ABSTRACT

Recently, social media is playing a vital role in social networking and sharing of data. Social media is favored by many users as it is available to millions of people without any limitations to share their opinions, educational learning experience and concerns via their status. Twitter API, twitter4j, is processed to search for the tweets based on the geo location.

Student’s posts on social network gives us a better concern to take decision about the particular education system’s learning process of the system. Evaluating such data in social network is quite a challenging process. In the proposed system, there will be a workflow to mine the data which integrates both qualitative analysis and large scale data mining technique. Based on the different prominent themes tweets will be categorized into different groups.

Naïve Bayes classifier will be implemented on mined data for qualitative analysis purpose to get the deeper understanding of the data. It uses multi label classification technique as each label falls into different categories and all the attributes are independent to each other. Label based measures will be taken to analyze the results and comparing them with the existing sentiment analysis technique.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................ ii

LIST OF FIGURES ............................................................................................................ v

1 BACKGROUND AND RATIONALE ................................................................. 1
   1.1 Existing System ................................................................................................. 2
   1.2 Related Work ..................................................................................................... 2
   1.3 Introduction to Proposed System ................................................................. 3

2 NARRATIVE .............................................................................................................. 5
   2.1 Problem Statement ........................................................................................... 5
   2.2 Motivation .......................................................................................................... 5
   2.3 Project Objective .............................................................................................. 6
   2.4 Project Functionalities ..................................................................................... 6

3 PROPOSED SYSTEM DESIGN ........................................................................ 7
   3.1 System Design and Architecture .................................................................. 8
      3.1.1 System Authentication ....................................................................... 9
      3.1.2 Creating Twitter Application ........................................................ 9
   3.2 Data Collection .............................................................................................. 10
   3.3 Data Pre-processing ....................................................................................... 11
   3.4 Prominent Themes ......................................................................................... 11
   3.5 Bayes Rule ..................................................................................................... 12
   3.6 Naïve Bayes Classifier ................................................................................. 13

4 IMPLEMENTATION ............................................................................................... 14
   4.1 Inductive Content Analysis ........................................................................ 14
   4.2 Authentication ............................................................................................... 17
4.2.1 Login ........................................................................................................... 17
4.2.2 Registration ................................................................................................. 18
4.3 Collection of the tweets: ........................................................................................ 19
4.4 Location function .................................................................................................. 21
4.5 Tweet Pre-processing: .......................................................................................... 22
4.6 Remove Duplicates ................................................................................................ 23
4.7 Classification and Analysis .................................................................................... 24
4.8 Class Maximum Posterior Probability: ................................................................. 25
4.9 Label-Based Evaluation Measures: .............................................................. 26
5 REQUIREMENTS .................................................................................................... 28
5.1 Environment ........................................................................................................... 28
5.2 System Requirements............................................................................................. 28
5.2.1 Software Requirements ............................................................................... 29
5.2.2 Hardware requirements ............................................................................... 29
5.3 Java Swing ............................................................................................................. 30
5.4 Xamp ...................................................................................................................... 30
5.5 Netbeans ide ........................................................................................................... 30
6 TESTING AND RESULTS....................................................................................... 31
6.1 Test Case1: Tweet Collection with Search ID ....................................................... 32
6.2 Test Case2: Tweet collection with geo location .................................................... 33
6.3 Test Case3: Preprocessing the tweet ...................................................................... 34
6.4 Test Case4: Duplicate Removal ............................................................................. 35
6.5 Tweet Analysis- Multi Label Classifier .................................................................. 36
7 CONCLUSION ......................................................................................................... 41
8 FUTURE WORK ...................................................................................................... 42
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Social Media Impact [1]</td>
<td>1</td>
</tr>
<tr>
<td>3.1</td>
<td>Data Flow Diagram</td>
<td>8</td>
</tr>
<tr>
<td>3.2</td>
<td>Twitter4j Integration to Application [20]</td>
<td>9</td>
</tr>
<tr>
<td>3.3</td>
<td>Twitter Application</td>
<td>10</td>
</tr>
<tr>
<td>3.4</td>
<td>Bayesian Network for Naïve Bayes [18]</td>
<td>13</td>
</tr>
<tr>
<td>4.1</td>
<td>Twitter App Creation</td>
<td>14</td>
</tr>
<tr>
<td>4.2</td>
<td>Details of the Application</td>
<td>15</td>
</tr>
<tr>
<td>4.3</td>
<td>Xampp Control Panel</td>
<td>16</td>
</tr>
<tr>
<td>4.4</td>
<td>Login</td>
<td>17</td>
</tr>
<tr>
<td>4.5</td>
<td>Registration</td>
<td>18</td>
</tr>
<tr>
<td>4.6</td>
<td>User Interface Mainframe</td>
<td>19</td>
</tr>
<tr>
<td>4.7</td>
<td>Collected Tweets</td>
<td>20</td>
</tr>
<tr>
<td>4.8</td>
<td>Categories of Tweets for Different Sections</td>
<td>21</td>
</tr>
<tr>
<td>4.9</td>
<td>Geo Location of Tweet</td>
<td>22</td>
</tr>
<tr>
<td>4.10</td>
<td>Tweet Pre-Processing</td>
<td>23</td>
</tr>
<tr>
<td>4.11</td>
<td>Remove Duplicates</td>
<td>24</td>
</tr>
<tr>
<td>6.1</td>
<td>Tweet Collection Based on Hashtag</td>
<td>32</td>
</tr>
<tr>
<td>6.2</td>
<td>Turning off the Location in Twitter</td>
<td>33</td>
</tr>
<tr>
<td>6.3</td>
<td>Result for Location Turnoff</td>
<td>34</td>
</tr>
<tr>
<td>6.4</td>
<td>Preprocess the Tweet</td>
<td>35</td>
</tr>
<tr>
<td>6.5</td>
<td>Duplicate Removing Test</td>
<td>36</td>
</tr>
<tr>
<td>6.6</td>
<td>Results from the Naïve Bayes Classifier</td>
<td>37</td>
</tr>
<tr>
<td>6.7</td>
<td>Label Based Analysis</td>
<td>38</td>
</tr>
</tbody>
</table>
Figure 6.8 Label Based Measures........................................................................................................... 39

