**Description of Features**

Features are each tokens in the present in the string of text attribute. Some list of features are below

Ranked attributes:

0.212776 18039 Raymond

0.202032 24910 Perdita

0.194909 24439 Adrian

0.19037 24744 Idris

0.185894 17323 Gilman

0.183151 16721 Arkham

0.183151 17482 Innsmouth

0.178616 25119 Windsor

0.175861 2201 Street

0.172392 2507 West

0.171629 18469 Whateley

0.170976 18759 aout

0.170309 20486 folk

0.168934 16726 Armitage

0.166754 18968 bearded

….

….

The probability of a word given the class

-----------------------------------------

EAP HPL MWS

Raymond 5.022601707684581E-6 1.819979737558922E-5 0.0014502533503812092

Perdita 5.022601707684581E-6 6.066599125196411E-6 8.997490173793628E-4

Adrian 5.022601707684581E-6 6.066599125196411E-6 7.340057773357955E-4

Idris 5.022601707684581E-6 6.066599125196411E-6 6.392953544537574E-4

Gilman 5.022601707684581E-6 3.9432894313776656E-4 5.919401430127382E-6

Arkham 5.022601707684581E-6 3.5792934838658796E-4 5.919401430127382E-6

Innsmouth 5.022601707684581E-6 3.5792934838658796E-4 5.919401430127382E-6

Windsor 5.022601707684581E-6 6.066599125196411E-6 4.321163043992992E-4

Street 2.0090406830738327E-5 5.884601151440513E-4 5.919401430127382E-6

West 2.0090406830738327E-5 5.459939212676764E-4 5.919401430127382E-6

Whateley 5.022601707684581E-6 2.3053076675746357E-4 5.919401430127382E-6

aout 5.022601707684581E-6 2.2446416763226707E-4 5.919401430127382E-6

folk 5.022601707684581E-6 2.1839756850707055E-4 5.919401430127382E-6

Armitage 5.022601707684581E-6 2.062643702566778E-4 5.919401430127382E-6

bearded 5.022601707684581E-6 1.8806457288108856E-4 5.919401430127382E-6

…

…

**Feature Selection**

First each string is broken into vector of tokens using Weka function StringToWordVector. Then using each feature as an attribute Gain Ratio attribute evaluator is used for weighting for each feature. Then Ranker is used to generate ranking order of each feature with respect to their weights.

84.88% is accuracy achieved after apply above techniques. Also Using Information Gain attribute evaluator same accuracy was achieved.

Moreover filtering out features using ranked results was resulting in lesser accuracy that’s why it is omitted.

Correctly Classified Instances 4986 84.8825 %

Incorrectly Classified Instances 888 15.1175 %

**Classifier**

Multinomial Naïve Bayes classifier

**Confusion Matrix**

=== Confusion Matrix ===

a b c <-- classified as

1959 123 228 | a = EAP

180 1420 128 | b = HPL

168 61 1607 | c = MWS