**Sentiment Analysis**

The task is to detect the sentiments an article expresses.

**Dataset**

The dataset contains 27731 rows, each containing a sentence/article and a ***sentiment*** seperated by semicolon (***;***). The dataset contains data of 12 different sentiments. The sentments are given below:

['Love', 'Like ', 'Consciousness ', 'Protestant ', 'Smiley ', 'Angry ', 'Blush', 'Skip ', 'Rocking ', 'Fail ', 'Shocking ', 'WOW', 'Bad ', 'HaHa', 'Sad ', 'Skeptical ', 'Evil ', 'Provocative ']

**Dataset Analysis**

In our analysis the dataset is very bad to work with for the task at hand.

1. The dataset claimed that document , sentiment pairs are seperated by semicolon (***;***), but actually it contained semicolon (***;***) inside documents too. An example is given below.

জীবনের নিরাপত্তা ও জীবনের মুল্য ২টিই বেশী; বাংলাদেশে গরূর মুল্য বোধ হয় মানুষ থেকে বেশী হওয়ার কথা।;HaHa(হা হা)

This made reading the dataset a bit hard for us.

1. We found the dataset inconsistant and meaning less. The same sentence is tagged with different sentiments at different times.

থ্যাংকস সুমন ভাই।;Like (ভাল)

থ্যাংকস সুমন ভাই।;Smiley (স্মাইলি)

থ্যাংকস সুমন ভাই।;Angry (রাগান্বিত)

থ্যাংকস সুমন ভাই।;Consciousness (চেতনাবাদ)

থ্যাংকস সুমন ভাই।;Skip ( বোঝতে পারছি না )

থ্যাংকস সুমন ভাই।;HaHa(হা হা)

থ্যাংকস সুমন ভাই।;WOW(কি দারুন)

থ্যাংকস সুমন ভাই।;Love(ভালবাসা)

থ্যাংকস সুমন ভাই।;Blush(গোলাপী আভা)

.

.

.

What sentiment exactly the sentence ‘থ্যাংকস সুমন ভাই।’ represents???

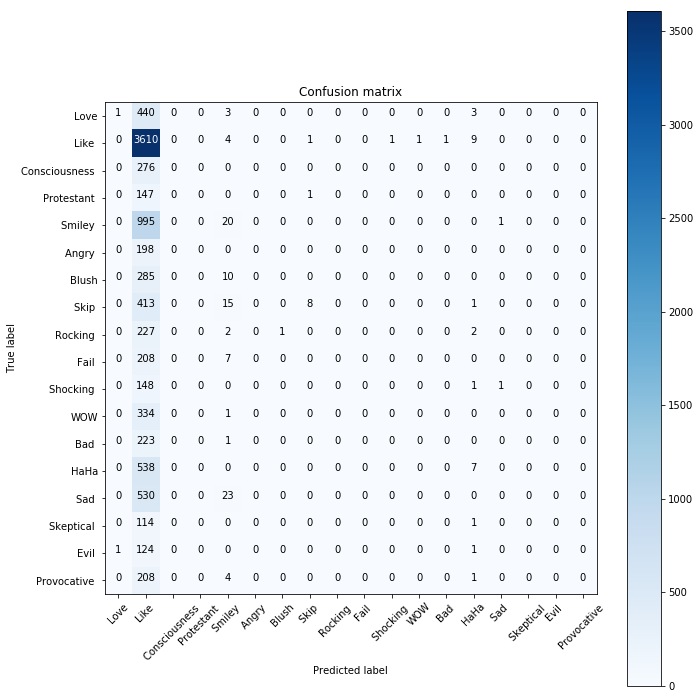
1. The same document-sentiment tuple occurs multiple times. The tuple

থ্যাংকস সুমন ভাই।;Like (ভাল)

occurs 17 times in this dataset !!!

**Experiment**

Despite the shortcomings of the dataset we actually did some experiments on it. Our feature was TFIDF value of the dataset. We used Naïve Bayes as our classifier. As expected the accuracy was very low, giving us an accuracy rate of only 39.8%. The confusion matrix is given below.



Then we thought that, may be the task is not a classification one. Then we built 12 separate classifiers, one for each of the sentiments to check whether an article expresses the sentiment or not. The results were still not very convincing.