Socially Mediated Patient Portal

By

Jamil Hussain

2011-NUST-MS-PhD-IT-041

Supervisor

Dr. Hafiz Farooq Ahmad

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of
Masters of Science in Information Technology (MS IT)
In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(April 2014)

Approval

It is certified that the contents and form of the thesis entitled **"Socially Mediated Patient Portal**" submitted by **Jamil Hussain** have been found satisfactory for the requirement of the degree.

Advisor: ___Dr. Hafiz Farooq Ahmad_____

Signature:_____  _____

Date: _____18/4/2014_____

Committee Member1: __Mr. Bilal Ali_____

Signature_____

Date: _____

Committee Member2: ___Mr. Muhammad Bilal

Signature _____

Date: _____

Committee Member3: _Ms. Sana Khalique__

Signature _____  _____

Date: _____19/4/2014_____

Dedication

I dedicate this thesis work to my parents who from the start encouraged and supported me. Also to my brother who is an endless source of motivation and guidance.

It is also dedicated to my friends and to my teachers with whom I have an exceptional and admirable relationship.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by an- other person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name:__Jamil Hussain___

Signature: _____

Acknowledgment

First and foremost praises be to Allah, the Almighty, on whom ultimately we depend for direction and guidance.

I wish to express my deep sense of gratitude to my supervisor Dr. Hafiz Farooq Ahmad for giving me a chance to work with him, in a research environment and polish my skills.

Words are inadequate in offering my thanks to the members of advisory Committee and Muhammad Afzal for their valuable input whenever required. Finally, yet importantly, I would like to express my heartfelt thanks to my colleagues and fellow students who helped me directly or indirectly.

Table of Contents

Contents

List of Tables

List of Figures

List of Algorithms

Abstract

Online Social-Networking Websites (OSNWs) are web-based communities used by individuals to make an online profile, provides a platform through which people share information in very cost effective way and easily expresses their opinions like Facebook, Twitter and MySpace. OSNWs have tremendously altered the way of communication. Face-to-face communications got replaced by posting the status updates, leaving comments and like. These have millions of users, having huge amount of user-generated content (UGC) that can be used in health-related human behaviors study in a cost-effective manner. In worldwide the mental illness is a primary cause of disability. Actually, for the diagnosing of psychological illness, there are no pathological tests; the treatment methodology is highly dependent upon the behavioral actions reported by patient himself and his closed ones. The preciseness and accuracy of information received from patient are subject to his/her non-artificial behaviors. In order to handle the challenge, we found that OSNWs can be used as a screening tool for discovering an affective disorder in individuals. This study investigates how Facebook user's profile can help in Life care Decision Support System (LCDSS) and how to use Facebook as screening tool to expose the user‘s mental illness from his/her profile. We propose a way to automatically classify Facebook user personally written text: status updates and comments using a Support Vector Machines and other classification algorithms of supervised machine-learning techniques, build a model to set up, train the classifier and identify depressive symptom–related features from user's profile. The assay of results depicts that statistics of deprived user's meet DSM-IV criterion for a depression symptoms or Major Depressive Disorder (MDD).

Chapter 1
Introduction

This chapter gives the basic idea of the concepts involved in this research. It also presents the background and motivation for this study. Moreover, it provides the hypothesis, gives an idea of expected results, and methodology to get and evaluate the results. Finally, it presents the structure of this thesis document.

1.1. Facebook

Facebook is well-known OSNWs is, was created in February 4, 2004 by Mark Zuckerberg (Harvard University student) [20]. Firstly its usage was limited to Harvard University students, but later on extended it Stanford University colleges, the Ivy League and Boston area. It increasingly added support for students at a variety of other universities. Now anyone can be registered to it except 13 years old. It can only use by register user, after that they can create a profile, add users as friends, post status updates, send messages, and much more. Once logged on, user's may engage in various activities, including wall posting, viewing user's profile, posting links, browsing and uploading photos, videos, and playing games etc. While posting status updates FB asking one question: What's on your mind?

The statistics estimation of Facebook that there are more than 1.223 billion monthly active users, 946 million users of them were using Facebook on their mobile devices; daily active users are 664 million, monthly mobile active user are 751 million, 20 billion minutes are spent per day by users worldwide; 4.75 billion content are shared per day; average 40 pages are liked per user reported at end of December 2013.

Facebook provides many APIs: Graph API, FQL, open Graph, Chat, ADS, Dialogs etc. we were used Graph API and FQL for data collection. The **Graph API** is HTTP-based API that provides access to social graph of the Facebook that representing connections among

objects the graph. Many other Facebook APIs are based on it. FQL syntax is much closer to SQL to query Facebook user table data. It gives advanced features that are limited in Graph API such as result generated by one query is used in another multi-query.

1.2. Motivation

On world level the disability is generally caused by mental illness. According to WHO (2001) report almost three thundered million people are the victims of mental depression. The statistical results based on various reports reveals that during a year North America is having MDD for male 3 to 5 % and for female are 8 to 10% [17]

Still, there is insufficient global support and service for exposing the mental illness (Detels, 2009). Although 87% of countries offer budget to deal with mental illness, and 28% have no budget specially to deal with mental health (Detels, 2009).

Actually, for mental illness medical science is still incompetent and no trustworthy techniques have been formulated which should be relied upon in this regard; usually, it's questioners based, patient's self-reported, behaviors mentioned by its closed one. Questionnaires frequently depend on a person's memory, which are subject to high degree of inaccuracy.

In order to overcome these challenges, we found social media as a tool in finding and predicting behaviors in individuals. The social-media activities overcome some of the problems regarding patients' self-reporting. By mining the online social media user's activities, we may get closer to the natural behavior of the user and his way of thinking.

We are mainly concerned to the Major Depressive Disorder (MDD), which is the most common mental illness. Most common symptoms of MDD are sadness, tried, sleep problem, losing interest in activities etc. Recently the cheaper availability of internet has made an incredible increase in the use of social media websites. On these websites the users communicate in very natural way through posts relating to their present past and future. The content the user post truly depict his current behavior, thinking style and mood. From social media user profile we can retrieve behavioral attributes which related to person's mood,

thinking style, interaction and activities. User's posting on Facebook might show guilt feeling, feeling of tried, sleep related problem, helplessness and worthlessness that are the symptoms of MDD. Also, depressed person have low social activities. Such kind of changes in user activity may be significant with changes in its social media activities.

1.3. Objective of this thesis

This study investigates how Facebook user's profile can help in Life Care Decision Support System (LCDSS), how to use Facebook as screening tool to expose the user's mental illness from its profile. We propose a way to automatically classify Facebook user personally written text: status updates and comments using Support Vector Machines and other classification algorithms, build a model to set up and train the classifier and identify depressive symptom–related features from user's Facebook profile.

Actually, to build classifier for Facebook status updates, backend is need where training and test dataset might be set up. The main goal is to make a quality training dataset, in order to achieve with high accuracy.

After building the classifier, it necessary to be estimate the accuracy of model. The objective is getting accurate results more than 80% accuracy). After that frontend will be build. It must offer a suitable way to register user's with SMMP portal , fetch profile data based on user permission using Facebook login and classify users as Depression Displayer and non-Displayer according CESD-R Scale

The purpose of this study is to assess user's from its Facebook profile to expose that that meet DSM-IV criterion for a depression symptoms or Major Depressive Disorder (MDD).

1.4. Challenges

There are different challenges in information filtering on Facebook status updates. Challenges are follows:

- **Short texts** Facebook status is short having sparse data; so it is hard to classify them.

- **Informal language** another problem is informal language used on Facebook posting. It is less structured, it might contains abbreviations, shorten words, modified words and emotions. So the vocabulary is very huge and any external source such as WordNet or Wikipedia or will not be enough to get additional information about the text. In addition it is required to recognize most common words and keywords that will be useful for text classification. Because of informal language simple stop word removal is not sufficient common words on Facebook should be identified.
- **Constantly vocabulary changing** the vocabulary on Facebook status is changing constantly with new phrases and words. So that the classification system should be dynamic that reflects to changing in the language.
- **Different Languages** Facebook is used worldwide; thus numerous languages are used in posting status update and comments.

1.5. Our Approach

A web portal was developed that fetch user profile data using Facebook API and expose mental related problem from its profile data using data mining techniques. For this purpose, many Facebook attributes such as status update, number of friends, number of user likes etc. was considered. Depression symptom reference from Status updates were considered if they meet one of the depression criteria by synonym or keyword. Such as, one symptom of MDD is "Sadness," thus a posting like "I am feeling sad" probable considered as depression reference. The term "insomnia" is a synonym of "sleep" so, a Facebook posting such as "I have insomnia problem" would be considered a depression reference. Status updates that show reference as another person (i.e. "Alice looks sad Today") not considered as depression references.

At First step user's profile were classified as two groups: user's Profile lacking any depression references were considered as "non-displayers" Profiles having depression references were considered as "depression Displayers"

Also we were further classifying the Depression Displayers status messages into 9 categories according CESD-R Scale, in order to find the Depression scale. According to CESD-R depression scale MDD is characterized as nine symptoms categories. First of all, we focus on extracting the main two categories as Depression Displayer and Non-Displayer from status updates then system further classifies the Depression Displayer status updated into nine symptom categories according CESD-R depression scale. However, some symptom categories such as sleep and tried that are more similar to each other, to overcome this problem; we suggest scoring feature to compares frequencies in the created word vector. With the help of scoring feature, we are capable to differentiate keywords associated to each category, and we can measure how it related to the key listed categories.

1.6. Contribution

The outcomes of this thesis are following:

1. **Dataset**. We created a dataset having a list of Facebook users (100 users) and their status updates (approx. 40 statuses per User). The status update set was manually categorized for train the classifier.
2. **Platform**. The most important assignment of thesis was to build the web portal which classifies status messages according CESD-R scale. We provide a web interface that generate Facebook personal analytics' and also classify the status update of a given user, and a backend to setup and test the classifier. The frontend will be made available publicly afterwards.
3. **Comparison.** We compare depressed and non-depressed user behaviors using these measures. Our findings show that individuals with depression displayer references having lower social activity and negative emotion.
4. **Scientific contribution.** We show that supervised machine learning systems such as Support Vector Machines and K-NN can be used to classify status update messages. While those statuses update messages totally differ from documents in regard to used words and length. We examined numerous approaches of feature selection, and find out the optimal cost-parameter setting for K-NN and SVMs Classifier.

