```
library(randomForest)


#train_tf_idf = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_train_tf_idf.csv")
train_wc = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_train_wc.csv" )
#test_wc = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_test_wc.csv")

features = train_wc[0:3000,c(-1,-502)]
label = as.factor(train_wc$Predict[0:3000])
features_v = train_wc[3000:4000,c(-1,-502)]
label_v = as.factor(train_wc$Predict[3000:4000])

bestmtry <- tuneRF(features,label, ntreeTry=100,
     stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE, dobest=FALSE)

rf <- randomForest(x=features, y=label, mtry=163, ntree=500,
     keep.forest=TRUE, importance=TRUE)

rf2 <- randomForest(x=features, y=label, mtry=163, ntree=500, classwt=c(2474,526), importance=TRUE)
rf3 <- randomForest(x=features, y=label, mtry=163, ntree=500, classwt=c(526,2474), importance=TRUE)



#=====================================================

rf.pr = predict(rf2,newdata=features_v)
error = mean(rf.pr!=label_v)
library(Epi)
ROC(form=label_v~rf.pr, plot="ROC")
important_varibles = importance(rf,type=1)
selected_features = important_varibles[order(important_varibles,decreasing=T),]
top_features = selected_features[selected_features>4]
# top_features = selected_features[:225,]

selected_new_features = features[,selected_features>4]
bestmtry <- tuneRF(selected_new_features,label, ntreeTry=100,
     stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE, dobest=FALSE)

rf_selected_features <- randomForest(x=selected_new_features, y=label, mtry=33, ntree=500,
     keep.forest=TRUE, importance=TRUE)
#=====================================================
train_tf_idf = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_train_tf_idf.csv" )
train_wc = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_train_wc.csv" )
#test_wc = read.csv("/Users/ouyamei/Documents/GitHub/kaggle-crisis/data/kaggle_test_wc.csv")

features_tf = train_tf_idf[0:3000,c(-1,-502)]
#label_tf = as.factor(train_tf_idf$Predict[0:3000])
features_wc = train_wc[0:3000,c(-1,-502)]
label = as.factor(train_wc$Predict[0:3000])
features = cbind(features_tf,features_wc)
#label = cbind(label_tf,label_wc)

features_v = train_wc[3000:4000,c(-1,-502)]
label_v = as.factor(train_wc$Predict[3000:4000])

bestmtry <- tuneRF(features,label, ntreeTry=100,
     stepFactor=1.5,improve=0.01, trace=TRUE, plot=TRUE, dobest=FALSE)

rf <- randomForest(x=features, y=label, mtry=163, ntree=500,
     keep.forest=TRUE, importance=TRUE)

rf2 <- randomForest(x=features, y=label, mtry=163, ntree=500, classwt=c(2474,526), importance=TRUE)
rf3 <- randomForest(x=features, y=label, mtry=163, ntree=500, classwt=c(526,2474), importance=TRUE)
```