Figure 6.9 Comparison Result ................................................................................................................. 40
1 BACKGROUND AND RATIONALE

Social media has become one of the comfortable medium for people to share their feelings instinctively. According to the survey sharing of the data is high in the social sites like twitter and facebook [1]. Students pay more attention to share their feelings spontaneously in a relaxed, informal environment more than the formal classroom environment. An information educational researcher can easily understand students experience from outside of the class and they can get full details about the education system.

It is very helpful for an institution to understand the difficulties of the student he/she facing in the learning system. However this data or information is used to upgrade the education system of the particular institution. These social network data mining provide an opportunity to make changes in education system ultimately to make an impact on economic growth as students play a vital role in the future workforce.

Social Media Sharing on Web

Figure 1.1 Social Media Impact [1]
1.1 Existing System

There are so many traditional methods available such as questionnaires, surveys and interviews to analyze the student’s learning barriers in an educational institution. But the main problem with these methods is these techniques are time consuming and cannot be performed efficiently with high frequency as the analysis has to be performed manually. One more important problem is the data collected in formal manner may not be authentic as the student may not convey what they feel properly when compared to an informal medium like social media.

In the existing projects sentiment analysis is one kind of approach followed for opinion mining which comes under three class classification technique, classifies the results as positive, negative and neutral. For understanding the tweet in a deeper manner this sentiment analysis[6] is not sufficient. So there is a need for analysis of social media data qualitatively with the help of integrating both the workflow and algorithmic approach.

1.2 Related Work

- **Public Discourse in an Informal Environment:**

According to Goffman’s Theory [2] people will be themselves at heart and say whatever they feel at that particular moment with their own style of spontaneous actions when they are back-stage. The same theory is applied to social media where a student can post in an informal environment whatever he/she feels bravely without thinking what others
may feel, and can say the things which he was not able to express in the formal class room surroundings. So the posts will be more authentic to proceed further.

- **Mining Twitter data:**
  Analysis of tweets was previously done based on histogram that is at what point the person becomes more active [3]. So previously based on the requirement different algorithms were implemented to perform linguistic, content and network analysis.

- **Learning Analytics - Educational Data Mining:**
  Previously, researchers performed educational mining, which tells how much time the student spent in virtual learning like blackboard etc. The performance data will be mined based on the categories like time spent in discussion forums and the grades acquired by them [4].