1.7. Structure of this thesis

In Chapter 2 we will discuss related work done so for. In chapter 3 we will discuss the state of art of the text classification and explain why SVM and K-NN classifier of choice for our problem. In chapter 4, we will discuss system architecture and brief introduction of its sub-modules. Chapter 5 gives detail discussion regarding the experiments and results of experiments. Lastly, chapter 6 gives final conclusions and future works can be done.

Chapter 2
Literature Survey

University of Missouri Researcher discovers a connection among social anhedonia and Facebook activities. They propose that patient's Facebook profiles can help therapists to understand their mental illnesses. They used 211 college students Facebook profiles and their activities for research purpose. Their findings show that therapists would be capable to gather information regarding patient's by analyzing their Facebook activity as an alternative that are based on patient self-reporting. [1]

The activities of social-media can be helpful in psychological diagnosis that eliminates several problems related with patients' self-reporting. Elizabeth Martin said that "Questionnaires frequently depend on a person's memory, which is subject to high degree of inaccuracy. By sharing patient's a Facebook activity on request gives us a capability to perceive how they naturally expressed themselves." It's the saying and common belief of the teenagers medicine and mental health experts that dark posting is challenge for the symptoms of depression and early signal of the timely involvement. It's the still debatable to interact through Social Media with patients.[23]

In 2013, 200 student Facebook profile were tested for the purpose of determining symptoms and depression level by the researchers at the University of the Washintan and the University of the Wisconsin Madison. The finding of that research reported that 30 student's posted updates show the indication hopelessness, insomnia, or excessive sleeping.

Their results concluded that the college students are facing more depression as compared to the other people. Similarly other research finding also reported that college students are more exposed to weakening depressive each year.

Dr. Megan A Moreno an assistant professor of pediatrics at the university of Wisconsin-Madison and Principle in the Social Media studies said that's everyone can spot adolescent and teenage on Facebook the people who are being exposed to the heavy heartedness and gloominess.

Gamon, M [2] used Twitter like a screening tool for quantifying the MDD in individuals using (CESD-R). First they collected standard labels on depression using crowd sourcing and proposed a diversity of an attributed like emotion, style, and user engagement in order to find depression Related symptoms. The results indicate that depressed persons have small social network, more negative feeling, greater concerns with drugs and intense expression of religious ideas. They build a modal using SVM classifier that predicts depression of an individual. The accuracy of classifier was 70%.

Moreno [3] evaluated Facebook displayed depression references and their association with depression and peer opinion. They used first-year college students' Facebook profiles for analysis purpose and categorized as Depression Displayers and Non-Displayers. Participants completed a depression screen and were interviewed about Facebook displayed depression references. Their finding shows that depression displayers were twice as to meet clinical criteria for depression.

Moreno, M. a [4] examined the association of the absence or presence of depression disclosures on Facebook by using the user's activities. In addition, they checked a connection of many depression references with Facebook user's demographics by applied the binomial regression models.

Park, M., Cha, C., & Cha, M [5] show that it is possible to use online social network data for clinical studies. They performed sentiment analysis on the twitter tweets by using the LIWC [19]. They developed a multiple regression model by using all the sentiment categories and examined how these variables are associated to the CES-D score.

Hamouda, S. Ben, & Akaichi, J[6] In this paper, they investigated the utility of sentiment classification on collection of dataset which is Tunisian Facebook users. Their finding show that Facebook statuses have unique characteristics compared to other corpuses (Reviews, News, etc.), and gives similar performance using machine learning algorithms to classify FB statuses. They used the most well-known machine learning algorithms, and performed a comparative experimental in between the Naïve Bayes and the SVM algorithms.

Kosinski, M., Stillwell [7] they investigated that a large range of user's personal attributes that accurately and automatically exposed from Facebook likes. They found a linked

between the Facebook like and other digital records. Additionally, the large range of attributes was predicted. And also suggest that if training dataset will improve then it might be possibly to disclose other attributes too. The logistic/ linear regression was used for predicting private traits in an individual's from its Facebook likes.

Farnadi, G.[8] In this study they applied a machine learning (ML) techniques on Facebook status updates in order to find users' personality traits. This study show that Facebook status updates can be use expose hidden information. They use Big-Five Model to inter the user's personality.

Rahman, M. M. (n.d.)[9] In this research, they used Facebook as key data source and collected many attributes from Facebook as raw data like comments, wall post , about me and date of birth. This study show that analyzed the mined knowledge is used for pattern recognition, decision making, human behavior prediction and product marketing. They used data mining approaches to excavate intellectual knowledge.

Chapter 3
Methodology

This chapter gives an outline of material and an overview of text classification. The first section explains the dataset that was utilized to train and evaluate the developed models. The final section described the algorithms related to machine learning applied in the experiments. We used many data mining techniques in our research reveal useful information from UGC. Firstly a collection of attributes has been chosen from Facebook user's profile that is related to user behavior and activities. The attributes lists are given in Table 3.1-3. Collected data has been stored in our personalize database.

3.1. Dataset

For Research purpose data collection is not easy and simple job from Facebook. There are assumptions and decisions to be made. We used Facebook search feature in order to collect the domain related corpus from Facebook user walls.

3.1.1. Training Dataset

For training the classifier: Depression Displayer and non-Displayer, two datasets are used. Depression Displayers data contains depression related symptoms, while non-Displayer doesn't show depression related symptoms. Depression Displayer data is data that have depression related emotions and symptoms.

A total of 4000 Facebook status updates were collected from Facebook public profiles using Facebook API. After data collection, data was separated into two categories (Depression Displayer and Non-Displayer) according to DSM-IV Criteria. Status update that contains depression related symptoms are coded as Depression Displayer and other were coded as non-displayer.

3.1.2. Non-English status updated removing

We have only train the classifier for only English Language. We used PHP Language detector Pear Package to detect the English phrases, it compared the string to find their language; In that case the string are Facebook status. PHP Language detector Pear Package give confidence level range from 0–1. The table 3.1 shows data distribution of English and non-English of two categories. Non-English status update ware removed in order to avoid classifier confusions that were used later in experiment.

	English	Non-English	% English
Displayer	1400	700	66.66
Non-displayer	1300	1600	44.83
Total	2700	2300	54

Table 3.1-1 Training Dataset Distribution

Below table show some example of status updates related to Depression displayer

I felt depressed.
My heart keeps breaking.
I am unhappy all day.
I am so tired and I could cry.
Too many emotions I don't know what to do!!!
I hate this life.

Table 3.1-2 Status updates examples related to Depression displayer category

Attribute	Data Type
ID	String
Name	String
about_me	String
activities	String
birthday	Timestamp
first_name	String
friend	Object
friend_count	int32

friend_request_count	int32
note	Object
notes_count	int32
status	Object
wall_count	int32
like	Object
photo	Object
video	Object

Table 3.1-3 Attribute List

3.2. Introduction to Text Classification

This section gives an outline the current state-of-the-art in text classification.

3.2.1. Classification

Classification is supervised data mining (machine learning) technique that assign a label to unlabeled input data items. There are two types of classification base on the number of classes

- Binary classification – classify dataset into 1 of the 2 classes.
- Multi-class classification – classify dataset into several classes.

3.2.2. Text Classification Text

In text classification, classification algorithms are applied on text documents; assign a text document into one (or more) classes. For a classifier to train how to classify the text documents, it required manual labeling, the input dataset are divided into training and test dataset. Training dataset consist those documents that are already labeled. Test data sets are those that are unlabeled. The aim is to learn the information from the existing labeled training dataset and used that knowledge for a test data to predict the class label. The

Classifier is accountable for learning a classification function (f) that maps the documents (d) to the classes (C) [18]

$$\boxed{f: d \rightarrow c}$$

The classifier then uses this classification function to classify the unlabeled set of documents. This is called supervised learning as a supervisor (person who defines labels or classes in training dataset) serves like a teacher guiding the learning process.

The size of the training and testing dataset is extremely important. If small dataset are supply to classifier to train from, it may not get considerable knowledge to classify the test data properly. Alternatively, if the training dataset is too big as compare to test dataset, it will create a problem which is called "over fitting". In order to handle it, the document is too fairly adjusted according to the training dataset, so much so that their performances degrade on the test data.

3.2.3. Preprocessing

It will perform on text before converting into the vector space. In particular, we used tokenizers, lower case conversion, word stemming and stop-words-removal preprocessing operators.

- **Tokenizer** is the process of converting a sequence of characters to a sequence of tokens. We tokenize a Facebook status update messages to generate tokens which are smaller pieces of text (words), we also applied the stop-words-removal to generated tokens in order to remove the unwanted words
- **Stemming** reduces the English words to its root form. Such as the words 'Jumping and 'Jumped is stem to 'Jump'. This technique is mostly valuable for information extraction and indexing. In 1968 Stemming algorithm was firstly developed by Julie B. Lovins. In 1980 M.F Porter made another algorithm that created based line for word stemming. M.F Porter has made 62 rules which might or might not apply on the given word. I have applied Porter stemmer for my project.

