Abhishek Sharma (2419831s) Text as Data Coursework

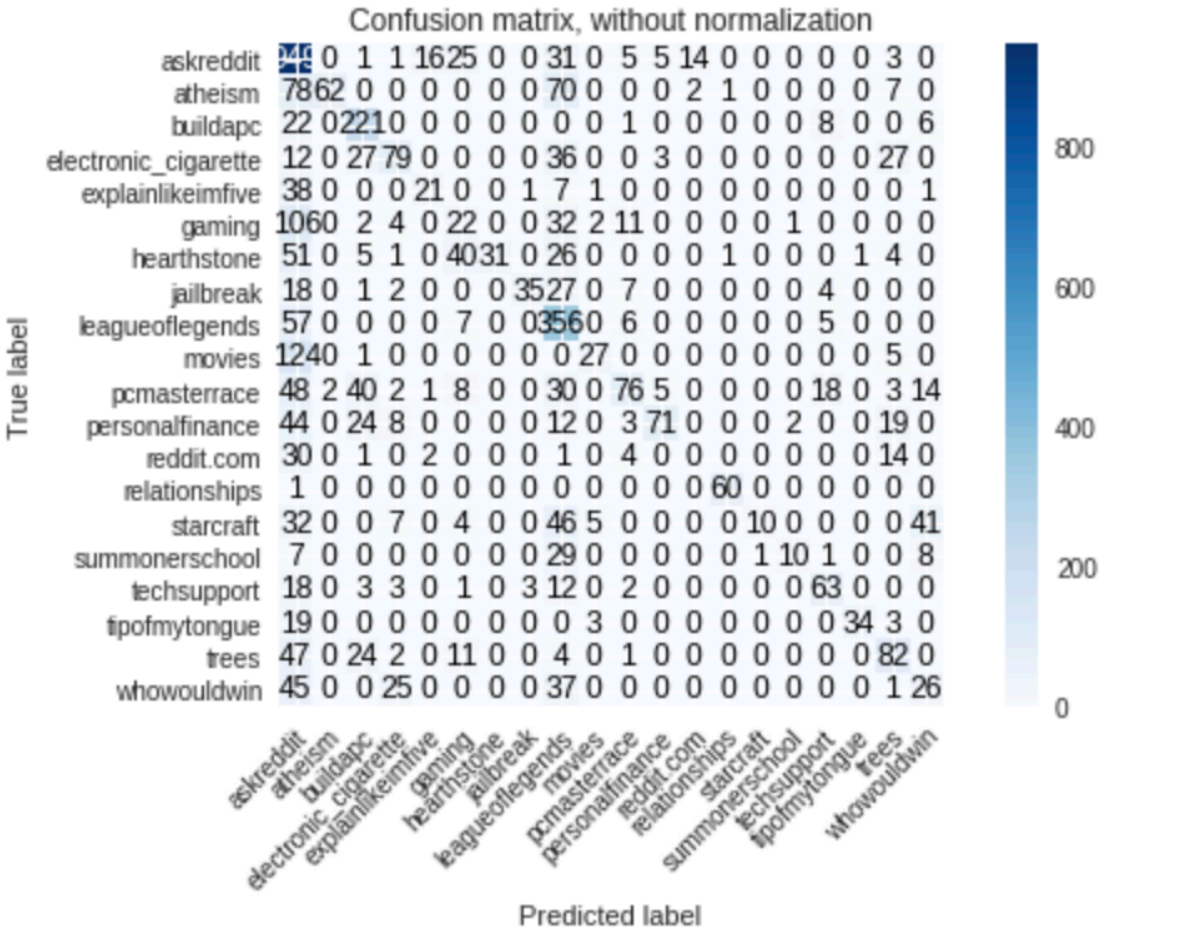
Q1

1) Tokenization and Normalization, they are two of the very important elements of text preprocessing. In order to make the text in machine-readable format for further steps, normalization is used. This step would remove punctuations, numbers or white spacing which would make the process of tokenization easier.

- Tokenization is the process in which the given text is divided or split into small tokens, this will help the model to analyze the text better.

2) Logistic Regression on TF-IDF got the highest Macro F1 score.

