1. Data Collection

The dataset for training is “Sentiment140”, which originated from Stanford University. The dataset can be downloaded from the below link.

http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

By looking at the description of the dataset from the link, the information on each field can be found.

0 — the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)

1 — the id of the tweet

2 — the date of the tweet

3 — the query . If there is no query, then this value is NO_QUERY.

4 — the user that tweeted

5 — the text of the tweet

The first five columns have been dropped as they are of no use for this particular problem.

2. Data Cleaning

A) All the ‘@’,’#’ and the links have been removed using Regular Expressions.
B) The punctuations have also been removed.
C) The text is tokenized using TweetTokenizer which is a special tokenizer for tweets.
D) Lastly the stop words are removed and steamming is done.

E) Finally the tokens are broken down into four grams.

F) Finally another column of sentence is added to dataframe which combines the first 3 grams.

3. Data Analysis

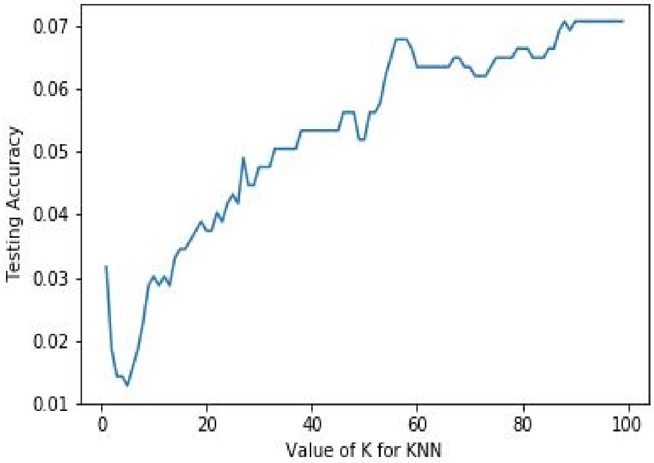
The data analysis has been done in two parts:-

A)Using TF-IDF

i) The sentence column is vectorized and taken as our independent variable and the 4th gram is our dependent variable.

ii) Now various classifier models have been used like knn, Random Forest, SVM,etc.

iii)K for KNeighborsClassifier is decided by plotting graph between Value of K and Testing Accuracy.



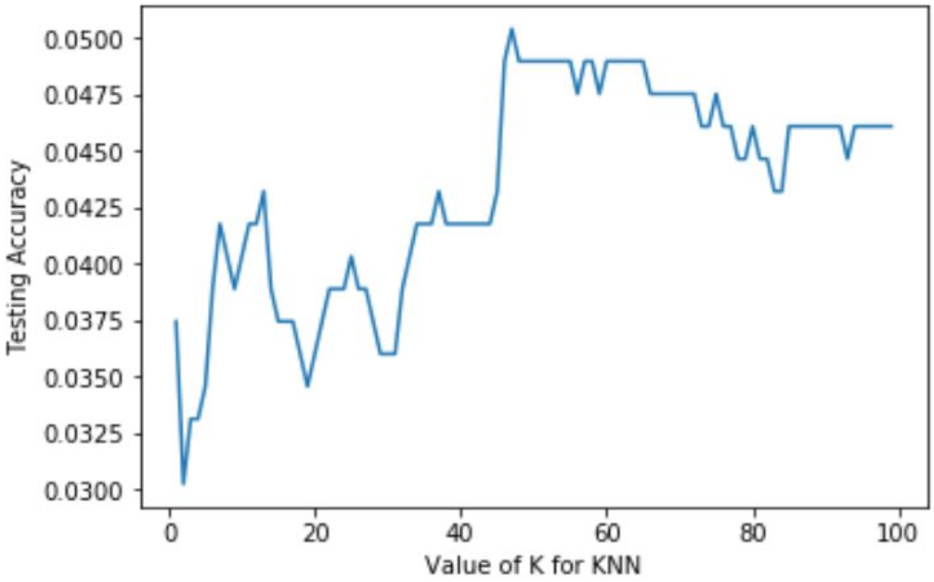
iv)For svm its many variations have been used to get maximum accuracy.

B)By using simple label encoding

i)The first 3 grams are taken to be our independent variables and are label encoded.

ii)Similarly various classifiers are used to solve our problem.

iii)K for KNeighborsClassifier is decided by plotting graph between Value of K and Testing Accuracy.

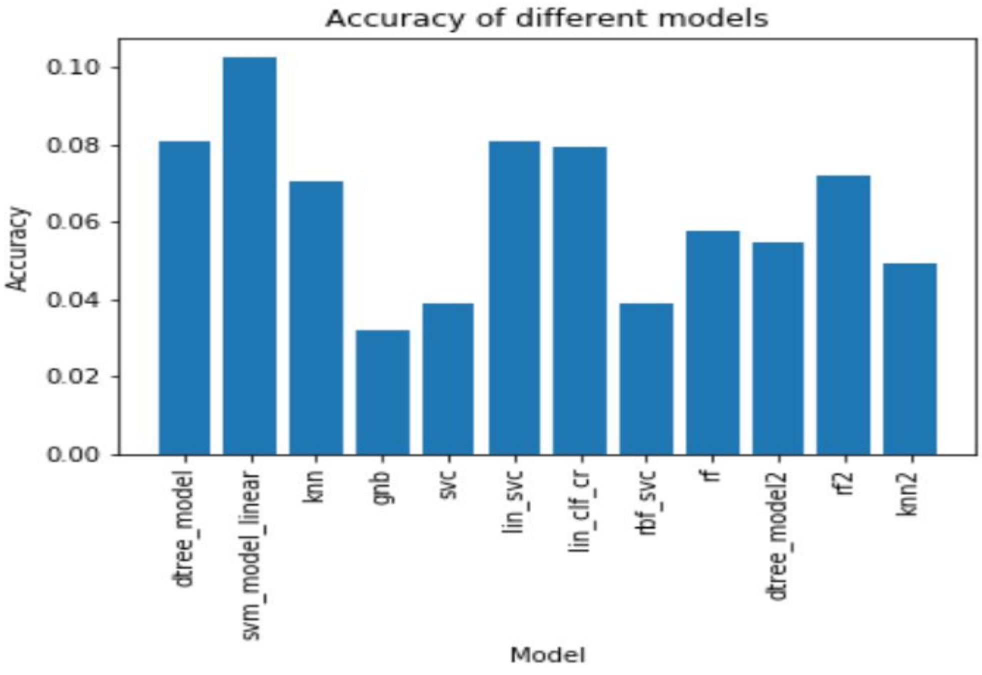


4. Results

Finally the accuracy of all the model is plotted using matplotlib and a histogram is made.

5. Conclusion

Therefore by looking at the graph we can say that maximum accuracy is obtained by the Linear SVC model which implements “one-vs-the-rest” multi-class strategy.
The highest accuracy achieved is 10.2%.


Accuracy of different models
Accuracy
dtree_model
svm_model_linear
knn
gnb
svc
lin_svc
lin_clf_cr
rbf_svc
rf
dtree_model2
rf2
knn2
Model