3.2.4. Features

Feature selection is most important task in text classification: terms selection from train dataset. It has two advantages: it accelerate the training process. As well, decreases the noise by eliminating unrelated features, hence classification accuracy is increased. First we made the vector space of the documents that will generate a word dictionary. For that purpose we selected the entire terms from existing documents then converted them into the vector space by removing the unwanted terms using filter stop-words operator then keywords are generated

Below are the some documents from our scenario to define the document space:

Train Dataset:

d1: I am sad all day. d2: I am tired of crying. d3: I am feeling tired. d4: I think Insomnia is stalking me. d5: Happy New Year ☺

Test Dataset:

d6: Feeling Sad today. d7: feeling Happy, Happy Birthday ☺

Now, we have created words dictionary from documents **d1,** d2, d3, and d4 of train dataset, index dictionary represented as ***G (t): t*** is the term:

$$G(t) = \begin{cases} 1, \text{ if t is "Sad"} \\ 2, \text{ if t is "Day"} \\ 3, \text{ if t is tire} \\ 4, \text{ if t is "cry"} \\ 5, \text{ if t is "feel"} \\ 6, \text{ if t is "think"} \\ 7, \text{ if t is "Insomnia"} \\ 8, \text{ if t is “stalking”} \\ 9, \text{ if t is “happy”} \\ 10, \text{ if t is “new”} \\ 11, \text{ if t is “year”} \\ 12, \text{ if t is “☺”} \end{cases}$$

From dictionary stop words were removed and normalize it then test dataset are converted into the vector space in order that every term of vector is indexed as training dictionary index, thus from vector the 1st term are illustrated as “sad”, 2nd is “day” and so on. After that I have used the **term-frequency (tf)** that measure terms occurrence in our Dictionary G**(t)** in the documents ***d6*** or ***d7***, tf is define as:

$$\text{tf}(t, d) = \sum_{x \in d} \text{frq}(\text{x}, \text{t})$$

Where function ***frq(x,t)*** defined as:

$$frq(x, t) = \begin{cases} 1, & if\ x = t \\ 0, & Otherwise \end{cases}$$

So, ***tf(t,d)*** gives total term ***t*** occurrence in the document ***d***. in given example ***tf(“sad” , d7)=1*** so that the term ***“sad”*** is present one time. After we were created the word vector, which is represented by:

$$\overrightarrow{v_{dn}} = (tf(t_1, d_n), tf(t_2, d_n), tf(t_3, d_n), tf(t_4, d_n), \dots, tf(t_n, d_n))$$

Vector representations of test dataset are:

$$\overrightarrow{v_{d6}} = (1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$$

$$\overrightarrow{v_{d7}} = (0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 2\ 0\ 0\ 1\ )$$

In the resulting vector vd_7 illustrate that, 0 occurrences of term "sad", 1 occurrence of term "feel "and so on. We have collection of text status updates; then we converted it into Document-Term matrix. The rows and columns represented the documents and words respectively within these text documents.

$$M_{|D|\times f} = \begin{bmatrix} 1 & 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0 & 0 \\ 0 & 0\ 0\ 0\ 1\ 0\ 0\ 0\ 2\ 0\ 0 & 1 \end{bmatrix}$$

Where |D| respesent Document Space Cardinality

3.2.5. Term Frequency-Inverse Document Frequency

Our document space is defined since $D = \{d1, d2, \dots, dn\}$ where n is the total number of documents in dataset, and in our scenario $D_{Training} = \{d1, d2, d3, \dots, d5\}$ and $D_{Testing} = \{d6, d7\}$. Our document space cardinality of is defined $|\boldsymbol{D_{Training}}| = \mathbf{5}$ and $|\boldsymbol{D_{Testing}}| = \mathbf{2}$, because we have 5 documents for training and 2 documents for testing.

Formula for idf (inverse document frequency) is defined as:

$$idf(t) = \log \frac{|D|}{1 + |\{d: t\ \in d\}|}$$

Where $|\{d: t\ \in d\}|$ represent term t appearance in **number of documents**, 1 are added in order to avoid division by zero.

Formula for the tf-idf is defined as:

$$td - idf(t) = tf\ \times idf(t)$$

In an agreed text document set, every term has a special importance in a certain document. The TF-IDF compute the weight for each term in a document according to all text document set. If the term more occurs in a document as compare to the other documents in set then it assign higher weight to it.

3.2.6. Terms N-grams

It is n consecutive characters of a string. It is very helpful for language detection and speech recognition. I have used a term n-Gram that is a series of adjacent tokens of length n, in order to detect the negation of sentence. The negation "not" could change entire sentence polarity. Such as "I am happy" is apparently non-displayer, and "I am not happy" is depression displayer sentence. By using unigrams, the 1st sentence features would be {"I", "am", "happy"} and 2nd one {I, am, not, happy}. Both sentences contain the word "happy", then the classifier put it into non-displayer class but actually it is related to depression displayer Class.

This problem is overcome by applying the terms n-gram by appending the negation word, which will reflect the change in sentence polarity. Therefore for negative sentence the features would be {"I", "am−not", "not−happy"}.

3.3. Classification methods

In general, a text classification is divided into supervised and unsupervised. In supervised classification the pre-existing knowledge are used to perform the classification process, but in unsupervised classification has no pre-existing knowledge. In my thesis work supervised classifier were used. Let first discuss the most famous supervised classifier.

3.3.1. Naive Bayes Classifier

It was one of the primary models that are based on probability for information retrieval that was made by C.T. Yu and G. Salton in the 1970.Generally, each document is expressed by a vector as:

$$d_{(t)} = \{d_1, d_2, \dots, d_n\}$$

Where

$$d_{(t)} = \begin{cases} 1, & if\ term\ t\ is\ exist \\ 0, & Otherwise \end{cases}$$

The relevance probability of a document can be calculated by the Bayes rule. In practice it gives good results due to the simple assumptions used by it.

Basic algorithm of Naïve Bayes classifier (multinomial mode) [9] is follow:

Naïve Bayes classifier (multinomial mode)

```
TRAINMULTINOMIALNB(ℂ, 𝔻)
 1  V ← EXTRACTVOCABULARY(𝔻)
 2  N ← COUNTDOCS(𝔻)
 3  for each c ∈ ℂ
 4  do N_c ← COUNTDOCSINCLASS(𝔻, c)
 5     prior[c] ← N_c / N
 6     text_c ← CONCATENATETEXTOFALLDOCSINCLASS(𝔻, c)
 7     for each t ∈ V
 8     do T_ct ← COUNTTOKENSOFTERM(text_c, t)
 9     for each t ∈ V
10     do condprob[t][c] ← (T_ct + 1) / Σ_t'(T_ct' + 1)
11  return V, prior, condprob

APPLYMULTINOMIALNB(ℂ, V, prior, condprob, d)
1  W ← EXTRACTTOKENSFROMDOC(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4     for each t ∈ W
5     do score[c] += log condprob[t][c]
6  return arg max_{c∈ℂ} score[c]
```

Algorithm 3.3-1 Naïve Bayes classifier

3.3.2. K-NN Classifier

The ‘k Nearest Neighbor’ approach allocates a class to every training document. Then new document is classified according to the k nearest neighbors. Such as, if the value k is set to 1, then new document in the same class as 1 neighbor will assign. If the value of k is set to 5, then the algorithm will choose the class that occurs most frequently in the close to 5 neighbors. K-NN does relatively well if k value greater than 1.

Basic algorithm of K-NN [17] is following:

Algorithm 2: K-NN

```
TRAIN-KNN(ℂ, 𝔻)
1  𝔻′ ← PREPROCESS(𝔻)
2  k ← SELECT-K(ℂ, 𝔻′)
3  return 𝔻′, k

APPLY-KNN(ℂ, 𝔻′, k, d)
1  S_k ← COMPUTENEARESTNEIGHBORS(𝔻′, k, d)
2  for each c_j ∈ ℂ
3  do p_j ← |S_k ∩ c_j|/k
4  return arg max_j p_j
```

Algorithm 3.3-2 K-NN

We have used the cosine similarity distance in which score is calculated as:

$$score(c,d) = \sum_{d' \in s_k(d)} I_c(d') cos\left(\vec{v}\,(d'), \vec{v}\,(d)\right)$$

Where $s_k(d)$ represent a set of d's k-nearest neighbors and if d's is in class c then "$I_c(d') = 1$" and "0" otherwise. Documents are assigning to the class having higher score. The weighting accuracy by similarities is frequently greater than simple voting.

3.3.3. Support Vector Machines

In Support Vector Machines (SVMs) are then train dataset are divided into two categories, trained SVM model assigns new item into one of two categories. SVMs model makes a hyperplan which divides an items by a higher gap.

Below figure-3.3-1 have 2 categories: blue and red, hyperplane are divided into 2 categories by the greater margin as below figure. The margin determined by items is so called the Support Vectors (1 red and 2 blue items in the given example).

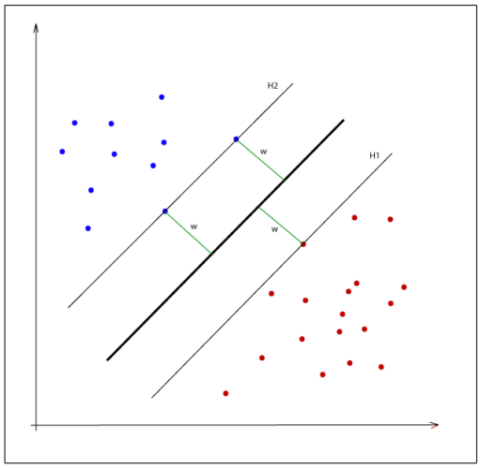

Figure 3.3-1 SVM hyperplane

Following are some agreed training dataset ***"D"*** as

$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^p, y_i \in \{+1, -1\}\}_{i=1}^{n}$$

Where $y_i = \begin{cases} +1, & if\ (x,y)\ is\ part\ of\ class \\ -1, & otherwise \end{cases}$

Which can be separated by the hyperplane?

$$(\vec{w}.x) - b = 0$$

The $\vec{w}$ represents normal vector of hyperplane separation. Our goal is to maximize the $\vec{w}$ length. In order to achieve this, need to introduce two other hyperplanes such as:

$(\vec{w}.x) - b \geq 1\ if\ \ y_i = +1$ ---------- (i)

And

$(\vec{w}.x) - b \leq\ 0\ if\ \ y_i = -1$ ----------(ii)

From eq(i) and eq(ii) we get.

$$y_i\left[(\vec{w}.x) - b\right] \geq 1, where\ i = 1, 2, 3, \ldots., n$$

Those points that satisfy the inequality are so called Support Vectors.