	Accuracy	Precision	Recall	F1-score
LR One hot Encoding	0.436	0.278	0.411	0.254
SVC One hot Encoding	0.269	0.057	0.095	0.034
BernoulliNB One hot Encoding	0.377	0.162	0.475	0.151
LR TFIDF	**0.557**	**0.412**	**0.652**	**0.450**
SVC TFIDF	0.261	0.050	0.013	0.021
BernoilliNB TFIDF	0.377	0.162	0.475	0.151



4) Encoding and classifier were effective as a combination were effective, as a combination of Logistic Regression which is the classifier and TF-IDF which is the encoder helped in increasing the numbers whereas the score got amplified in terms of precision, recall, f1-score and support. I

   Parameter 1 = 10000
   Parameter 2 = 5000
   Parameter 3 = 5000
   Parameter 4 = None

Q2)

1) After dividing the parameters into 4 parts and performing a grid search. There were some changes after tuning the Logistic regression on TFIDF, the precision increased in almost of every subreddit whereas it decreased in very few. Same goes for F1-score as well. Overall even Macro F1 score increased If we compare if it with the non-tuned model.

```
Evaluation for: LR TFIDF after Tuning
Classifier 'LR TFIDF after Tuning' has Acc=0.586 P=0.485 R=0.599 F1=0.506
                      precision    recall  f1-score   support

           askreddit      0.840     0.618     0.712      1427
             atheism      0.318     0.795     0.455        88
           buildapc       0.868     0.617     0.721       363
electronic_cigarette      0.500     0.609     0.549       151
   explainlikeimfive      0.348     0.296     0.320        81
              gaming      0.156     0.147     0.151       191
         hearthstone      0.312     0.893     0.463        56
           jailbreak      0.585     0.902     0.710        61
     leagueoflegends      0.787     0.600     0.681       565
              movies      0.325     0.614     0.425        83
        pcmasterrace      0.316     0.549     0.401       142
     personalfinance      0.574     0.700     0.631       150
          reddit.com      0.058     0.054     0.056        56
       relationships      1.000     0.744     0.853        82
           starcraft      0.193     0.875     0.316        32
       summonerschool      0.357     0.690     0.471        29
         techsupport      0.638     0.615     0.626       109
       tipofmytongue      0.661     0.812     0.729        48
               trees      0.596     0.543     0.568       188
        whowouldwin       0.261     0.307     0.282       114

           micro avg      0.586     0.586     0.586      4016
           macro avg      0.485     0.599     0.506      4016
        weighted avg      0.666     0.586     0.606      4016
```

2) After conducting an error analysis in the following scenario, I found out that if the classifier would work more efficiently if the post depth would be more and if less acronyms were used in the subreddits.

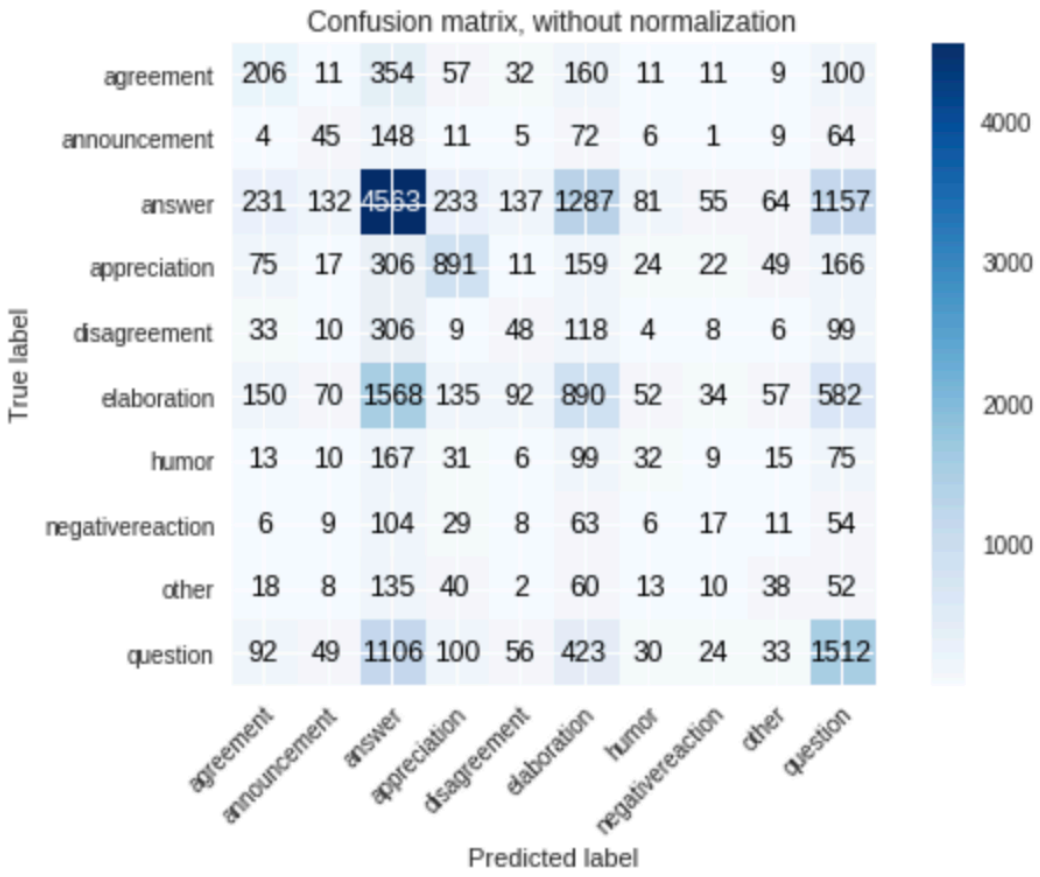
3) The feature that has been implemented is post_depth, which would help in analyzing the depth of the post the classifier is trying to reach.

