**Model building**

Approach1 :- LGBM algorithm Light gradient boosting machine

Feature building :-

Response Variable :- effectiveness

(1) effectiveness-> Driven from Rating column of dataset where We have make to category 1, drug which are having rating above 5 and rest is 0.

effectiveness = 1 [rating >5]

effectiveness = 0 [rating <5]

Predictor variable :-

(2) usefulCount -> Dataset column

(3) review -> Dataset column

(4) review\_clean-> Driven from review column eliminated below properties

* removing white space from both side of line
* lower case conversion
* removing of Apostrophes
* removing exclamatory sign
* removing of m percent sign
* removing quotation marks
* Removing all the special Characters
* Removing all the non ASCII characters
* Removing the leading and trailing Whitespaces
* Replacing Two or more dots with one

(5) day

(6) month

(7) year

(8) sentiment -> Driven from review column

(9) sentiment\_clean -> Driven from review\_clean column

(10) count\_word -> Driven from review\_clean, totoal number or word in review

(11) word\_unique\_count -> Driven from review\_clean, total unique words

(12) count\_punctuations-> Driven from review\_clean

(13) count\_letters -> Driven from review\_clean

(14) count\_words\_upper -> Driven from review

(15) count\_stop\_words -> Driven from review

(16) mean\_word\_len-> Driven from review\_clean

(17) drugNameNumeric -> Numeric format of Drug Names

(18) conditionNumeric -> Numeric format of conditions

(19) reviewNumeric -> Numeric format of review [not used]

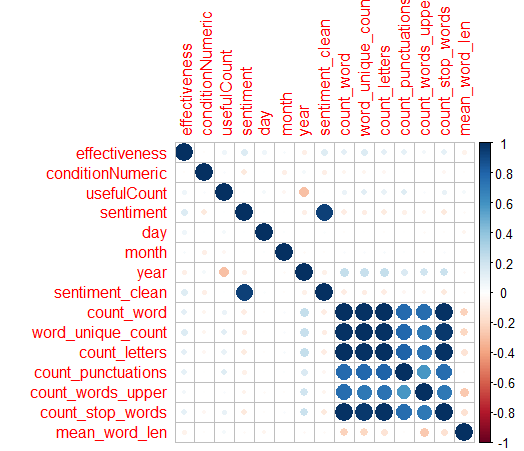
(20) ratingNumeric -> Numeric format of rating

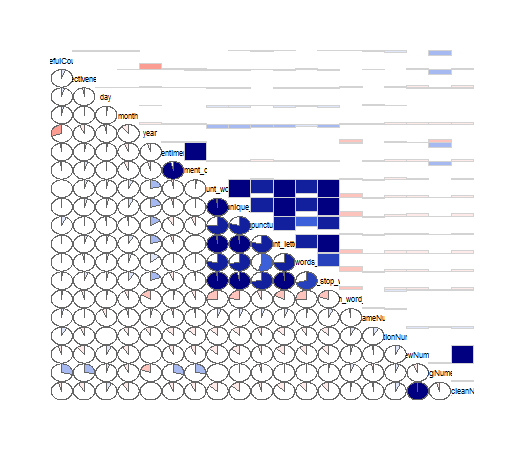
(21) review\_cleanNumeric -> -> Numeric format of review clean [Not used ]

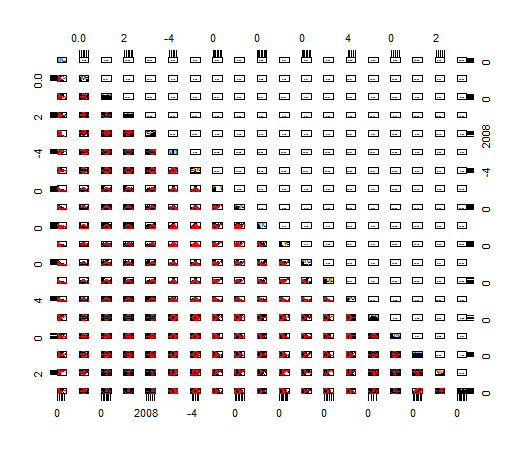
Sample Method and ration -> sample(2,nrow(df7),replace = T, prob = c(0.8,0.2))

Benefit -> It will shuffle the dataset and from random places makes the two set of data in 80:20 ratio

Plots for understanding the correlation in different features







Accuracy Got -> .433

**Approach 2:- Redesign the dataset**

**Step1-> Calculate the sentiments**

describe(df6$sentiment)

df6$sentiment

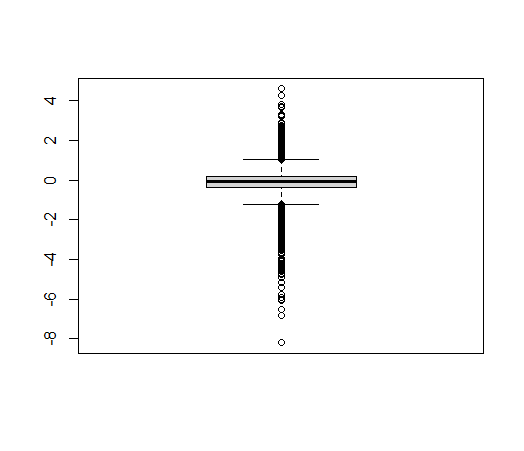
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

213786 0 3786 1 -0.09908 0.5681 -0.95833 -0.70000 -0.36667 -0.07143 0.20 0.485 0.70