1.3 **Introduction to Proposed System**

The proposed system is different from the existing system in terms of analysis technique of data. The proposed system has to perform the qualitative analysis using classification algorithm instead of sentiment analysis. Sentiment analysis considers the opinion of the user about a system or product and categorizes it to neutral, negative or positive mood [5] [6]. In the proposed system, searching the information based on the keywords such as engineer, students, campus, class, professor and lab in the twitter data as per the geo location, keyword and search id.
Searching keywords in twitter is one of the hardest tasks because of the diversity of the language and the Internet slang used is different.

A multi label classification algorithm is implemented to analyze content as per the category and the results will be reported to a policy maker which helps the person to get an overview of student’s problems and their experience in learning, so that an institution can make a proper decision making policy to improve the performance of the students as they are the future of society.
2 NARRATIVE

2.1 Problem Statement

Methods used to analyze the data include surveys, interviews, questionnaires, classroom activities about the student educational experiences and problems they are facing. But these traditional methods are time consuming and very limited in scale. The manual analysis does not make sense of analyzing student learning experiences which are huge in volume with different Internet slang and the timing of the student posting on the web. The sentiment analysis [5][6] of the tweets does not cover much relevant experience because even for a human judge to determine what student problems a tweet indicates is a more complicated task than to determine just the sentiment of a tweet.

2.2 Motivation

An educational institution always tries to get better day by day and make their students feel comfortable and safe. They also want to know student’s opinion about the education system and the learning problems they are facing. Social media provides a better environment for sharing user experience in an interactive and informal way so most of the students show interest to post content about what they really feel.

Some interesting methods are used to collect data from the social media like Twitter, Facebook. Collecting the data from these conversations is authentic for analysis purpose because the majority of the posts by students will be instinctive. Pure manual
analysis and automatic analysis cannot deal with the growing scale of data. Thus the need for the qualitative analysis is increasing.

2.3 Project Objective

The main objective of the proposed system is to design a workflow of social media data analysis for student learning problems and experience and applying the mining classification algorithm to get the results which will be useful for policy makers to take proper decisions to improve the education system of the institution.

2.4 Project Functionalities

This system provides a more user friendly functionalities and attractive user interface, some of the functionalities are listed below,

- Connectivity between social media,
- Collecting data from social media,
- Identifying keywords,
- Classification Algorithm
- Clustering the data.
3 PROPOSED SYSTEM DESIGN

As the twitter content is very concise and provides free APIs it will be easy to mine using twitter4j [6] and analyze the data. The proposed system is suggested to implementing each module of the architecture presented below. At first the data will be collected from large volumes of dataset [26], later Inductive content analysis will be performed on it once the Data sampling and data analysis takes place in the data flow. In model training and evaluation the multi level classifier [7] is implemented upon some prominent theme categories. Results will help the policy makers to understand the learning problems of students also to take proper decisions to overcome them.

The login system is designed to begin the process to fetch the data and to mine as shown in Figure 3.1.Once the data is collected first pre-processing of the text should be done by removing the repeated letters, non lettered symbols, hash tags etc. Once the data is preprocessed Naive Bayes Multi-Label Classifier is implemented to analyze the data and later on evaluating the measures for classifier in terms of performance. The words here are mutually exclusive that is independent of each other’s category.
3.1 System Design and Architecture

System architecture defines the structure and behavior of a system. Authorized users after the verification process, enters into main system and collects the data and performing qualitative analysis that is naïve bayes classifier after pre processing and removing the duplicates the data is supposed to evaluate the label based measures as shown in the Figure 3.1.

Figure 3.1 Data Flow Diagram
3.1.1 System Authentication

Registered user should have to use their username and the password to get login into the system. If the user is not registered, user can register an account to get login into the system. After passing the system authentication, then only user can see the main frame of the system.

3.1.2 Creating Twitter Application

Millions of opinions will be tweeted daily on twitter. These tweets can be collected using API (Application program interface) provided by twitter. Using the Twitter APIs [16][6] twitter4j the integration of the project to twitter is done and the data in the form of tweets are collected as shown in Figure 3.2.