Q3)
After implementation of the tuned classifier the Marco - precision, recall and F1 effectiveness dropped on the different data set.

```
Evaluation for: LR TFIDF
Classifier 'LR TFIDF' has Acc=0.416 P=0.242 R=0.260 F1=0.248
                  precision    recall  f1-score   support

       agreement      0.217     0.249     0.232       828
    announcement      0.123     0.125     0.124       361
          answer      0.575     0.521     0.547      8757
    appreciation      0.518     0.580     0.547      1536
    disagreement      0.075     0.121     0.092       397
     elaboration      0.245     0.267     0.256      3331
           humor      0.070     0.124     0.089       259
negativereaction      0.055     0.089     0.068       191
           other      0.101     0.131     0.114       291
        question      0.441     0.392     0.415      3861

       micro avg      0.416     0.416     0.416     19812
       macro avg      0.242     0.260     0.248     19812
    weighted avg      0.437     0.416     0.425     19812
```



After conducting an error analysis in the following scenario, I found out that if the classifier would work more efficiently if the total comment count could be given and post depth count which as well would help the model to run in a better way.

Q4)

Implemented Features -

a) Total Comments-

I have created a feature which is giving the total number of comments, this will help the model in getting the total number of comments as it was discovered in error analysis that the comments were not very clear in terms of the number.

```
Classifier 'feature 1' has Acc=0.460 P=0.205 R=0.298 F1=0.210
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classificatio
  'recall', 'true', average, warn_for)
                  precision    recall  f1-score   support

                      0.086     0.242     0.127       813
       agreement      0.134     0.555     0.215       229
    announcement      0.000     0.000     0.000         0
          answer      0.848     0.465     0.600     14479
    appreciation      0.544     0.735     0.625      1274
    disagreement      0.000     0.000     0.000        11
     elaboration      0.191     0.281     0.227      2469
           humor      0.000     0.000     0.000         2
negativereaction      0.000     0.000     0.000         0
           other      0.021     0.471     0.041        17
        question      0.431     0.527     0.474      2803

       micro avg      0.460     0.460     0.460     22097
       macro avg      0.205     0.298     0.210     22097
    weighted avg      0.667     0.460     0.522     22097
```

b) Post Depth

This feature will provide the post depth of the post, as we noticed in the error analysis that the deeper the post is, the better the classifier would work so to check the depth we implemented this feature.

```
  'recall', 'true', average, warn_for)
Classifier 'feature 2' has Acc=0.460 P=0.205 R=0.298 F1=0.210
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classificatio
  'recall', 'true', average, warn_for)
                  precision    recall  f1-score   support

                      0.087     0.245     0.129       811
       agreement      0.134     0.552     0.215       230
    announcement      0.000     0.000     0.000         0
          answer      0.848     0.464     0.600     14508
    appreciation      0.545     0.736     0.626      1273
    disagreement      0.000     0.000     0.000        11
     elaboration      0.189     0.280     0.225      2447
           humor      0.000     0.000     0.000         2
negativereaction      0.000     0.000     0.000         0
           other      0.021     0.471     0.041        17
        question      0.431     0.527     0.474      2798

       micro avg      0.460     0.460     0.460     22097
       macro avg      0.205     0.298     0.210     22097
    weighted avg      0.668     0.460     0.522     22097
```

