Why Machine Learning for our problem?

There are certain human tasks which cannot be resolved using deterministic, rule-based solution. Large number of overlapping factors are involved to come to an accurate answer. Machine learning is used for such types of tasks. Moreover, these tasks are oftentimes scalable in nature.

Model selection and how we have trained it?

Target and Features:

* There are 2 components when training a machine learning model – target and input features.
* Target is/are the values you want to predict. For our application, target is the ‘tweet sentiment’ which has the following values:
  + Negative (-1)
  + Neutral (0)
  + Positive (1)
* Input features are the preprocessed, algorithm specific data which you feed into the model so it can be trained to determine the target values. We are fetching tweets (raw-data) corresponding to specific airline via Twitter APIs. The ‘tweet text’ attribute of the csv is then cleaned and converted into a document-term matrix of token counts. For token counts, we have used ‘Bag of words’ approach.

The classifier

Ours is a supervised learning classification problem. We have chosen Random Forest Classifier for our project.

Random Forest Classifier:

This classifier divides a random subset from the training set into multiple decision trees. The final class is the aggregated vote from these decision trees.

Decision tree are used for predicting a target which is formed of discrete values. It takes factors like entropy, impurity, and information gain in its prediction. Higher the information gain lower the error rate of the tree. Decision tree stops its process when there are no features available to classify.

The process of multiple decision trees is used because a single tree may contain noise and leads to wrong results. On a contrary, aggregate of different decision trees reduces the effect of noise and leads to right prediction.

Some of the primary parameters to this classifier are no. of the trees to divide the sample set, criteria of split, etc. We have divided our training sample set in 200 decision trees.

Before providing the data to the classifier for training, we need to clean it. Raw tweets were cleaned for stop words and common words. Then a document term matrix was prepared. The classifier was trained using this matrix and another similar type of matrix was prepared for testing classifier and prediction. The testing data needs to use train matrix for its transformation. Failing to do so will lead to inaccurate results.

Other classifiers and why did we choose Random Forest classifier?

Several other classifiers are available for classification supervised learning problems - Logistic Regression, KNeighbors Classifiers, SVC, Decision Tree, AdaBoost, GaussianNB,etc.

We trained our data with all the classifiers and observed that Random Forest Classifier gave the highest accuracy. The reason is it works well with numerical data and gives better results with large training set, though its training time is more.