![Twitter4j Integration to Application](image-url)

Figure 3.2 Twitter4j Integration to Application [20]
To create a twitter application, user have to login into the twitter account and by visiting the website https://apps.twitter.com, user can create a twitter application as shown in the Figure 3.3. After creating the twitter application, secret tokens will be generated. The secret tokens will be used to collect tweets from the twitter. Twitter will return Consumer Key and Consumer Secret for application which acts as credentials for the application and later on to update the settings. Twitter4j jar file should be placed in the class path.

![Twitter Application](image)

**Figure 3.3 Twitter Application**

### 3.2 Data Collection

Tweets will be collected from the account by using the secret tokens of the twitter application. Twitter authenticates the secret tokens and allows the user to access the twitter to collect the tweets. From this text mining [25], the list of keywords that have to be used to collect the tweets will be saved in the database and that keywords will be
retrieved from the database to query data from the twitter. Tweets basically will be selected based on the university hash tag, keyword hash tag and geo location.

3.3 Data Pre-processing

The data in the form of tweets will be saved to a text file along with the database. After that, preprocessing of the text should be done before using classifier. Different cases followed to filter the data are described below [27].

- All the hashtags will be removed that is # sign.
- Negative words with negative emotions will be identified and the words ending with “n’t” and other words like none, cannot will be replaced with “negtoken”, because negative emotions are useful for the analysis purpose later on.
- Punctuations and symbols or special characters have to be removed.
- RT which is used to retweet and all the hyperlinks will be removed.
- Compress the duplicate letters if they appear more than twice.

3.4 Prominent Themes

Various types of emotions related to learning experiences are termed as prominent themes. Below are the few themes related to the project [27],

- Heavy Study Load: Students who face problems which are dominated by labs, exams, homework and classes will be grouped on to heavy study load category.
Lack of Social Engagement: Students feel that they give up their freedom and joy time for the sake of the academic works.

Negative Emotion: Expressing the anger, sickness, depression and disappointment will come under negative emotions.

Sleep Problems: Students frequently suffer from lack of sleep and nightmares due to heavy study load and stress.

Diversity Issues: Anti social image of mingling with others is termed as the diversity issue. Without knowing the background all the negative comments will passed on which may cause the diversity issue.

Others-The Long Tail: Other tweets in this category do reflect various large number of less common problems or noisy tweets appearing in very low frequency.

3.5 Bayes Rule

Influence of one event’s occurrence on the probability of another event is known as conditional probability [17]. From the probability theory, Bayes theorem, allows for the calculation of conditional probability. Usually in data mining Bayes theorem will be used to decide among alternate hypothesis.

Bayes' theorem formula for the conditional probability of A given B, is as follows,

\[ P\left(\frac{A}{B}\right) = P\left(\frac{B}{A}\right) \times \frac{P(A)}{P(B)} \]

Where P (A) is prior probability P (A/B) is the posterior probability of A given B [22].
3.6 Naïve Bayes Classifier

If there is more than one attribute best classifier to use is naïve bayes because all the attributes coming from a class are independent of each other as they are mined from twitter. Simple Bayesian network can be denoted as naïve Bayesian classification assuming all the attributes are independent to each other when represented in Bayesian network [12]. Naïve bayes classifier’s structure is shown in Figure 3.4 below where the network consists only one parent and various child nodes, considering the parent node stating which child node each object in the database should belong to.

![Bayesian Network for Naïve Bayes](image)

**Figure 3.4 Bayesian Network for Naïve Bayes [18]**

Every attribute from x1…xn is independent of each other given a class label. The effect of an attribute on a class has no influence over other attributes as all of them are independent [11]. First the prior probability is determined by checking frequency of each label in the training set, to get the likelihood for label. The posterior probability is calculated for all classes, and the class with the highest probability will be the instance’s label which will be seen in the implementation phase in next section.
4 IMPLEMENTATION

4.1 Inductive Content Analysis

Once the developer login to the twitter to get access to the java project first they are supposed to create an application as shown in the Figure 4.1. Configuration Builder class is one of the class available in twitter4j jar collection, which is responsible to collect the data from twitter by creating an interface between the application and twitter.

Figure 4.1 Twitter App Creation
The application will be successfully created and returns access tokens to the developer as shown in the Figure 4.2. Parameters of Authentication required to configure Twitter4j:

- Consumer Key
- Consumer Secret
- Access Token
- Access Token Secret

Figure 4.2 Details of the Application
Each tweet can fall into multi labeled classification that is multiple categories of emotions. As all the categories are mutually exclusive the harmonic mean will be calculated between two label sets to determine the closeness between them.

