Data Mining Project Steps :

Abstract:

- The ease of access and rapid improvements in social media has enabled more than half of the world's population using it in their daily life
- With the flexibility of anyone can share anything over the platform, Social Media is more prone to the spread of irrelevant and misleading information.
- This extensive spread of spam has the potential for extremely negative impacts on individuals and the society.
- Therefore, detecting the spam or false information on social media has gained attention of many researchers.

1) Dataset Collection:

Collected the dataset of 2011 UK riots which turned out to be a major issue due to the spreading of fake news .

Get the data

- Tweet Text
- Tweet Date/Time
- Retweets.
- Location of Tweet
- User Details

How ?

- Using the Streaming API and Trends API over a given period of time.
- Scrape the HTML web page by scrolling infinitely to get the historical tweets

2) Data preprocessing/Cleaning:-

We will load the extracted file using the **read_csv()** function in R . Once the file is loaded, we have modified the datatypes of each column of our dataset using **as.datatype()** function in R and removed unwanted characters/white spaces from each tweet using **gsub() pattern replacement** function in R. Then we checked for fields having missing values in our dataset, and removed the missing rows using **na.omit()** function in R.

3) Feature Selection :

a) **No of URL's** : We found the number of occurrences of URL's in each tweet in our extracted data using **str_count()** function in R **.**

b) **No of Emoticons** : We found the number of occurrences of Emoticons such as smiley etc. in each tweet in our extracted dataset using **str_count()** function in R.

c) **Length of Tweet** : We found the length of each tweet in our raw data using **nchar()** R function.

d) **No of @ Mentions** : Found the number of occurrences of '@' symbol in each of the tweet using **str_count() function.**

e) **No of Hashtags :** Found the number of occurrences of '#' symbol in each of the tweet using **str_count()** function.

f) **Length of User Screen Name :** Found the length of the username that appears on the user's screen on his/her twitter account using **nchar()** function.

g) **Frequency of spam words** : We looked for the spam words (ex: London Zoo etc.) corresponding to the event UK Riots and found the occurrence of these words using **coll() pattern searching function** in R .

h) **Frequency of swear words:**We looked for the spam words (ex: burnt etc.) corresponding to the event UK Riots and found the occurrence of these words using **coll() pattern searching function** in R .

4) Feature Extraction Techniques:

- **PCA (Principal Component Analysis):**
  - The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset.
- **Fisher Linear Discriminant**
  - This algorithm is used to increase the distance between the classes; thereby increasing the separability between them. Since the gap between the classes increases, the classification will be easy.

5) Modelling/ Classification:

SVM:

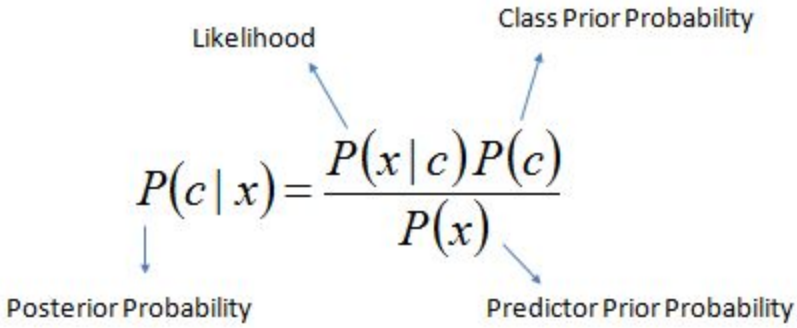
- SVM is the major algorithm used for modelling our preprocessed data.
- SVM is a classification algorithm used for classifying the data and then rank the data using SVM rank algorithm.
- We currently use Linear kernel for the SVM algorithm implementation.
- Find the hyperplane to segregate the classes as Spam or Non-Spam news in the given training data, i.e., given the labeled training data (*supervised learning*), the algorithm outputs an optimal hyperplane which categorizes new examples(test data).The notation used to define a hyperplane is:

$$f(x) = \beta_0 + \beta^T x,$$

where β is known as the *weight vector* and β_0 as the *bias*.

Naive Bayes:

Naive bayes is the classification technique used to derive the classification model for extracted data. Using prior probability and likelihood for the feature , we will be compute the probability of each class for the data using posterior probability. Based on the calculated posterior probability, data in the data set is classified.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Decision tree :

Decision Tree is another supervised learning algorithm that is used for classification. The tree is constructed by computing the Information gain of each of the attributes with respect to the target variable and selecting the attribute having the maximum Information gain. In our project, we are using Decision Tree to compare our model output results i.e accuracy with that of the **decision tree** inbuilt function in R.

KNN:

KNN can be used for both classification and regression predictive problems. For each data points we take the voting among its neighbours. The data point is assigned to the class which has got more votes from the neighbour nodes. In our project, we are classifying the data into credible and noncredible, the data is classified as credible if the data is surrounded by credible data points or else classified as non credible.

6) Post Processing:

Based on the rank that we get from the output of SVM algorithm, we have planned to Assign weights to the classified data .This is used to tell how much credible a given tweet is.