**Approach**

The give data is studied and an approach and methodology are built.

The first step involved is cleaning the data

Figure 2

All the data not related to any of the banks(A,B,C,D) are removed

Cleaned Dataset

Dataset

From the cleaned dataset, Data related to banks are separated. This is crucial step.

All the junk files are removed.

For example, comment no 8 is removed as it contains “8|8/15/2015|2015|8|twitter|i just invested in tsemah on twit\_hndl. listen and invest at INTERNET# tradiio” This could be related to any of the banks but because of the scrubbing, it is impossible for us to find out which bank this is related to or what comment it is referring to. Other junk messages such as 69 “- facebook wife wife? Ÿ ’ Ÿ˜:- \* Ÿ ’ – Ÿ ” ¥ Ÿ ’ ¦ Ÿ ’ ¦” are also removed.

From the output from this is

Data containing the bank IDs(Tweeter) are placed in another file named Bank\_tweeter

The left out Data related is classified further into

Data related only to bank A, only to BankB, only to bankC, only to bankD.

Data related only to bank A and B(Data which includes both bank A and bank B only )

Similarly Data related to banks A and C, data related to banks A and D etc

Data related only to only three banks ABC,ABD,BCD,etc

Data related only to all banks(ABCD)

This is done to ease the analysis.

For instance, when we perform sentiment analysis on the bank A data. It reflects the sentiments or substance on bankA. Similarly, for other banks we can get the direct substance and topics discussed.

When data related to two banks is analyzed, its relation can be found out.

Similarly when other data is used to interpret relations between banks, comparisons etc

Mostly, when data involves all the banks, it reflects the trend of the banking sector in general.

Twitter data contains tweets tweeted by banks, they could be an advertising

Other twitter data includes tweets directed to ID of bank, such comments are either complaints or praises.

Only the data in which IDs of banks are referred without mentioning the name of the bank are placed here.

This competes the steps of data preparation and cleaning.

Cleaned

Data

**BANK\_A**

**BANK\_B**

**BANK\_C**

**BANKD**

**BANK\_A\_B**

**BANKA\_C**

**BANKA\_D**

**BANKB\_C**

**Twitter data with bank IDs**

**BANK\_A\_B\_C\_D**

**BANK\_A\_B\_C**

**And so on**

The next step is putting this data into a compiler which removes the symbols, stop words etc

The terms such as “INTERNET” and “Name” make very little sense as they are scrubbed.

Such scrubbed data can be neglected.

Final step involved producing a summary/substance of each comment for the respective bank.