c) Subreddit Author

This feature will provide the author of the subreddit which will help in identifying which subreddit belongs to which author.

```
Evaluation for: feature 3
Classifier 'feature 3' has Acc=0.382 P=0.216 R=0.261 F1=0.227
                  precision    recall  f1-score   support

                      0.161     0.179     0.169      2059
         agreement    0.169     0.268     0.208       600
      announcement    0.088     0.147     0.110       217
            answer    0.594     0.473     0.527      9984
      appreciation    0.528     0.602     0.562      1509
      disagreement    0.045     0.124     0.066       234
       elaboration    0.211     0.236     0.223      3248
             humor    0.028     0.109     0.045       119
  negativereaction    0.065     0.200     0.098       100
             other    0.074     0.168     0.103       167
          question    0.410     0.363     0.385      3860

         micro avg    0.382     0.382     0.382     22097
         macro avg    0.216     0.261     0.227     22097
      weighted avg    0.429     0.382     0.401     22097
```

d) Subreddit

This is the feature which will tell subreddit which the post would come from and after implementation it helped in identifying which posts belongs to which subreddit. The score did not improve

```
Evaluation for: feature 4
Classifier 'feature 4' has Acc=0.385 P=0.218 R=0.261 F1=0.228
                  precision    recall  f1-score   support

                      0.167     0.180     0.173      2119
         agreement    0.177     0.273     0.214       616
      announcement    0.079     0.130     0.099       223
            answer    0.591     0.481     0.530      9741
      appreciation    0.536     0.596     0.565      1546
      disagreement    0.045     0.119     0.066       244
       elaboration    0.212     0.232     0.221      3312
             humor    0.026     0.099     0.042       121
  negativereaction    0.052     0.178     0.081        90
             other    0.082     0.208     0.118       149
          question    0.429     0.373     0.399      3936

         micro avg    0.385     0.385     0.385     22097
         macro avg    0.218     0.261     0.228     22097
      weighted avg    0.429     0.385     0.403     22097
```

Total score of all the features combined –

```
Classifier 'LR TFIDF' has Acc=0.503 P=0.213 R=0.444 F1=0.221
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classif
  'recall', 'true', average, warn_for)
                  precision    recall  f1-score   support

       agreement       0.086     0.522     0.148       157
    announcement       0.000     0.000     0.000         0
          answer       0.853     0.504     0.634     13427
    appreciation       0.530     0.768     0.627      1187
    disagreement       0.006     0.286     0.012        14
     elaboration       0.211     0.311     0.252      2461
           humor       0.007     0.429     0.013         7
negativereaction       0.003     0.500     0.006         2
           other       0.024     0.562     0.046        16
        question       0.412     0.555     0.473      2541

       micro avg       0.503     0.503     0.503     19812
       macro avg       0.213     0.444     0.221     19812
    weighted avg       0.689     0.503     0.560     19812
```