3.4. Classifier selection for text categorization

The most important decision is the selection of the optimal classifier for our thesis work.

Dumais et al [11] and Yang et al. [12] equally stated that K-NN and SVM classifiers give best results for text classification. Rios and Zha [13] found that SVM working well while classifying spam e-mails, and Joachims [14] stated that SVM are extremely fitted for text classification.

We are tested all classifier, we found that SVM are best classifier for binomial classification(Displayer and Non-Displayer) and for polynomial classification K-NN gives best results with k value equal to 4 using Rapidminer Tool

Chapter 4
Architecture

This chapter explains the architecture SMPP and the implementation. Section 4.1 explains the architecture for the SMMP and also explains how the applications are implemented and how the final product works.

4.1. SMPP Architecture

Figure 4.1-1 show that how internally SMPP portal works. The client send request via HTTP protocol to the Web Server. The web server send request to the Module called PHP interpreter that are responsible for the request processing such fetching data from Facebook and data storing in MySQL database as well the classification process. After request processing, sent results back to client.

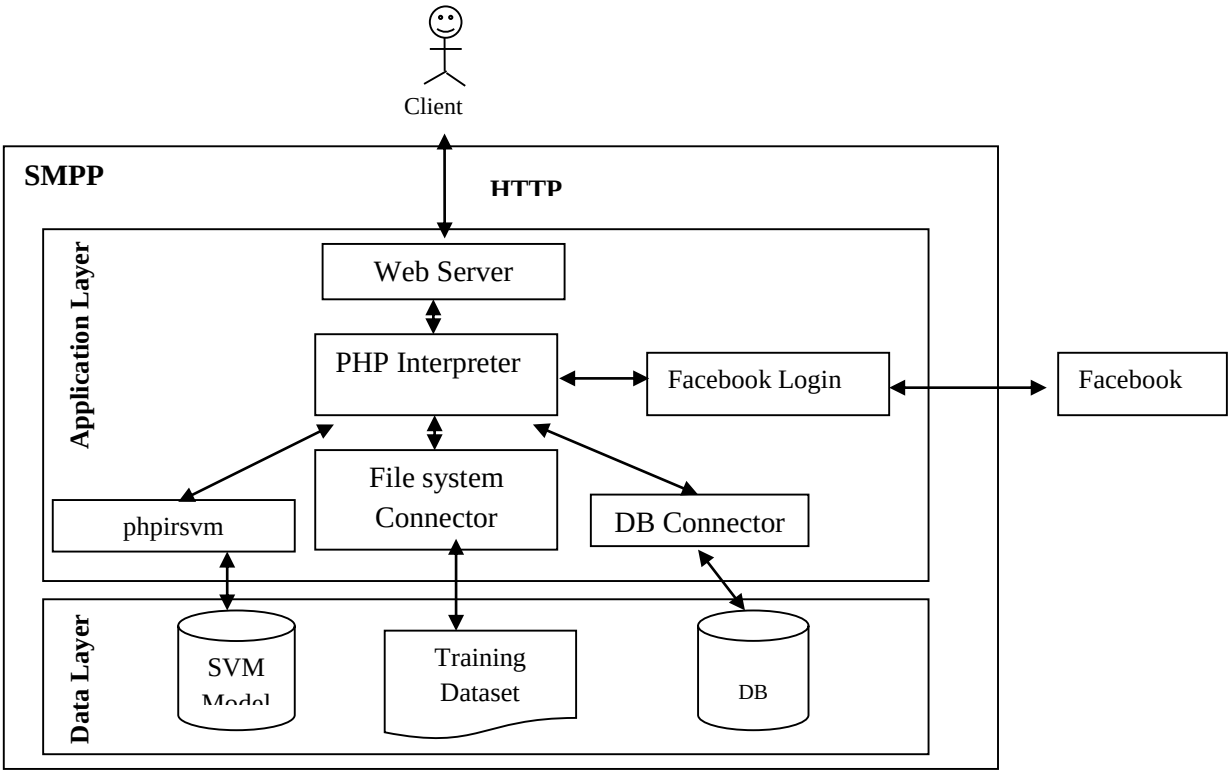

Figure 4.1-1 SMPP Application diagram

The system architecture has many sub Modules that are **Data Collection Module (DCM)**, **Data Processing Module (DPM), Data Classification Module (DCM)** and **Data Visualization Module (DVM).** Fig. 4.1-2 represents complete system architecture.

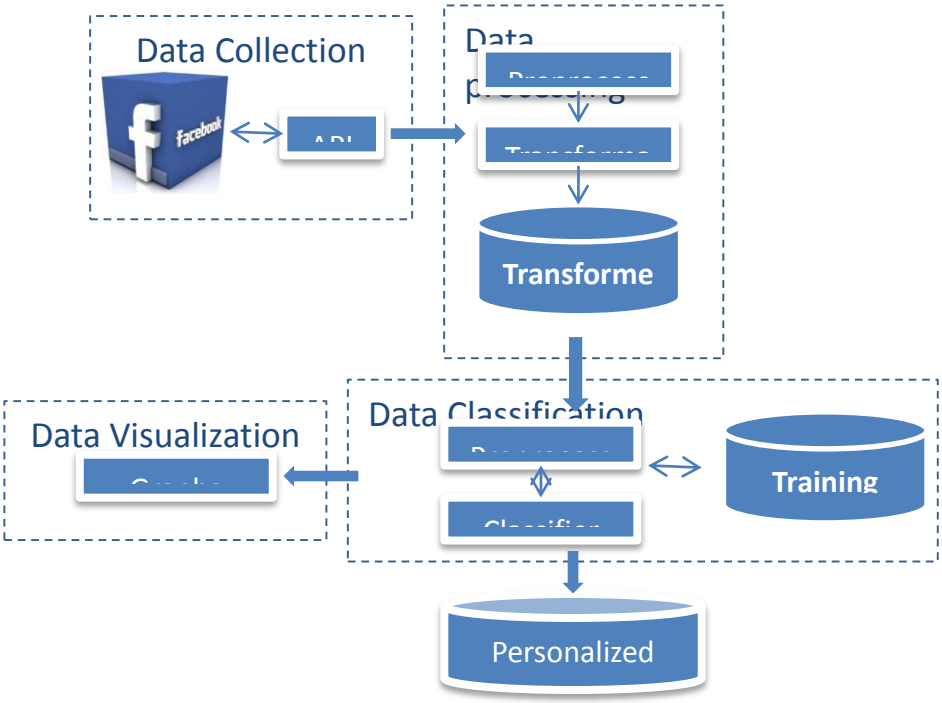

Figure 4.1-2 System Architecture

Different Modules of our system are described in brief:

4.1.1. Data Collection Module (DCM)

It gathers data from Facebook. It has data fetching engine and store data temporary using session variable. In order to collect data, we created and registered a Facebook application, called *SMPP*. Allowing the APP to require specific Facebook privileges and hosted the SMPP at my smpp.chinarportfolio.com, making it available to anyone. *SMPP is* a PHP-based web portal that uses the Facebook Graph API and FQL, which simply use URL-based queries to return result, sets such as status updates, gender, comments, likes, photo, video etc.

DCM used the following algorithm

Data Collection

FB_D: Fetch user profile Data from Facebook,
FB_{DF} : Format the facebook data,
FB_q: Facebook queries
Data_fetching (FB_D, FB_D, FB_{DF})

1. user ← facebook_appid and facebook_app_secret
2. user ← FB.getUser();
3. **If** (user) then
4. Logouturl ← FB.getLogoutUrl()
5. Result_Set ← "null";
6. **For** each ($F_{Bq} \in F_{BD}$)
7. do Result_Set ←SDB; store FB_D in Session Variables
8. end **for**
9. end **if**

Algorithm 4.1-1 Data Collection

4.1.2. Data Processing Module (DPM)

Data transformations are done by DMP that is used by Classification Module. It has preprocessor, transformed data storage. Preprocessor used temporary data stored in session variables and transformed it into normalized form. The transformed data are generated from normalized data that have no noisy, missing information and errors.

4.1.3. Data Classification Module (DCM)

Data classification key module in our system. It contains preprocessor such as Transformation (tokenization, stop words, Stemmer, Normalizer), classifier (SVM and K-NN) and training dataset (pre-classified labels). Training dataset were made based on existing Status updates collection from public Facebook status then labeled it by filtering for known emotion equivalence classes [24] and The 20 items in CESD-R Questioner measure symptoms of depression into nine different groups [21] such as sadness (Dysphoria), sleep, tired(fatigue), loss of interest (Anhedonia), appetite, thinking / concentration, guilt(worthlessness), movement(Agitation), suicidal ideation. Classification Module collects data from Data Preprocessor Module. Firstly it processes status update text data of user's in order to generate tokens from text then Stop words, stemmer, lower case conversation and n-gram terms operators are applied. It created keyword, tf-idf using

equation (1). Lastly features are selected based on tf-idf that is used for the experiment. Then SVM and K-NN classifier are used for text classification. Firstly we have train the classifier using train dataset, and then we applied the train classifier on unseen data (test dataset).Other attributes are also used by classifier such as friend count, wall count, photo count, messages count etc.

Text Preprocessing

Input: D_t : Document set (Facebook Status Updates), WM: weight matrix

Preprocessing (D_t)

1. **for** each D_t do
2. **Tokens ← Tokenization**
3. end of **for**
4. **for** each Tokens do
5. Stop-words Removal
6. end of **for**
7. **for** each Tokens do
8. **Stemming ← Porter Stemmer**
9. end of **for**
10. **for** each Tokens in the word vector do
11. **weight matrix ← Calculate TF/IDF**
12. end of **for**
13. **for** each in WM do
14. **Set the threshold ← 'c'**
15. Calculate (DF_t) for each term
16. **If** DF less then c then
17. Delete term from WM
18. End **if**
19. End **for**

Algorithm 4.1-2 Text Preprocessing

4.1.4. Data Visualization Module (DVM)

It is used to extract intellectual knowledge from classified data. It uses many Javascript data visualization libraries such as Google Graph, D3.js, Jquery etc for classified data visualization. It visualizes classified data as map, chart, graph and tables. It reveals many hidden patterns. DVM used the following algorithm getting the final Result of individual's according to CESD-R Scale.

