Assignment 2 – Report

The given data was first formatted to suit each model separately and then separated as 75% and 25% data for both training and development purpose.

The IMDB data was divided as 9375 POSITIVE & NEGATIVE reviews as 75% data and 3125 POSITIVE & NEGATIVE reviews ad 25% data.

E-Mail data divided with 6978 HAM documents and 9675 SPAM documents as 75% data; 4055 HAM documents 1496 SPAM documents as 25% data.

Naive Bayes

Sentiment Analysis

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* POSITIVE precision: 0.844007858546169
* POSITIVE recall: 0.68736
* POSITIVE F1 score: 0.7576719576719576
* NEGATIVE precision: 0.7363022941970311
* NEGATIVE recall: 0.87296
* NEGATIVE F1 score: 0.7988286969253295

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* POSITIVE precision: 0.8390748031496063
* POSITIVE recall: 0.7274666666666667
* POSITIVE F1 score: 0.7792949780037708
* NEGATIVE precision: 0.7594614950103559
* NEGATIVE recall: 0.86048
* NEGATIVE F1 score: 0.806821023153473

Below are the results obtained by using 100% of the data for training and the same for development

* POSITIVE precision: 0.9315494173500216
* POSITIVE recall: 0.86336
* POSITIVE F1 score: 0.8961594353331949
* NEGATIVE precision: 0.8726798360044726
* NEGATIVE recall: 0.93656
* NEGATIVE F1 score: 0.9034921859926683

SPAM Analysis

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* SPAM precision: 0.8878224355128974
* SPAM recall: 0.9893048128342246
* SPAM F1 score: 0.935820423648435
* HAM precision: 0.995880535530381
* HAM recall: 0.9538840937114673
* HAM F1 score: 0.9744300289709031

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* SPAM precision: 0.9471947194719472
* SPAM recall: 0.9789147286821706
* SPAM F1 score: 0.9627935346142117
* HAM precision: 0.9693417493237151
* HAM recall: 0.9243336199484092
* HAM F1 score: 0.9463028169014084

Below are the results obtained by using 100% of the data for training and the same for development

* SPAM precision: 0.9858490566037735
* SPAM recall: 0.9915853549368902
* SPAM F1 score: 0.988708885616102
* HAM precision: 0.9914296134208607
* HAM recall: 0.9855886884800145
* HAM F1 score: 0.9885005227035135

SVM

IMDB Data

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* POSITIVE precision: 0.8590561224489796
* POSITIVE recall: 0.86208
* NEGATIVE precision: 0.8615928066795119
* NEGATIVE recall: 0.85856
* POSITIVE F1 score: 0.8605654048873982
* NEGATIVE F1 score: 0.8600737297643853

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* POSITIVE precision: 0.824715345241826
* POSITIVE recall: 0.8421333333333333
* NEGATIVE precision: 0.8387272529148959
* NEGATIVE recall: 0.8210133333333334
* POSITIVE F1 score: 0.8333333333333334
* NEGATIVE F1 score: 0.8297757654161276

Below are the results obtained by using 100% of the data for training and the same for development

* POSITIVE precision: 0.8885850457066958
* POSITIVE recall: 0.90984
* NEGATIVE precision: 0.9076305220883534
* NEGATIVE recall: 0.88592
* POSITIVE F1 score: 0.8990869204316376
* NEGATIVE F1 score: 0.8966438605724465

SPAM Analysis

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* POSITIVE precision: 0.7186293436293436
* POSITIVE recall: 0.9953208556149733
* NEGATIVE precision: 0.9979879275653923
* NEGATIVE recall: 0.8562268803945746
* POSITIVE F1 score: 0.8346412556053812
* NEGATIVE F1 score: 0.9216883461640563

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* POSITIVE precision: 0.9547860275215485
* POSITIVE recall: 0.652609819121447
* NEGATIVE precision: 0.6652390438247012
* NEGATIVE recall: 0.9571510461450272
* POSITIVE F1 score: 0.775294695481336
* NEGATIVE F1 score: 0.7849335997179457

Below are the results obtained by using 100% of the data for training and the same for development

* POSITIVE precision: 0.9385705847607797
* POSITIVE recall: 0.9957031599677737
* NEGATIVE precision: 0.9953636627064619
* NEGATIVE recall: 0.934016133417928
* POSITIVE F1 score: 0.9662931109373643
* NEGATIVE F1 score: 0.9637145796315347

MEGA M

IMDB Data

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* POSITIVE precision: 0.8074282147315855
* POSITIVE recall: 0.82784
* NEGATIVE precision: 0.8233749179251477
* NEGATIVE recall: 0.80256
* POSITIVE F1 score: 0.8175067151208721
* NEGATIVE F1 score: 0.8128342245989304

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* POSITIVE precision: 0.8106508875739645
* POSITIVE recall: 0.8183466666666667
* NEGATIVE precision: 0.8166056429032953
* NEGATIVE recall: 0.8088533333333333
* POSITIVE F1 score: 0.8144805987578959
* NEGATIVE F1 score: 0.8127110015540432

Below are the results obtained by using 100% of the data for training and the same for development

* POSITIVE precision: 0.8385271317829457
* POSITIVE recall: 0.86536
* NEGATIVE precision: 0.860909090909091
* NEGATIVE recall: 0.83336
* POSITIVE F1 score: 0.8517322834645669
* NEGATIVE F1 score: 0.8469105691056911