Weighing of the classifier on different classes -

y=agreement top features		y=announcement top features		y=answer top features		y=appreciation top features		y=disagreement top features		y=elaboration top features		y=humor top features		y=negativereaction top features		y=other top features		y=question top features	
Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature	Weight[?]	Feature
+4.867	x167	+1.206	x224	+2.195	x4977	+10.150	x4425	+2.765	x2954	+1.553	x199	+1.461	x61444	+0.722	x2039	+1.642	x5529	+5.369	x4854
+3.953	x4972	+0.957	x1186	+2.150	x26710	+6.776	x4424	+1.435	x2939	+1.393	x4444	+1.053	x62319	+0.683	x1821	+1.376	x61656	+4.707	x2150
+2.581	x168	+0.897	x4823	+2.000	x224	+3.326	x381	+1.270	x2319	+1.314	x4427	+1.029	x41484	+0.608	x1914	+1.283	x28899	+3.901	x4875
+2.112	x3776	+0.832	x2999	+1.641	x2939	+2.886	x1953	+1.189	x1273	+1.271	x2382	+0.906	x60504	+0.580	x61931	+1.256	x45298	+3.813	x247
+1.964	x4594	+0.750	x4504	+1.604	x4427	+2.878	x1925	+1.150	x2046	+1.154	x2218	+0.774	x61041	+0.555	x4251	+1.169	x19494	+3.326	x250
+1.770	x1537	+0.608	x4427	+1.570	x2187	+2.590	x2929	+1.084	x4426	+1.065	x4504	+0.742	x61072	+0.513	x1117	+1.087	x3472	+3.077	x3047
+1.763	x4453	+0.481	x5529	+1.458	x23509	+2.426	x2174	+0.890	x649	+1.004	x2325	+0.710	x25762	*... 5244 more positive ...*		+0.896	x62180	+2.360	x1307
+1.678	x4517	+0.461	x2218	+1.422	x20101	+2.069	x1016	+0.851	x4427	+0.984	x4847	*... 5385 more positive ...*		*... 57404 more negative ...*		+0.830	x60836	+2.060	x1299
+1.671	x3698	*... 9316 more positive ...*		+1.397	x3367	+2.018	x2618	+0.827	x4444	+0.960	x2187	*... 57263 more negative ...*		-0.526	x3047	+0.773	x3259	+2.045	x4862
+1.664	x2174	*... 53332 more negative ...*		+1.370	x15667	+1.984	x1902	+0.811	x1299	+0.889	x77	-0.715	x438	-0.526	x4906	+0.765	x19754	+2.001	x4920
+1.538	x4977	-0.480	x2319	+1.348	x4985	+1.910	x4946	+0.794	x4423	+0.867	x4458	-0.733	x4453	-0.549	x2041	+0.763	x4880	+1.901	x4944
+1.214	x4976	-0.483	x16728	+1.333	x26992	+1.683	x4426	+0.791	x4449	*... 21263 more positive ...*		-0.772	x3047	-0.551	x681	*... 5032 more positive ...*		+1.836	x1246
+1.176	x1180	-0.526	x15667	*... 30833 more positive ...*		+1.539	x4453	+0.773	x2325	*... 41385 more negative ...*		-0.780	x2999	-0.554	x4504	*... 57616 more negative ...*		+1.711	x1186
+1.111	x4989	-0.529	x1299	*... 31815 more negative ...*		+1.477	x967	+0.723	x60703	-0.814	x4453	-0.793	x289	-0.557	x1762	-0.775	x1299	+1.707	x289
+1.095	x3843	-0.592	x20101	-1.398	x250	+1.428	x1762	+0.685	x1307	-0.861	x247	-0.813	x1762	-0.567	x3020	-0.858	x2174	+1.630	x4046
+0.982	x1029	-0.647	x18146	-1.614	x4854	+1.393	x209	+0.676	x4594	-0.914	x381	-1.010	x4426	-0.592	x649	-0.918	x4426	+1.596	x208
+0.905	x4426	-0.648	x28661	-1.635	x4875	+1.323	x2003	*... 8117 more positive ...*		-0.971	x4977	-1.065	x2319	-0.650	x77	-0.934	x4504	+1.589	x3259
+0.867	x2708	-0.657	x23509	-1.697	x2150	*... 12649 more positive ...*		*... 54531 more negative ...*		-1.096	x167	-1.101	x2325	-0.672	x2218	-0.960	x224	+1.534	x1994
... 10073 more positive ...		-0.734	x19914	-1.828	x187	*... 49999 more negative ...*		-0.749	x4425	-1.144	x2150	-1.194	x224	-0.789	x2325	-1.160	x2319	+1.437	x2041
... 52575 more negative ...		-0.868	x4982	-1.865	x4453	-1.326	x2187	-0.757	x4453	-1.286	x4424	-1.264	x4504	-1.036	x224	-1.204	x2325	*... 21139 more positive ...*	
-1.084	x4425	-0.953	x26710	-2.559	x4424	-1.329	x3047	-0.986	x2873	-1.424	x4854	-1.276	x4427	-1.257	x2174	-1.384	x77	*... 41509 more negative ...*	
-1.120	x4982	-1.148	x60504	-4.263	x4425	-1.396	x224	-1.525	x2174	-2.370	x4425	-1.358	x2174	-1.363	x4427	-1.581	x4427	-1.769	x4972