Insomnia Frame (Late Night Setting based on user's activities)

A_t : User Facebook activities based on fast two week data, T_1 : Insomnia Frame which are set in between 9PM and 6AM; $Count_{lt}$: Total late night setting

InsonniaFrame (A_t , T_1)

1. activities_time_set ← "null";
2. **for** each(A_t) do
   activities_time ← A_t User activities time
3. end **for**
4. **for** each(activities_time)
5. **if** the activities_time in between 9PM and 6AM ***do***
6. **$Count_{lt}$ ← late night setting from A_t**
7. ***return*** $Count_{lt}$

Algorithm 4.1-3 Late Night Setting

CESD-R Determining categories based on two week Data user profile

CAT_{Stat} : User facebook status updates after categorization using classifier(SVM, K-NN), T_1: Status Update creation time, T_2: Two weeks ago time stamp

CESD_R_Scale(CAT $_{Stat}$)

1. SADCount ← "null"; SLCount ← "null"; TIRCount ← "null"; THCount ← "null"; SATTCount ← "null"; MOVCount ← "null"; GUICount ← "null"; LOSCount ← "null";
2. **foreach (CAT $_{Stat}$)**
3. **If(T_1 greater then T_2)**
4. **Switch CAT $_{Stat}$**
5. case "Sadness": SADCount ++;
6. case "Loss of Interset": LOSCount++;
7. case "Sleep":SLCount++;
8. case "Tired":TIRCount++;
9. case "Thinking":THCount++;
10. case "Suicidal Ideation": SATTCount++;
11. case "Movement": MOVCount++;
12. case "Guilt": GUICount++;
13. **end switch**
14. **end if**
15. **end foreach**
16. CESD_Result_set ← "null";
17. **If ((SADCount >=1 || LOSCount >=1) && (countother >=4))**
18. CESD_Result_set ← "Meets criteria for Major depressive episode";
19. **else if(($sadnessCount >=1 || $LOSCount >=1) && ($countother ==3))**
20. CESD_Result_set ← "Probable major depressive episode";
21. **else if((SADCount >=1 || LOSCount >=1) && (countother ==2))**

22. CESD_Result_set ← "Possible major depressive episode";

23. else
24. CESD_Result_set ← "No clinical significance";
25. *return* CESD_Result_set

Algorithm 4.1-4 CESD-R determining categories based on two week data user profile

Chapter 5
Experimental Results and Evaluation

In this chapter, we present and analyze the experimented results. Different machine learning classifiers used for our experiments named, Naïve Bayes (NB), K-NN and Support Vector Machine (SVM) which are provided by RapidMiner environment [17][18].These selected classification methods achieve the best accuracy among other classification methods in our data sets to classify the instances. We explained the machine environment, and the tools used in our research.

We have made experiment using Rapidminer and PHP NLP tools. First we setup an experiments using Rapidminer, test different classifiers, after that we developed online portal using PHP with php-nlp tools used same classifiers that was setup using Rapidminer [22].

5.1. Experiments Setup

In this section, a description about the experimental environment, tools used in experiments, measures of performance evaluation of classifiers.

5.1.1. Experimental Environment and Tools

We applied experiments on a machine with properties that is Intel (R) Core$^{(TM)}$ i3 @ 2.50 GHz, 2.00 GB RAM, 500 GB hard disk drive and Windows 7 operating system. To carry out our thesis (including the experimentation), special tools and programs were used:

- **RapidMiner application program:** used to build our approach, and conduct experiments practical and extracting the required results.
- **Microsoft Excel**: used to organize and store datasets in tables, do some simple preprocessing and analyze the results.
- **Adobe Dreamweaver**
- **Adobe Photoshop**

- **Apache Server**
- **PHP Language**
- **HTML and CSS**
- **Javescript / Jquery / d3.js**

Our experiment has two phases:

- Training phase
- Testing phase

In Training phase classifier are train based on train dataset. On the bases on training model, unlabeled status updates are classified in testing phase.
Preprocessing steps are applied in Training and Testing Phase such as stop-words, stemming, N-gram terms etc. In order to validate, we used cross validation having 10-fold cross validation.

Model categorized Facebook status as depression symptom **displayers** or **non-displayers**. These are the texts that were un-scored in the test dataset are score between 0 and 100, where higher score indicates a more **Displayer**. Displayed depression references were associated with depression; Displayed references are further Classify into sub categories according to CESD-R scale

5.1.2. Explanation of each step

Explanation of each step shown in above in figure 5.1 of work flow is following

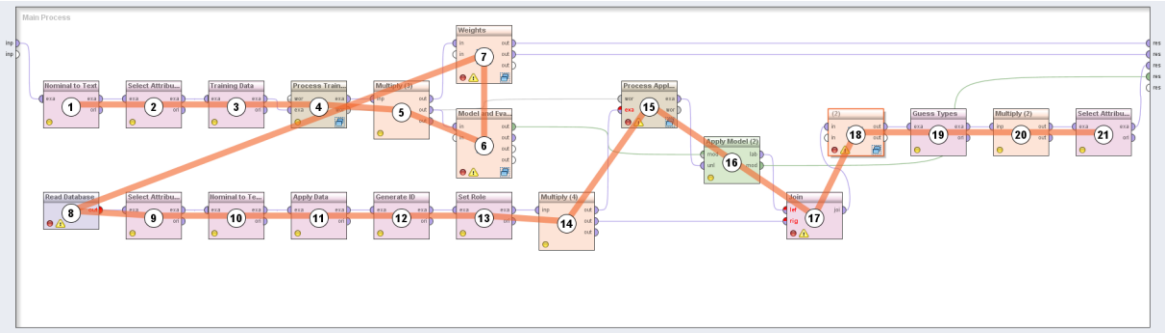

Figure 5.1-1 workflow for SMPP that classify the Facebook status updates

1. Get the train dataset

2. Convert the data type from Nominal to Text for text mining.
3. Select the Attribute, in training dataset there are three attribute, such as Facebook Status , Sentiment(**displayers and Non- displayers**)and Categories (**Sadness**, **Sleep, Tired , Loss of Interest**, **Appetite, Guilt**, **Suicidal ideation** and **Thinking / concentration**). For sentiment I have removed the categories attribute. First we want to train the Model by using SVM for two categories (Displayer and Non-displayer)
4. **Filters attribute** which filter the no-missing-label class for filter, which filter the no-missing-labels rows in training dataset so that model will train on labels data.
5. Then we created the word vector from train Facebook status dataset.

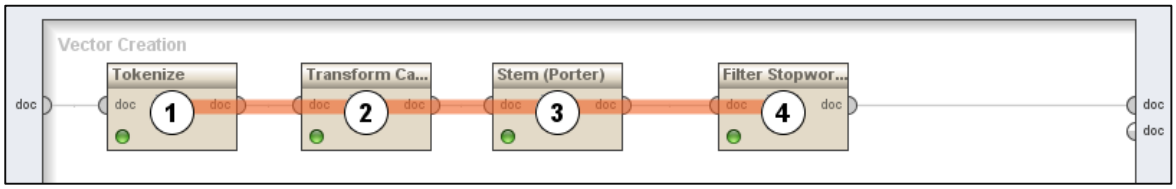

Figure 5.1-2 workflow for word vector creation using TF-IDF matrix

6. Then we applied the Multiply, This operator copies its input object to all connected output ports. It does not modify the input object.
7. Then train the Model using training dataset, for that we applied sub-process operator, this operator introduces a process within a process. Whenever a Sub-process operator is reached during a process execution, first the entire sub-process is executed. Once the sub-process execution is complete, the flow is returned to the process (the parent process). A sub-process can be considered as a small unit of a process, like in process, all operators and combination of operators can be applied in a sub-process. That is why a sub-process can also be defined as a chain of operators that is subsequently applied.

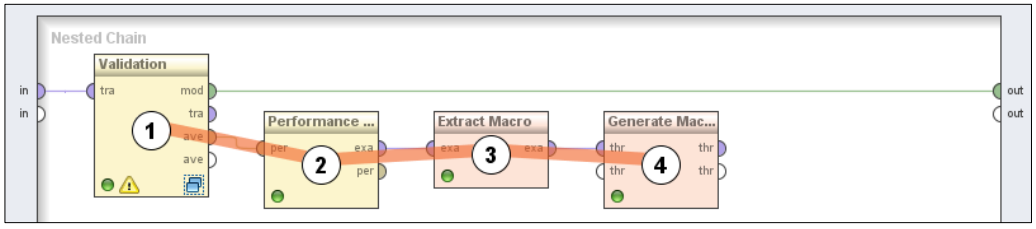

Figure 5.1-3 workflow for Model and Evaluate

Then we applied cross-validation operator. It guesses the statistical performance by a cross-validation. It is generally used to estimate model accuracy. It has two sections (1) **Training** (2) **Testing**

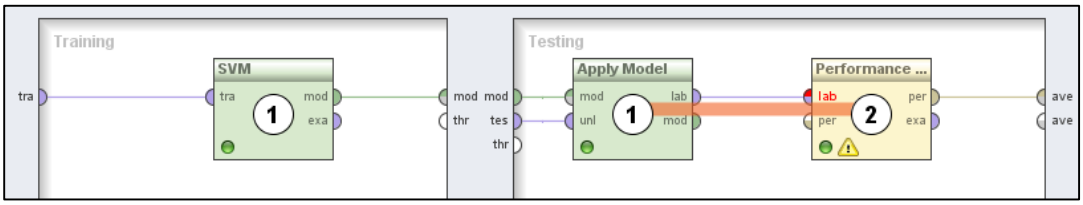

Figure 5.1-4 workflow for the cross-validation to Model and Evaluate

we applied the SVM classifier. This operator is an SVM (Support vector machine) Learner. It is based on the Java libSVM. Those parameters are supplied to SVM classifier such as **SVM Type**: C-SVC, **Kernl Type**:Linear, **C**:100, **Cache Size:** 80, **Epsilon**: 0.001

8. Then we applied the sub-process operator in order to get the weights of **displayers and Non- displayers** key words that are can indentify topics or terms relevant to depression and non depression domain. The weight sub-process work flow is follows.