SPAM Analysis

Below are the results obtained by using 75% of the data for training and remaining 25% for development

* POSITIVE precision: 0.9682054417609294
* POSITIVE recall: 0.982015503875969
* NEGATIVE precision: 0.974561403508772
* NEGATIVE recall: 0.9552880481513327
* POSITIVE F1 score: 0.9750615763546797
* NEGATIVE F1 score: 0.9648284845853234

Below are the results obtained by using 25% of the data for training and remaining 75% for development

* POSITIVE precision: 0.9566144334292719
* POSITIVE recall: 0.9275452196382429
* NEGATIVE precision: 0.9036028602860287
* NEGATIVE recall: 0.9416738320435655
* POSITIVE F1 score: 0.9418555835432411
* NEGATIVE F1 score: 0.9222456140350878

Below are the results obtained by using 100% of the data for training and the same for development

* POSITIVE precision: 0.9608204049434657
* POSITIVE recall: 0.9812908423596813
* NEGATIVE precision: 0.9806391848077813
* NEGATIVE recall: 0.9594851808211728
* POSITIVE F1 score: 0.970947741364039
* NEGATIVE F1 score: 0.9699468572475719

When 75% of data was used for testing and remaining 25% used for development Maximum entropy model performed better for both sentiment analysis and spam classification.

Comparing the result values of each technique it can be seen that the Maximum entropy model performs the best but SVM is consistent and precision in SVM is high. Main strength of SVM is in using kernel to create new dimensions where we can put maximum margin but here we are using linear SVM, hence its performance is restricted and Maximum entropy model outperforms SVM. Also Maximum entropy model is a special case of Bayesian Inference with a posterior. The posterior in Maximum Entropy model makes it perform better than Naïve Bayes.

SPAM analysis performs better even though the number of e-mail documents are less than the number of IMDB reviews because the number of unique tokens in each E-mail is extensive and this helps to obtain an accurate probability measure whether the message is SPAM/HAM given the message.

When 25% of the data was used for training the performance of SVM model dropped by an average of 10% for sentiment analysis because the data provided was not random enough to create a margin that could match the actual classification but the percentage drop for spam detection was only 3%, this can be attributed to the random distribution of spam and ham emails between 75% and 25% training. For Maximum Entropy model the drop in both sentiment analysis and spam detection was around 2%. Naïve Bayes performance improved in performance of sentiment analysis this might be due to the number of unique words present in each review, which is more in the 25% training set. But the system performance drops by 4% for spam detection.

In SVM and Maximum Entropy model the system performs better when the system is trained with 75% of the data i.e. more the training data available to the system better are the results. When 100% of the data is used and the same is used for development the Maximum Entropy model trumps SVM by a very small percentage in F1 score. Hence we can conclude that Maximum Entropy model is robust.

Since the POSITIVE and NEGATIVE reviews in IMDB data is split uniformly the sentiments analysis does not perform well with 75% and 25% training data, the random split of SPAM and HAM emails from E-Mail data makes it perform better and hence the F1 scores of SPAM/HAM is better in all the cases compared to F1 scores of POSITIVE/NEGATIVE.

##### Naïve Bayes scripts in repository

nblearn.py

nbclassify.py

###### Other scripts in repository

#### For Naïve Bayes implementation

1. inputconv.py – This script is used to label the IMDB data as POSITIVE and NEGATIVE
2. nbsplit.py - This scripts splits the IMDB data into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
3. input100.py – This script is used to convert all the IMDB data into testing data and known output sample
4. metrics.py – This script is used to calculate precision, recall and F1 score
5. emailData.py – This script used to create e-mail data set using enron.vocab and the spam/ham e-mails
6. splitsh.py – This scripts splits the e-mail into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
7. input100sh.py – This script is used to convert all the e-mail data into testing data and known output sample

#### For SVM

1. inputconv\_svm.py - This script is used to label the input labels as +1 for POSITIVE/SPAM and -1 for NEGATIVE/HAM
2. emailData\_svm.py – This reads all the emails and creates E-Mail data with +1 and -1 labels
3. metrics\_svm.py - this is used to calculate POS/NEG and SPAM/HAM precision, recall and F1 score
4. svm\_output.py – This script converts the output from svm\_classify to a form suitable for visualizing
5. splitshsvm.py – This scripts splits the e-mail into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
6. input\_svm.py - This scripts splits the IMDB data into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
7. emailData100out.py - This script is used to convert all the e-mail data into testing data and known output sample
8. input100\_svm.py - This script is used to convert all the IMDB data into testing data and known output sample

#### For MEGA M

1. inputconv\_mega.py - This script is used to label the input labels as 1 for POSITIVE/SPAM and 0 for NEGATIVE/HAM
2. emailData\_mega.py – This reads all the emails and creates E-Mail data with 1 and 0 labels
3. metrics\_megam.py - this is used to calculate POS/NEG and SPAM/HAM precision, recall and F1 score
4. megam\_output.py – This script converts the output from maximum entropy to a form suitable for visualizing
5. splitshmegam.py – This scripts splits the e-mail into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
6. input\_megam.py - This scripts splits the IMDB data into 75% training/testing data and 25% training/testing data, and also creates 75%/25% test sample data for comparison
7. inputconv\_megam100.py - This script is used to convert all the IMDB data into testing data and known output sample