and to classify the data into positive, negative and neutral category. Analysis of student feedback reviews helps to understand the student feedback on academic curriculum more precisely. This will help the management of the university to know how much students are satisfied with the different features of the university those are teaching, course content, examination, practical work, library facilities and extracurricular activities. In this work, different feature extraction and machine learning algorithm is used to classify the reviews of students into positive, negative and neutral category.

5.2 FUTURE SCOPE

The main challenge in this research work is the textual nature of comments illustrated in natural language processing and it is very hard to differentiate between the subjective and objective information. In the future work, one can further make improvements to perform emotion mining on the feedback comments on the E-learning as well as may extract the features of the collective dataset automatically than predefined features by using the clustering technique.

REFERENCES

- [1] F. F. Balahadia, "Teacher' s Performance Evaluation Tool Using Opinion Mining with Sentiment Analysis," pp. 95–98, 2016.
- [2] G. a. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [3] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006.
- [4] B. R. Feldman, "Techniques and Applications for Sentiment Analysis." *Communications of the ACM*, pp.82-89, 2013
- [5] M. Yassine and H. Hajj, "A framework for emotion mining from text in online social networks," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1136–1142, 2010.
- [6] A. El-halees, "Mining Opinions in User-Generated Contents to Improve," pp. 107–115, 2011.
- [7] H. Shi and X. Li, "A SENTIMENT ANALYSIS MODEL FOR HOTEL REVIEWS BASED ON," pp. 10–13, 2011.
- [8] R. R. Kabra, "Performance Prediction of Engineering Students using Decision Trees," vol. 36, no. 11, pp. 8–12, 2011.
- [9] A. Rashid, "Feature Level Opinion Mining of Educational Student Feedback Data using Sequential Pattern Mining and Association ... Feature Level Opinion Mining of Educational Student Feedback Data using Sequential Pattern Mining and Association Rule Mining," no. November, 2013.
- [10] N. Altrabsheh, "Sentiment analysis : towards a tool for analysing real-time students feedback," 2014.
- [11] A. Kaur and V. Gupta, "N-gram Based Approach for Opinion Mining," pp. 81–88, 2014.
- [12] V. B. Raut, "Opinion Mining and Summarization of Hotel Reviews," 2014.
- [13] M. Bouazizi and T. Ohtsuki, "Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis," *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM '15*, pp. 1594–1597, 2015.
- [14] E. M. G. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs

- using Open Source Tools : An Empirical Study,” vol. 112, no. 5, pp. 44–48, 2015.
- [15] V. Jha, N. Manjunath, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, “HOMS: Hindi opinion mining system,” *2015 IEEE 2nd Int. Conf. Recent Trends Inf. Syst. ReTIS 2015 - Proc.*, pp. 366–371, 2015.
- [16] G. Tripathi, S. Naganna, G. Noida, and G. Noida, “FEATURE SELECTION AND CLASSIFICATION APPROACH FOR,” vol. 2, no. 2, 2015.
- [17] M. Arora, “A Framework for Informal Language : Opinion Mining,” 2015.
- [18] M. A. Rao, “Model for Improving Relevant Feature Extraction for Opinion Summarization,” pp. 1–5, 2015.
- [19] C. Chatterjee, K. Chakma, and C. Science, “A Comparison between Sentiment Analysis of Student Feedback at Sentence Level and at Token Level,” vol. 4, no. 3, pp. 482–486, 2015.
- [20] A. Kumar, “Sentiment Analysis and Feedback Evaluation,” pp. 433–436, 2015.
- [21] G. I. Nitin, “Analyzing Educational Comments for Topics and Sentiments : A Text Analytics Approach,” 2015.
- [22] A. Yang, J. Zhang, L. Pan, and Y. Xiang, “Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination,” 2015.
- [23] S. Ahmed, “A Novel Approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using Web Data,” pp. 0–4, 2015.
- [24] F. Koto and M. Adriani, “The Use of POS Sequence for Analyzing Sentence Pattern in Twitter Sentiment Analysis,” 2015.
- [25] Rao, Ashwini, and Ketan Shah. "An optimized rule based approach to extract relevant features for sentiment mining." In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, pp. 2330–2336, 2016.
- [26] M. Antony, N. Johny, V. James, and A. Wilson, “PRODUCT RATING USING SENTIMENT ANALYSIS,” pp. 3458–3462, 2016.
- [27] J. Islam, Z. A. Badhon, and P. C. Shill, “An Effective Approach of Intrinsic and Extrinsic Domain Relevance Technique for Feature Extraction in Opinion Mining,” pp. 428–433, 2016.
- [28] T. P. Sahu, “Sentiment Analysis of Movie Reviews : A study on Feature Selection & Classification Algorithms,” 2016.

- [29] V. Dhanalakshmi and D. Bino, “supervised learning algorithms,” pp. 1–5, 2016.
- [30] Bakshi, Rushlene Kaur, et al. "Opinion mining and sentiment analysis." *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on.* IEEE, 2016.
- [31] E. Deepak, G. S. Pooja, R. N. S. Jyothi, S. V. P. Kumar, and K. V Kishore, “SVM Kernel based Predictive Analytics on Faculty Performance Evaluation.”
- [32] “Correlation of Feature Score to Overall Sentiment Score for Identifying The Promising Features,” pp. 9–13, 2016.
- [33] P. Kumar, K. Manocha, S.-B. Tech, H. Gupta, and S.-B. Tech, “ENTERPRISE ANALYSIS THROUGH OPINION,” pp. 3318–3323, 2016.
- [34] N. Akkarapatty, “A Machine Learning Approach for Classification of Sentence Polarity,” pp. 316–321, 2016.
- [35] Ullah, Mohammad Aman. "Sentiment analysis of students feedback: A study towards optimal tools." *Computational Intelligence (IWCI), International Workshop on.* IEEE, 2016.
- [36] R. Menaha, “Student Feedback Mining System Using Sentiment Analysis,” vol. 6, no. 1, pp. 51–55, 2017.