Xampp\[9\] control panel includes Apache, MySQL and other installed components which need to be started before running the application as shown in the Figure 4.3

![Xampp Control Panel](image)

**Figure 4.3 Xampp Control Panel**

In MySQL admin all the databases related to the project are stored, viewed and can be fetched as per the classification technique need.
4.2 Authentication

To prevent the unauthorized users from accessing the application there is an authentication system to allow only authorized users who are registered to access the application.

4.2.1 Login

If the person is an existing user then he will get the access to the project using existing credentials with which he registered. Database is supposed to validate the username and password of the user as shown in the Figure 4.4.

Figure 4.4  Login
4.2.2 Registration

If the user is new to the system then he/she is supposed to enter the details accordingly as per Figure 4.5. All the fields will be validated and the data will be stored in database.

Figure 4.5 Registration

Once the login into the system is done, it will allow the user to do more operations on twitter data to proceed further with analysis. It will provide the user interface along with the menu bar to perform operations like collection, pre-processing and analysis of the data as shown in the Figure 4.6.
4.3 Collection of the tweets:

Tweets will be collected from the twitter account by using the keyword, hash tag and geo location of the university. Based on the keyword highest likelihood in the category of trained set, the tweets will be collected as shown in Figure 4.7 shown below.
The database will check for three categories that is based on geo location, keyword and search id all the tweets will be fetched under defined categories in database that is training set. Categories like student, campus, engineering problems, college problems, sleep problems and staff will be fetched accordingly as shown in the below Figure 4.8.

**Figure 4.7 Collected Tweets**
4.4 Location function

Based on the geo location, according the longitude and latitude of the location the outputs will be generated under search id according to the location from which the person access the twitter as shown in Figure 4.9.
4.5 Tweet Pre-processing:

Tweet pre-process will remove all the unwanted data from the raw data and train the data for the further work. The following will be the sample output, removing all @, hyperlinks, RT and # symbols from the raw data as shown in the Figure 4.8 will be resulted as shown in the Figure 4.10.
Figure 4.10  Tweet Pre-Processing

4.6 Remove Duplicates

The tweets will have the duplicated words which have to be avoided for further process that is words with more than two continuous letters will be wrapped into single word with the proper meaning. The output will be as shown in the Figure 4.11.
4.7 Classification and Analysis

Let’s consider each tweet as a document here, $V = \{p_1, p_2, \ldots, p_n\}$, there are $p_n$ words, Category $C = \{c_1, c_2, \ldots, c_m\}$, where there are $c_m$ categories in it. Based on the above information, probability of particular word in the category $c$ is,
\[ \rho(p|n|c) = \frac{\text{Mpn}c}{\sum_{k} \text{Mpn}k} \] [27], where, MpnC is the number of words in the category that is number of words divided by total number of words gives the probability of a particular word will be determined.

4.8 Class Maximum Posterior Probability:

Maximum posterior probability is the probability of occurrence of an event given some external event has occurred. Probability will be calculated for each class of the training set and the class with highest posterior probability will be termed as the class label.

\[
\text{C Maximum posterior} = \arg\max \Pr \left[ \frac{c}{\text{words}1} \right] = \arg\max \Pr \left[ \frac{\text{words}1}{c} \right] \Pr [c]
\] [23]

If there are bunch of words of single category break them as independent events. Consider a tweet from twitter. To know which category it belongs to, look for the text in a single category and multiply with the category frequency as shown in the above equation that is probability of word frequency in a category should be multiplied with the category frequency for example if there are 4 words in a filtered tweet “heavy work, can’t study”, then to find the probability of category in which the tweet falls, first find the probability of a particular category out of six prominent themes hat is 0.6 and multiply it with the probability of a word to be occurred in that particular category based on the local trained data set that is 0.25.
4.9 Label-Based Evaluation Measures:

From the categorized tweets, the system will display the tweets and the categorized section. User will be intimated whether it is true positive, true negative, false positive or false negative [10][23] based on the confusion matrix which allows visualizing the algorithm as shown in the Figure 4.12. Then the formula will be executed and will be displayed in graph mode. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, the true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly.