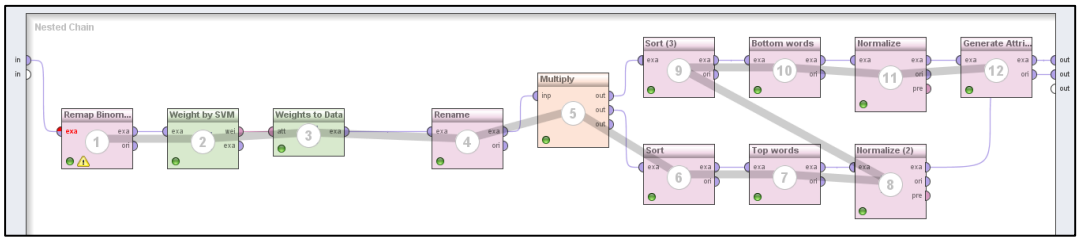

Figure 5.1-5 workflow for Weights assignment to keywords in between 0 to 100

9. In that step it loaded the test dataset in order to classify as **Displayer or Non-Displayer**
10. Convert the data type from Nominal to Text
11. Filter row no-missing-label class for filter
12. Generate ID operator, this operator adds a new attribute with id role in the input test Dataset.

13. Set Role Operator to make the Id as ID role. This operator is used to change the role of one or more attributes.
14. Then applied the **multiply** operator, it copies the input item to all output connected ports. It does not change the input item.
15. Process Document in order to create Word Vector for test DataSet.

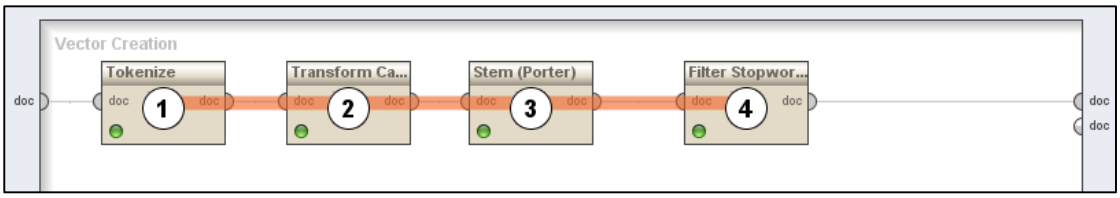

Figure 5.1-6 workflow for test dataset vector creation using TF-IDF

16. Then applied Apply Operator , this operator applies an already learnt or trained model on an test Dataset
17. Then join the facebook test dataset with trained test FB Status using Join Operator, This operator joins two ExampleSets using specified key attribute(s) of the two ExampleSets.
18. Then applied the sub process in order to get **Best and Worst case and score the Test Dataset**

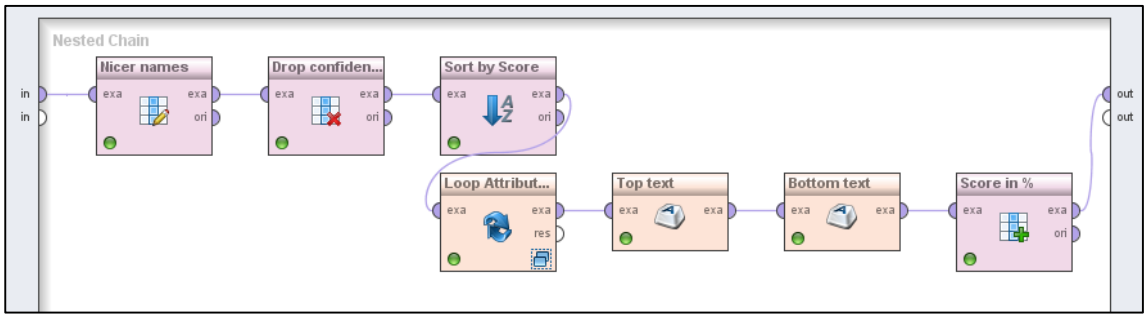

Figure 5.6: workflow for Best and Worst case and score the Test Dataset

19. Then applied the guess Operator, this operator (re-)guesses the value types of all attributes of the input ExampleSet and changes them accordingly.
20. Then applied the Multiply operator

21. Then applied Filter Example Operator, I have filter the Displayer row for further classification.
22. Then applied sub-process in order to classify the Displayer Dataset into Categories (**Sadness**, **Sleep, Tired, Loss of Interest**, **Appetite, Guilt**, **Suicidal ideation** and **Thinking / concentration**).

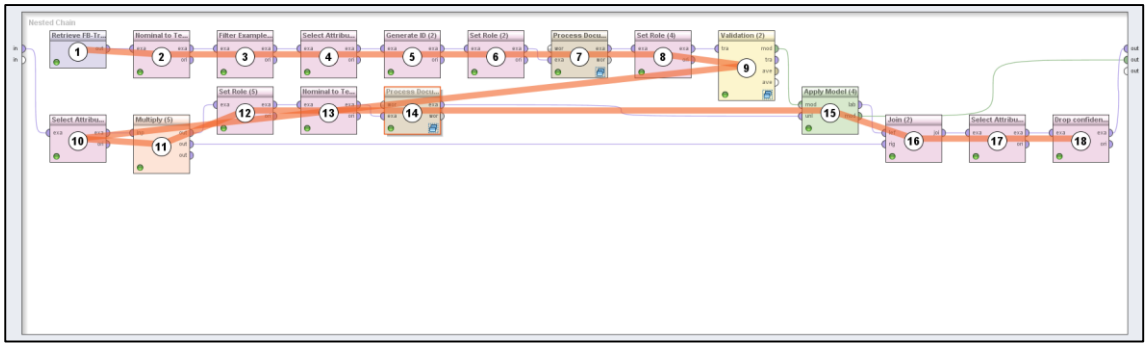

Figure 5.1-7 workflow for Best and Worst case and score the Test Dataset

That sub process also has 18 Main operators, for that we used K-NN classifier, K-NN support multi-class classification, this operator generates a k-Nearest Neighbor model from the input ExampleSet. This model can be a classification or regression model depending on the input ExampleSet. We have tested other classifier for multi-label classification, but K-NN give me best results, on the k value set to 4.

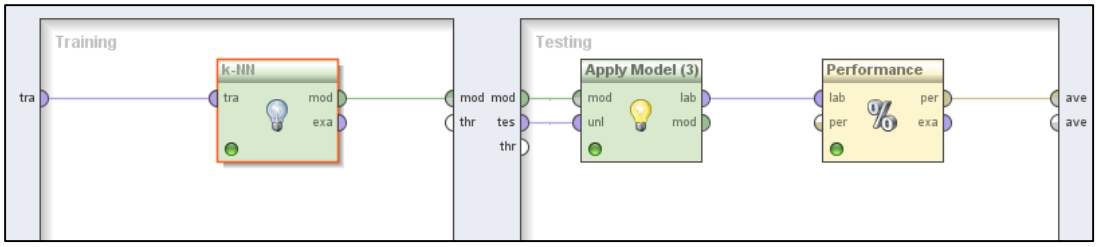

Figure 5.1-8 K-NN work flow

After executed the Model we achieved the following accuracy based on test dataset.

Table View Plot View

accuracy: 77.22% +/- 1.90% (mikro: 77.22%)

	true Displayers	true Non-displayers	class precision
pred. Displayers	541	180	75.03%
pred. Non-displayers	289	1049	78.40%
class recall	65.18%	85.35%	

Table View Plot View

precision: 78.40% +/- 1.27% (mikro: 78.40%) (positive class: Non-displayers)

	true Displayers	true Non-displayers	class precision
pred. Displayers	541	180	75.03%
pred. Non-displayers	289	1049	78.40%
class recall	65.18%	85.35%	

Table View Plot View

recall: 85.36% +/- 2.51% (mikro: 85.35%) (positive class: Non-displayers)

	true Displayers	true Non-displayers	class precision
pred. Displayers	541	180	75.03%
pred. Non-displayers	289	1049	78.40%
class recall	65.18%	85.35%	

Figure 5.1-9 Estimated performance with 10-fold cross-validation (accuracy, precision and recall respectively)

5.2. Experimental Results

5.2.1. Depression Reference Words

These words are the one which are most important for texts with Displayed **Depression References** in test Dataset. For this you can identify topics or terms relevant to depression domain.

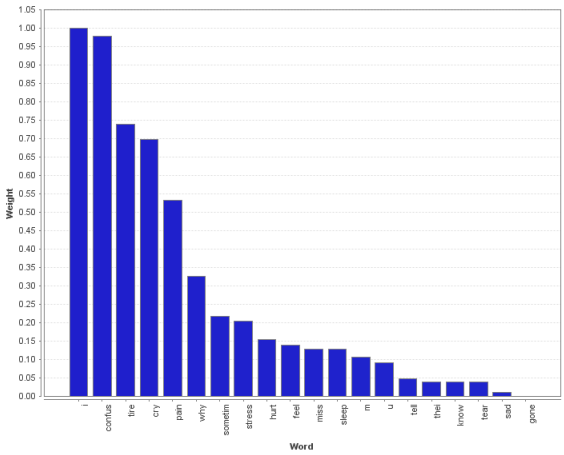

Figure 5.2-1 Important words related to depression displayers in corpus.

5.2.2. Non-Depression Reference Words

These words are the one which are most important for texts with **Non-Displayed depression references** in test Dataset. For this you can identify topics or terms relevant to non-Displayer depression domain.