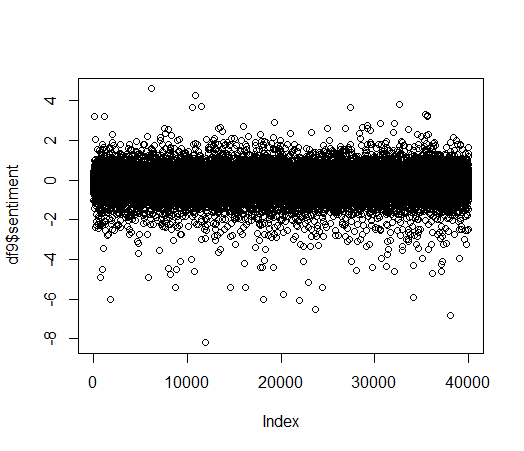
lowest : -8.550 -8.200 -7.300 -7.100 -6.925, highest: 4.750 4.850 5.100 5.200 5.600

**Boxplot and plot diagram**

**Boxplot Sentiments**



**Plot :- Sentiments**



I have processed the data for 40k records and after eliminating the outliers I have got total 38842 observation

Outlier Criteria :- **filter(df9,df9$sentiment >= -1.3 & df9$sentiment <= 1.3)**

Observation Sentiments below -1.3

Observation Sentiments above 1.3

**Challenges in this column :-** Not giving true relation between review and rating.

When I applied the test with one problem statement

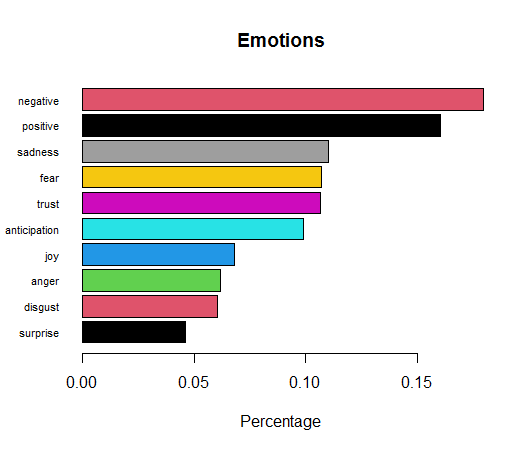
Find out the total count of observation which are having higher rating but negative sentiments

**filter(df\_filtr,df\_filtr$sentiment < -0.068 & df\_filtr$ratingNumeric > 5)**

72598 counts which satisfying the condition out of 213000 records

which is more than 33% percent of entire dataset.

**Step2-> Calculate the NRC sentiment**

****

**nrc\_sentiment formate**

anger anticipation disgust fear joy sadness surprise trust negative positive

0 4 0 1 1 1 0 3 3 3

In calculation used only two factor negative or positive

|  |  |  |  |
| --- | --- | --- | --- |
| Value1 | Condition | Value2 | Final value |
| negative | > | positive | -1 |
| negative | = | positive | 0 |
| negative | < | positive | 1 |

**Challenges in this column :-** Not giving true relation between review and rating.

When I applied the test with one problem statement

Find out the total count of observation which are having higher rating but negative nrc sentiments

**10500** out of **40000** which is 25% of entire dataset but some how I improved from simple sentiment statement.

**Step3:- Calculating polarity**

**Polarity output**

**all total.sentences total.words ave.polarity sd.polarity stan.mean.polarity**

**all 1 140 0.017 NA NA**

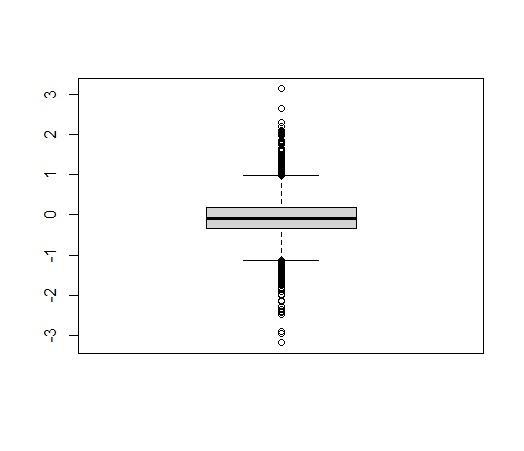
n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95

40000 0 8383 0.999 -0.08519 0.4725 -0.78928 -0.61632 -0.34641 -0.07958 0.18107 0.42701 0.58183

#

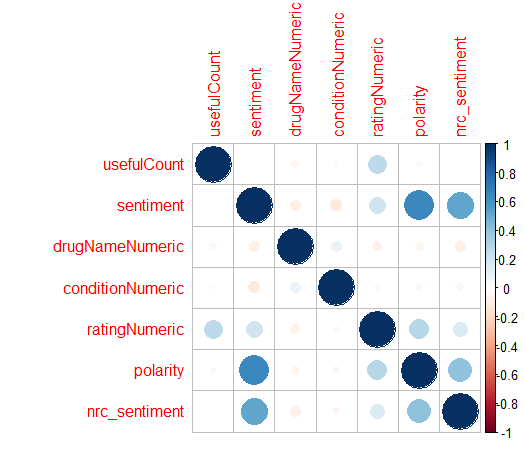
# lowest : -3.175217 -2.948123 -2.899138 -2.463361 -2.417546, highest: 2.200000 2.204541 2.300000 2.645751 3.130495

**Boxplot Polarity**



**Multiple Linear Regression**

**corplot**

****

Corgram