The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative. F1 represents relationship between different data sets that is how close two different labels are given to a single tweet in harmonic manner [11].

\[
F1 = \frac{2tp}{2tp + fp + fn}
\]

Where tp is true positive, fn is false negative and fp is false positive.
For single category C, [10]

- **Accuracy** \( a = \frac{tp + tn}{tp + tn + fp + fn} \) is proportion to the total number of predictions which are correct [11].

- **Precision** \( p = \frac{tp}{tp + fp} \) is proportion to the predicted positive case which is correct [10].

- The recall \( r = \frac{tp}{tp + fn} \) is the proportion of positive cases that were correctly identified which is similar to true positive (TP) [10].
The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, the true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly.

The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative. F1 represents relationship between different data sets that is how close two different labels are given to a single tweet in harmonic manner.

5 REQUIREMENTS

5.1 Environment

The proposed system is implemented in Java. Java Swing and twitter4j and xampp are mainly used. The programming environment used is Netbeans IDE.

5.2 System Requirements

System requirements will include software requirements and hardware requirements those necessary to support the project.
5.2.1 Software Requirements

This is the description of the software needed to be established by the user before using this application.

- Operating System : Windows XP and higher version
- Language : java
- Version : 7
- IDE : NetBeans 8.0
- Back-end : MySQL

5.2.2 Hardware requirements

To run this application these are the basic versions of hardware which are needed. It gives high speed of execution if the versions are updated.

- Processor : PENTIUM IV
- Clock Speed : 2.7 GHZ
- Ram Capacity : 1 GB
- Hard Disk Drive : 250 GB
5.3 Java Swing

Java Swing, which was released by Oracle, is a part of java foundation class and is a Graphical User Interface (GUI) toolkit to create java applications. This program lets programmers make GUI for java applications and provides platform independent components. Furthermore, there are more familiar components offered, which include labels, checkboxes lists tables and buttons. In addition, some of its components have drag and drop features to allow for further ease of use. All its class names begin with ‘J’ such as JList, JButton and JFrame.

5.4 Xamp

Xamp[9] is an open source cross platform Apache MySQL PHP and Perl apache distribution to create the local web server by the developers as it is intended to use as a tool only by developers because it configures the necessary components without any barriers.

5.5 Netbeans ide

NetBeans is an open source platform Integrated Development Environment (IDE) which is used to develop desktop, web and mobile applications in Java [5]. NetBeans IDE can run on any operating system platforms that support a compatible Java Virtual Machine (JVM). NetBeans IDE 8.0.2 is the latest version and is used to implement this project. NetBeans platform provides features such as User Interface management, project management, window management and storage management.
6. TESTING AND RESULTS

Mining of social media will give us a lot of information about our need. In the project, twitter data is collected and processed to find the issues that the students are facing in their life. Various test cases will be generated based on the flow of the implementation, like testing the pre-processed the data from the collected data, checking the categories in which each label set falls into.

Once the data is tested it will be evaluated based on different activities, input and the output and the activities occurring in between. Also evaluation includes so many indicators to assess the results efficiently. Once the user login to the system, as shown in the data flow diagram in the Figure 3.1, the flow of the application should be tested. First testing should be done to check whether tweets are collected in proper manner or not based on the geo location, keyword and search id aspects. Later testing is performed on collected tweets, whether data pre-processing is done or not according to the rules mentioned in the section 3.3. The filtered data is further processed to remove the duplicates by trimming the unwanted letters. And finally data was tested to check whether the tweets were able to get fetched if the user location is turned off.
6.1 Test Case 1: Tweet Collection with Search ID

When the user tweets the opinion as per the Twitter terminology, a tweet will end with the hash tag. Once the tweet is compared with the hashtag in the database depending on the categories specified like heavy study load, sleep problems, diversity issues, data will be collected for example the data which was tweeted with hash tag will be stacked onto the frame correctly as shown in the Figure 6.1.