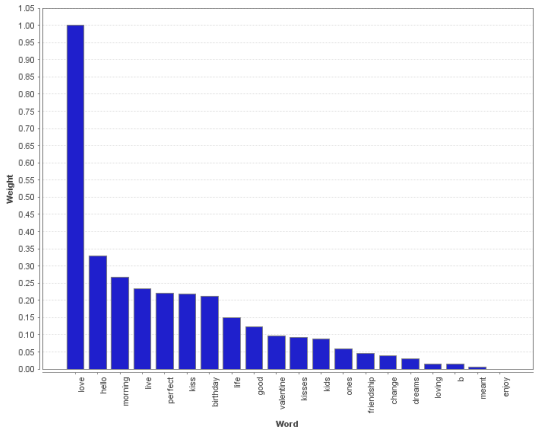

Figure 5.2-2 Important words of Non-Displayers sentiment in the corpus.

5.2.3. Test Data Set Classification Result using SVM

ExampleSet (54 examples, 3 special attributes, 1 regular attribute) Filter (54 / 54 examples): all

Row No.	id	Sentiment	Score	Facebook_Status
1	33	Displayers	100	i am near to cry, i am hurt :(
2	7	Displayers	100	I'm so tired my whole body hurts...did not realize it was possible to be so tired that your eyeballs could hurt too!!!
3	3	Displayers	99	Erika is really really really unhappy'
4	17	Displayers	99	I am sad
5	9	Displayers	98	..tired of all the stress.. just want to stop the world from spinning for a minute and just breathe..
6	48	Displayers	95	i know i cant have both.. what do i do?
7	4	Displayers	94	''John is sad that he doesn't have therapy to go to anymore'
8	14	Displayers	89	I love to sleep
9	52	Displayers	86	Is feeling extremely lonely, and neglected right now. But does anyone really care? :(
10	49	Displayers	83	Too many emotions I don't know what to do!!!
11	29	Displayers	80	Talent does what it can; genius does what it must.
12	31	Displayers	79	i have no idea what to do
13	27	Displayers	77	The best rule of friendship is to keep your heart a little softer than your head.
14	10	Displayers	74	s brain is officially on strike until it gets some rest.
15	25	Displayers	73	That lonely moment when the only text message you get all day is from your cell phone company. :(
16	12	Displayers	73	Sleep?? Whats that?? Oh, that's what I didn't get last night! Now I remember! "Yawn".
17	51	Displayers	73	My heart keeps breaking.
18	11	Displayers	73	[Error 403: User has become unresponsive due to excessive fatigue]
19	39	Displayers	70	A man is as unhappy as he has convinced himself he is.
20	47	Displayers	69	its raining, its pouring, Facebook is boring. I'm of to bed, to rest my head I'll see you all in the morning x
21	50	Displayers	68	Feels like going for a walk, alone, in the dark, with nothing but the moon to guide me.
22	2	Displayers	63	Mary has tears in her eyes
23	1	Displayers	51	Tom is pretty sad. Time for some whiny music...thank god for that'
24	32	Non-displayers	50	my hart and mind have different directions
25	8	Non-displayers	46	Ann needs to stop being lame and so tired so that she can go out and socialize more instead of having to sleep s
26	44	Non-displayers	42	Its a Serious request to all my fellows, juniors, seniors and everyone else. If you are doing or have done BS in CS.
27	5	Non-displayers	34	Just a little funny to make you smile on the last day of 2013...
28	30	Non-displayers	32	I reached home Safe :)
29	43	Non-displayers	29	Pakistan Army's Training Standards
30	24	Non-displayers	27	If pregnancy were a book they would cut out the last two chapters.
31	40	Non-displayers	27	True Friends look at you with no judgment in their eyes, they know you've made mistakes but they accept you for t
32	16	Non-displayers	27	:)
33	22	Non-displayers	27	:(
34	36	Non-displayers	27	hahaha.....
35	38	Non-displayers	27	:')
36	37	Non-displayers	27	Nice Picture :)
37	42	Non-displayers	27	Nice photo and nice caption :)
38	53	Non-displayers	26	Mommy ,mommy, mommy, mommy, mommy, mommy, mommy, mommy, mommy, mommy, mommy, WHAT!!! I L
39	28	Non-displayers	25	Other might have forgotten, But never can i, the Flag of my country Furls very high, Happy Independence Day
40	35	Non-displayers	18	i am not able to do anything, i am feeling lonely
41	46	Non-displayers	17	People who always want or expect something
42	21	Non-displayers	13	Real eyes can realise real lies...
43	18	Non-displayers	6	Eid Mubarak Ho
44	20	Non-displayers	5	God Bless me in 2014. Amen.
45	45	Non-displayers	5	God Bless me in 2014. Amen.
46	6	Non-displayers	5	''Jane doesn't feel like getting up today, or doing anything
47	54	Non-displayers	4	Happiness is about being thankful for the little things in life... and the best things are always free :)
48	13	Non-displayers	4	I want to go bed.
49	41	Non-displayers	3	love you Dad !

Figure 5.2-3 Test dataset classification using SVM

5.2.4. Sentiment Distribution

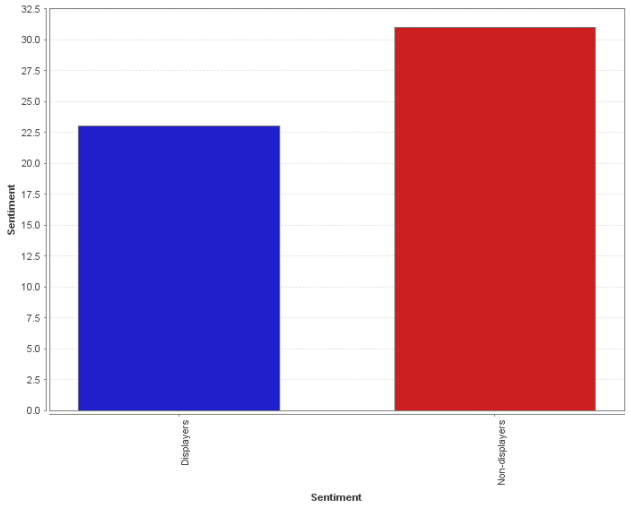

Figure 5.2-4 Number of Displayer versus Non-Displayers texts in Test Dataset.

5.2.5. Displayers Test Dataset Categorization according CESD-R Using K-NN Classifier

ExampleSet (23 examples, 2 special attributes, 1 regular attribute) Filter (23 / 23 examples): all

Row No.	id	prediction(Categories)	Facebook_Status
1	33	Sadness	i am near to cry, i am hurt :(
2	7	Sadness	I'm so tired my whole body hurts...did not realize it was possible to be so tired that your eyeballs cou
3	3	Sadness	Erika is really really really unhappy'
4	17	Sadness	I am sad
5	9	Sadness	..tired of all the stress.. just want to stop the world from spinning for a minute and just breathe..
6	48	Sadness	i know i cant have both.. what do i do?
7	4	Sadness	''John is sad that he doesn't have therapy to go to anymore'
8	14	Sleep	I love to sleep
9	52	Sadness	Is feeling extremely lonely, and neglected right now. But does anyone really care? :(
10	49	Thinking	Too many emotions I don't know what to do!!!
11	29	Sadness	Talent does what it can; genius does what it must.
12	31	Sadness	i have no idea what to do
13	27	Thinking	The best rule of friendship is to keep your heart a little softer than your head.
14	10	Tired	s brain is officially on strike until it gets some rest.
15	25	Sadness	That lonely moment when the only text message you get all day is from your cell phone company. :(
16	12	Sleep	Sleep?? Whats that?? Oh, that's what I didn't get last night! Now I remember! "Yawn".
17	51	Sadness	My heart keeps breaking.
18	11	Sadness	[Error 403: User has become unresponsive due to excessive fatigue]
19	39	Sadness	A man is as unhappy as he has convinced himself he is.
20	47	Sadness	its raining, its pouring, Facebook is boring. I'm of to bed, to rest my head I'll see you all in the mornin
21	50	Sadness	Feels like going for a walk, alone, in the dark, with nothing but the moon to guide me.

Figure 5.2-5 K-NN classifier for Displayer text Categorization

5.3. Online implemented SMPP Portal

SMPP was developed in the SSRG lab under the supervision of Dr. Hafaz Farooq. In order to get the user's report according CESD-R scale, user must be subscribed to Facebook SMPP APP to fetch user profile data based on login permissions. The home page of portal gives the Facebook personal analytics based on user's Facebook profile data such as friend list, friend's gender, post count etc. SMPP assess user's from its Facebook profile to expose that meet DSM-IV criterion for a depression symptoms or Major Depressive Disorder (MDD).

Below from figure 1 to 6 are the screenshots of SMPP web portal.

