**A. Description of Folders**

* **thesis\_code\_submission**: The folder containing sub-folders for Experiments and Proposed Solution.

1. **Experiments folder**:
2. **code\_files+dataset** : Folder containing all files for experimental module at Tweet Level.
   * + 1. **word2vec.model**: File containing pre-processed word vectors.
       2. **word2vec\_model.bin**: File generated by Word2vec trained embeddings.
       3. **glove.6B**:Folder containing pre-trained Glove embeddings.
       4. **users\_having\_info.csv**:This file contains Information of attributes related with the Tweet of users which will be further utilized to extract the user’s profile information under 1.1.5.
       5. **users\_info\_all\_combined.csv**:This file contains information of users combined using ten csv files from the **dataset** sub-folder. However, the file (**users\_info\_all\_combined.csv**) contains information extracted from each user’s profile which will be utilized for data visualization and analysis.
     1. all\_user\_details\_all\_eight.csv
     2. all\_user\_details\_all\_five.csv
     3. all\_user\_details\_all\_four.csv
     4. all\_user\_details\_all\_nine.csv
     5. all\_user\_details\_all\_seven.csv
     6. all\_user\_details\_all\_six.csv
     7. all\_user\_details\_all\_ten.csv
     8. all\_user\_details\_all\_three.csv
     9. all\_user\_details\_all\_two.csv
     10. all\_user\_details\_all.csv

|  |
| --- |
| Index(['id', 'conversation\_id', 'created\_at', 'date', 'time', 'timezone','user\_id', 'username', 'name', 'place', 'tweet', 'mentions', 'urls', 'photos', 'replies\_count', 'retweets\_count', 'likes\_count', 'hashtags', cashtags', 'link', 'retweet', 'quote\_url', 'video', 'near', 'geo', 'source', 'user\_rt\_id', 'user\_rt', 'retweet\_id', 'reply\_to', 'retweet\_date', 'translate', 'trans\_src', 'trans\_dest', 'target', 'category', 'lower\_username', 'person\_name', 'person\_username' 'person\_followers\_count', 'person\_listed\_count', 'person\_following', 'person\_favorites', 'person\_verified', 'person\_default\_profile', 'person\_location', 'person\_statuses\_count', 'person\_description', 'person\_geo\_enabled', 'person\_contributors\_enabled', 'account\_creation\_date', 'year', 'month', 'day', 'hour', 'text', 'source', 'direct\_reply', 'retweet\_status', 'retweet\_count', 'favorite\_count', 'hashtags\_count', 'urls\_count', 'user\_mentions\_count', 'lower\_username'], dtype='object') |

**Table 1 Attributes of user information (experiments)**

Table 1 shows the list of attributes of user information under the (**users\_info\_all\_combined.csv)** file.

1. **Dataset sub-folder**:
2. **df\_depressed\_tm.csv**: This file contains all the tweets with depression related keywords such as (suicide, kill, depressed etc.)
3. **df\_tweets\_happy\_tm.csv**: This file contains all the tweets with non-depressive keywords such as (amazing, cheerful, happiest etc.)
4. **lat\_long\_info.csv**: This file contains information of users’ location converted into latitude and longitude for accurate visualization. However, this file plays a vital role because some of the users having null or incorrect location gets removed. For instance, a user location: (“enjoying in heaven”) or a user with location:<URL>.
5. **Proposed Solution**:

**2.1 code\_files+dataset**:Folder containing all files for Proposed Solution module at User Level.

1. **clean\_dep\_ff.csv**: This file is generated by extracting and storing meaning of emojis with all the cleaned depressed tweets. This file is created because the extraction of emojis takes a lot of time which can be saved by loading this file directly into the python notebook and skipping the extraction of emoji step.
2. **clean\_happy\_3lakh.csv**: This file is generated by extracting and storing meaning of emojis with all the cleaned Non-depressed tweets. This file is created because the extraction of emojis takes a lot of time which can be saved by loading this file directly into the python notebook and skipping the extraction of emoji step.
3. **depressed\_features\_list\_322.csv**: This file is generated for storing the values of extensive feature engineering operated on 322 Depressed users. This file is created in order to save the time under pre-processing and the data cleaning by directly loading the file into the python notebook.
4. **proposed\_sol\_users.csv**: This file is generated for storing the values of extensive feature engineering operated on all 760 users under proposed solution. This file is created in order to save the time under pre-processing and the data cleaning by directly loading the file into the classification model.
5. **df\_tweets\_dep\_data.csv(extra file)** : This file was created to store the uncleaned tweets of all the 322 depressed users into a single file.
6. **Dataset sub-folder**:
7. **all\_user\_details\_proposed\_sol.csv**: This file contains all the attributes for

depressed and non-depressed user’s profile as following:

|  |
| --- |
| Index(['person\_name', 'person\_username', 'person\_followers\_count', 'person\_listed\_count', 'person\_following', 'person\_favorites', 'person\_verified', 'person\_default\_profile', 'person\_location', 'person\_time\_zone', 'person\_statuses\_count', 'person\_description', 'person\_geo\_enabled', 'person\_contributors\_enabled', 'account\_creation\_date', 'year', 'month', 'day', 'hour', 'text', 'tweet\_latitude', 'tweet\_longitude', 'source', 'in\_reply\_to\_screen\_name', 'direct\_reply', 'retweet\_status', 'retweet\_count', 'favorite\_count', 'hashtags', 'hashtags\_count', 'urls', 'urls\_count', 'user\_mentions', 'user\_mentions\_count', 'media\_type', 'tweet\_contributions'], dtype='object') |

**Table 2 Attributes of user information (Proposed solution)**

1. **df\_depressed\_users\_twint.csv**: This file contains the list of all the users who have self-declared for being in depression in their tweets, extracted by keyword matching using Twint.
2. **depressed\_tweets\_three/ tweets.csv**: This file contains list of all the users who have self-declared for being in depression in their tweets as well as posted tweets indicating symptoms of depression extracted by keyword matching using Twint. However, the file under 2.2.2 and 2.2.3 is further combined and processed together to generate a list of 322 unique users for the final dataset of Depressed Users.
3. **happy\_user(x)**: These are the folders generated by extraction of tweets for each Non-depressed user. All the folders are named as happy\_user(x), where range for x is (1,600). However, the total number of folders used by Non-depressed users for proposed solution are 438.
4. **user(y)**: These are the folders generated by extraction of tweets for each depressed user. All the folders are named as user(y), where range for y is (1,500). However, the total number of folders used by depressed users for proposed solution are 322.

**B. Description of Dataset**

1. **Purpose of Dataset**: The dataset for experimental module was utilized to analyze and predict depression in users on a Tweet Level architecture. However, the dataset for proposed solution module was carried under a User Level architecture for studying and predicting depression in users on social media. To add to it, the dataset at User Level was used to visualize and analyze the pattern of user engagement on social media using data analysis and machine learning.

2. **Data Collection**: The dataset for both the experimental module (Tweet Level) and the proposed solution module (User Level) are collected from Twitter. The datasets were collected using Twint (Twitter Intelligence Tool) and Tweepy (Python library).

**Table 3 Dataset used for Tweet Level Architecture (Experiments)**

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Data Collected | Data cleaning | Pre-Processed Extensive Data Cleaning (Final Dataset -Tweet Level) |
| Depressed Tweets | 38850 | 29997 | **29997** |
| Non-Depressed Tweets | 100496 | 90409 | **89243** |
| Total Tweets | 139346 | 120406 | **119240** |

Table 3 shows the final Dataset generated after data cleaning used at Tweet Level Architecture.

**Table 4 Dataset used for the User Level module (Proposed solution).**

|  |  |  |
| --- | --- | --- |
| Type | Data Collected | Pre-Processed Extensive Data Cleaning (Final Dataset-User Level) |
| Tweets of **322** Depressed Users | 3,40,034 | **86,630** |
| Tweets of **438** Non-Depressed Users | 3,32,352 | **2,12,909** |
| Total Tweets of **760** Users | 6,72,386 | **2,99,539** |

Table 4 Shows the records for the final dataset generated after data cleaning of the records used at user Level Architecture.

3. **Procedures on Dataset**: The dataset was operated using the following procedures:

1. Extensive Data Cleaning: Removal of duplicate users as well as removal of duplicate tweets.
2. Extraction of emojis: Extracting the meaning of emojis using emojipedia under proposed solution.
3. Location: Conversion of location in terms of latitude and longitude.
4. Word2vec embeddings: Generating Trained word2vec embeddings under experimental model.
5. Extensive Feature Engineering: Various features were extracted using the feature engineering under proposed solution.

4. **Features of Dataset**: The features of dataset for proposed solution model are as following:

1. Extreme Depress Status
2. Night Status
3. Polarity Contrast
4. Negative Polarity Tweets
5. Lexical Richness
6. User Mentions
7. Social Response Ratio
8. Depressive Words Intensity
9. Nouns per Sentence
10. Proper Noun
11. Adjective Ratio.
12. Adverb Quotient

Nevertheless, the features of dataset for Tweet Level represented under Table 1 (section 1.1) can be utilized by implementation the following:

1. Data Visualization (matplotlib, plotly, pyplot etc. )
2. TF-IDF (Term Frequency- Inverse Document Frequency)
3. LDA (Topic Modeling)
4. NLTK- NER (Named Entity Recognition)
5. NLTK- POS (Part of Speech) Tagging
6. Sentiment Analysis and Polarity
7. GloVe Embeddings
8. Word2Vec Embeddings