Figure 6.1 Tweet Collection Based on Hashtag
6.2 Test Case 2: Tweet collection with geo location

User turns off the location and tweets the opinion to check whether it is getting fetched or not as shown in the Figure 6.2. But while checking for the location based on longitude and latitude it will consider the tweets which are from Corpus Christi only. So if the user turns off the location and tweets in the output frame the data won’t be appeared as shown in the Figure 6.3.
Search ID: corpus_christi
The following are its fetched tweets,

If you don’t think every day is a good day... just try missing one and see...)
"@gerardway: Sending you the good vibes, hope your day is sweet if it isn’t, here’s a little (((((extra)))))
RT @Ree_SYE: @YoungReeez http://t.co/TurpsinG48
RT @Ree_SYE: @YoungReeez http://t.co/TurpsinG48
#wow is the biggest mahomie ever, Clarisa Ellondo. ☼☼☼ Twinning today with Austin Mahone... https://t.co/ScLsg6kT0P
Sums up my life ☼
It’s like I take 2 steps forward to get 3 steps pushed back ☼
I’m so nice... I just let people walk all over me and I’m getting tired of it ☼
Been staying quiet for too long ☼
Don’t want to read your problems on my timeline.
RT @theepilogue: @alexandrevreyl I miss you too.

Figure 6.3 Result for Location Turnoff

6.3 Test Case 3: Preprocessing the tweet

Twitter slang includes so many emoticons and symbols in a tweet. All those emoticons will be removed and the result will be in the form plain sentence for further processing as shown in the Figure 6.4.
6.4 Test Case4: Duplicate Removal

User’s tweet may be expressive with the extended letters but for the analysis purpose all the duplicate words need to be trimmed to match with the database keywords. In the Figure 6.5 the tweet saying “Happyyyyy studiessss” is trimmed to “happy studies”.

Figure 6.4 Preprocess the Tweet
6.5 Tweet Analysis- Multi Label Classifier

Running the naïve Bayes classifier algorithm on the filtered tweets will result the following output based on different prominent themes which are categorized after collecting the tweets.
It will define the tweet based on the problems categorized in each theme. The problems which are not related to the learning experience will be sometimes in a huge volume that is less important problems with high frequency rate. These kind of problems will be shown in others category as shown in the Figure 6.6.

![Tweet Classification](image)

**Figure 6.6 Results from the Naïve Bayes Classifier**

After this naïve bayes classification, tweets are supposed to be tested based on their classification and actual meaning as shown in the Figure6.7. These values will be further
used for label based graph to calculate the values for accuracy, recall, precision and F1[24]. Tweet will be fetched one after the other, upon clicking the prediction button as shown in the Figure 6.7. It analyzes, whether the tweet falls into the category as it was labeled or it has any other meaning. Later the result will be stored in database to calculate the accuracy, precision, recall values.

Figure 6.7 Label Based Analysis
Based on the true positive or false positive or true negative or false negative the graph will be generated for Accuracy, Precision, Recall and F1 as shown in the Figure 6.8.

![Label Based Measure Graph](image)

**Figure 6.8  Label Based Measures**

And finally based on the values of accuracy precision and recall of the existing system [6][17], comparison is made as shown in the table 6.1 and a graph is generated as shown in the Figure 6.9.

<table>
<thead>
<tr>
<th></th>
<th>Existing System</th>
<th>Proposed System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>30(TP)</td>
<td>17(TN)</td>
</tr>
<tr>
<td>Negative</td>
<td>11(FP)</td>
<td>29(TN)</td>
</tr>
</tbody>
</table>

**Table 6.1 Analysis by Comparing Both the Systems.**
In the existing system [6] where the sentiment analysis used to restrict the word to be
termed as positive, negative or neutral without any classification and each label falls into
single category.

![Label Based Measure Graph](image)

**Figure 6.9 Comparison Result**
7 CONCLUSION

Mining social media data is beneficial for researcher in an education system to identify the student learning experience. The workflow for analyzing educational content in the social media helps to overcome the limitations of the large data mining and manual quality analysis of user generated textual content in social media as in depth qualitative mining is performed in this system.

This work helps the organization and in an education system to and the present student educational experience. Based on this organization and participation can easily take decision in the engineering studies.
8 FUTURE WORK

Apart from the work done towards this system, future work mainly comprises of the following objectives:

- In future we can collect large student generated data other than texts which may include videos and images for analyzing the student experience with exact results. Graphs are used to show the both positive and negative experience results on a yearly basis.
- Building a tool based on the social media data and the student academic performance.
- Future work may include analysis on long tail as tradeoff as this category will bring essential information as well.


22. http://rspa.royalsocietypublishing.org/content/465/2109/2927