Figure 5.3-1 Login Screen

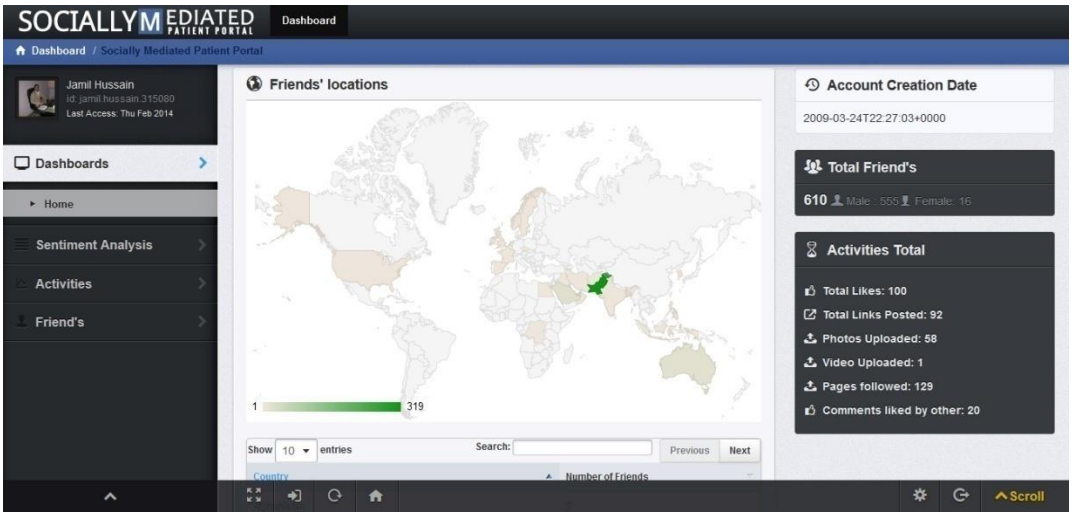

Figure 5.3-2 Dashboard Screen

Figure 5.3-3 Friend's Gender

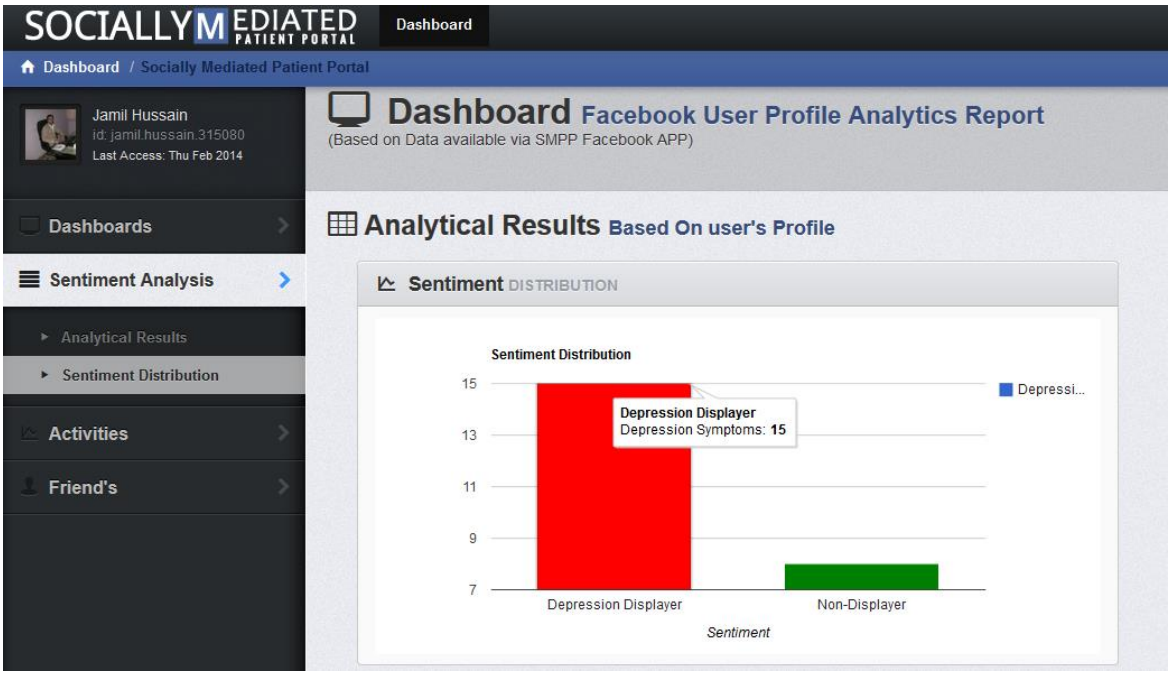

Figure 5.3-4 Sentiment Distribution

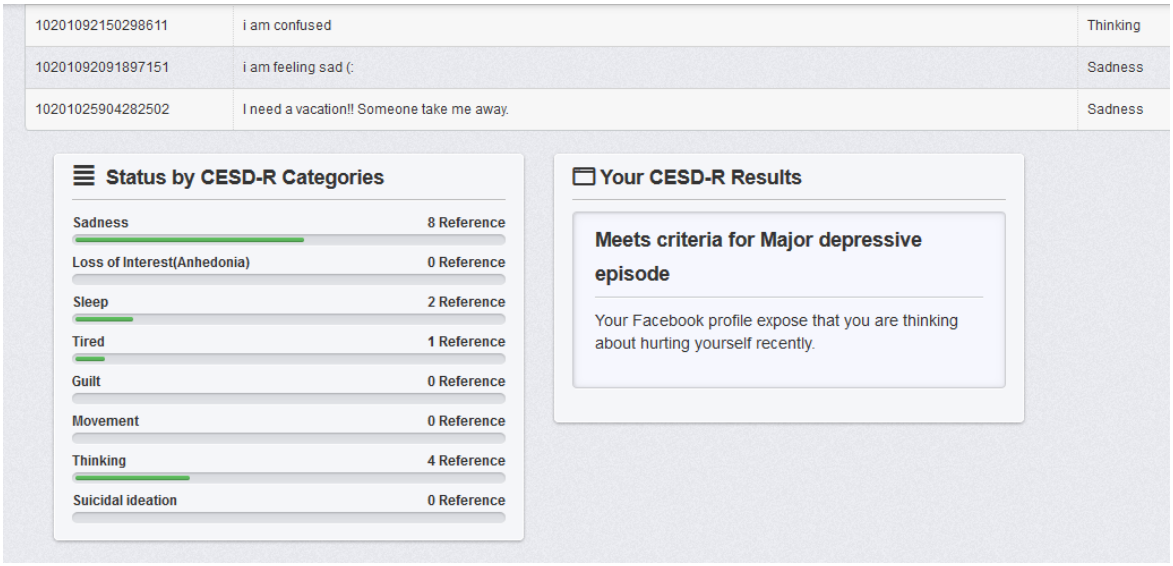

Figure 5.3-5 Status updates by CESD-R Categories and CESD-R result

Chapter 6
Conclusion and Future Work

6.1. Discussion

The key idea of our study was to find out the association among Facebook activities and Major Depression Disorder. We assume that Facebook activity can reveals the mental illness at initial stages. Our results fairly confirm our assumptions.

We envision a web portal that can provide the untimely alerts to an individual's based social data regarding life styles and mental illness. These tools possibly can be used for depression and mental illness diagnosis, additional to questioner techniques such as BDI CESD-R etc. Doctor can't get the complete information from depressed one using patients self-reporting; in one place at once, they can't be able to get correct info from the subject. The social-media activities overcome some of the problems regarding patient self-reporting. From user's social activities; we may get closer to the natural behavior of the user and his way of thinking. If we apply SMPP in the domain of psychology we can achieve enormous benefits which otherwise be either difficult to gain mistakably e.g. questioners, etc. The APP will help to automatically retrieve the natural position of subject over the time and also gives alerts if any changes in individual's behaviors.

6.2. Conclusion and Future Work

We have verified the likely the use of Facebook as a tool for assessing and predicting major depression in individuals. First we collected depression related status update and made the Training Dataset manually according the CESD-R Scale. After that we select a diversity of behavioral attributes which related to person's mood, thinking style, interaction and activities. User's posting on Facebook revealed the feeling of guilt, feeling of tried, sleeping related problem, helplessness and worthlessness that are the symptoms of MDD.

Finally, we used these selected attributes and build a classifier using SVM and K-NN Classifiers that can expose mental related problem in individuals from social media activities. The SVM classifier gives hopeful results having 77% classification accuracy.

Currently we were selected minimal attributes from Facebook user's profile, In future we will also include other attributes such as user likes etc.; in addition we haven't applied filter for self-referencing phrase in Model, and in our next version we will overcome that problem too.

Our current model is based on hand-labeling training dataset which is very expansive and required domain experts, for best performance we need large amount of hand-labeling training dataset. In future we will apply the bootstrapping techniques that are used to learn from huge sets of labeled data.

References

[1]. Gamon, M., Counts, S., & Horvitz, E. (n.d.). Predicting Depression via Social Media, 2.

[2]. Gamon, M., Counts, S., & Horvitz, E. (n.d.). Predicting Depression via Social Media, 2.

[3]. Moreno, M. a., Jelenchick, L. a., & Kota, R. (2013). Exploring Depression Symptom References on Facebook among College Freshmen: A Mixed Methods Approach. Open Journal of Depression, 02(03), 35–41. doi:10.4236/ojd.2013.23008

[4]. Moreno, M. a, Jelenchick, L. a, Egan, K. G., Cox, E., Young, H., Gannon, K. E., & Becker, T. (2011). Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depression and anxiety, 28(6), 447–55. doi:10.1002/da.20805

[5]. Park, M., Cha, C., & Cha, M. (2012). Depressive Moods of Users Portrayed in Twitter.

[6]. Hamouda, S. Ben, & Akaichi, J. (2013). Social Networks ' Text Mining for Sentiment Classification : The case of Facebook ' statuses updates in the " Arabic Spring " Era, 2(5), 470–478.

[7]. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences of the United States of America, 110(15), 5802–5. doi:10.1073/pnas.1218772110

[8]. Farnadi, G. (2013). How Well Do Your Facebook Status Updates Express Your Personality ?

[9]. Rahman, M. M. (n.d.). Mining Social Data to Extract Intellectual Knowledge

[10]. http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

[11]. Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. In- ductive learning algorithms and representations for text categorization. In CIKM '98: Proceedings of the seventh international conference on Information and knowledge management, pages 148–155. ACM Press, 1998.

[12]. Yiming Yang and Xin Liu. A re-examination of text categorization methods. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

[13]. Gordon Rios and Hongyuan Zha. Exploring support vector machines and random forests for spam detection. In CEAS, 2004.

[14]. Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Machine Learning: ECML-98, pages 137–142. Springer, 1998.

[15]. http://en.wikipedia.org/wiki/Tokenizing

[16]. Andrade L, Caraveo-A. 2003. Epidemiology of major de- pressive episodes: Results from the International Consorti- um of Psychiatric Epidemiology (ICPE) Surveys . Int J Methods Psychiatr Res.12(1):3–21.

[17]. http://nlp.stanford.edu/IR-book/html/htmledition/k-nearest-neighbor-1.html

[18]. http://nlp.stanford.edu

[19]. http://www.liwc.net4

[20]. http://en.wikipedia.org/wiki/Facebook

[21]. http://cesd-r.com/cesdr/
[22]. http://rapidminer.com/
[23]. Gannon, Megan. "Facebook profile may expose mental illness." (2013).
[24]. http://en.wikipedia.org/wiki/List_of_emoticons