# Section 1

See data sets generated and submitted on MyCourses.

# Section 2

## Part a)

The f1 score for the **random classifier** on the Yelp reviews dataset is 0.1975.
The f1 score for the **majority classifier** on the Yelp reviews dataset is 0.351.

## Part b)

See Jupyter notebook for this section.

## Part c)

For each algorithm, 50 models were created. The fine tuning was executed in the following matter: for each parameter, 50 random values were selected between a range predetermined.

### Naive Bayes

The values of **alpha** ranged between 0 and 1. Here is the list of alpha values tried:

[ 0.97979595919183837, 0.48429685937187439, 0.97159431886377279, 0.94258851770354068,
0.54770954190838173, 0.74954990998199644, 0.1940388077615523, 0.55411082216443286,
0.48429685937187439, 0.44628925785157031, 0.66453290658131625, 0.12642528505701139,
0.9613922784556912, 0.55791158231646332, 0.088817763552710538, 0.04160832166433287,
0.0040008001600320064, 0.21784356871374275, 0.56911382276455291, 0.21664332866573316,
0.60612122424484893, 0.61252250450090018, 0.50370074014802957, 0.70414082816563317,
0.90598119623924789, 0.55091018203640729, 0.93318663732746554, 0.48009601920384076,
0.53670734146829369, 0.96419283856771354, 0.92078415683136627, 0.70514102820564117,
0.77475495099019809, 0.2488497699539908, 0.084416883376675342, 0.78835767153430691,
0.011802360472094419, 0.45129025805161033, 0.55811162232446487, 0.30486097219443886,
0.81756351270254046, 0.33886777355471093, 0.9155831166233247, 0.76555311062212439,
0.78115623124624922, 0.74534906981396276, 0.24084816963392677, 0.80576115223044609,
0.86577315463092619, 0.11522304460892178 ]

The best f1 score obtained was **0.421** with alpha set to **0.0118023604721**.

**Decision Trees**

The values of **max depth** ranged between 1 and 32. Here is the list of max depth values tried:

[ 4.8385677135427088, 30.418683736747347, 19.082816563312662, 7.0276055211042205,
5.0866173234646928, 27.535107021404279, 9.5143028605721138, 1.5209041808361672,
18.772754550910182, 22.592718543708742, 22.555511102220443, 11.895579115823164,
5.0246049209841965, 22.753950790158029, 19.058011602320462, 31.981396279255851,
8.385677135427084, 11.095619123824765, 15.380676135227045, 11.120424084816962,
2.4200840168033606, 18.264252850570113, 11.49249849969994, 4.1378275655131027,
31.125625125025003, 29.017203440688135, 1.3286657331466294, 18.952590518103619,
23.956991398279655, 11.418083616723344, 28.366073214642928, 11.628925785157032,
28.18623724744949, 23.72134426885377, 4.6463292658531703, 28.793958791758349,
18.053410682136427, 8.1314262852570511, 11.883176635327064, 16.019403880776153,
17.972794558911783, 12.844368873774755, 16.552710542108422, 24.831366273254652,
7.5237047409481894, 24.756951390278054, 29.972194438887776, 8.5035007001400267,
28.235847169433885, 1.2046409281856372 ]

The values of **min samples split** ranged between 0.1 and 1. Here is the list of min samples split values tried:

[ 0.83112622524504909, 0.29353870774154833, 0.5500900180036008, 0.96165233046609322,
0.12394478895779157, 0.2571714342868574, 0.75155031006201245, 0.27229445889177839,
0.91574314862972594, 0.68133626725345076, 0.34646929385877179, 0.22530506101220246,
0.85219043808761752, 0.84192838567713546, 0.12790558111622324, 0.32378475695139031,
0.54072814562912586, 0.33638727745549113, 0.45953190638127628, 0.71158231646329273,
0.77981596319263857, 0.20748149629925988, 0.40156031206241249, 0.98613722744548915,
0.18623724744948991, 0.66891378275655133, 0.70852170434086814, 0.1455491098219644,
0.36213242648529709, 0.59527905581116225, 0.1111622324464893, 0.60464092818563708,
0.55135027005401083, 0.55549109821964393, 0.54702940588117632, 0.75137027405481094,
0.70744148829765952, 0.13114622924584918, 0.79421884376875374, 0.36123224644928986,
0.67971594318863771, 0.86587317463492697, 0.95787157431486303, 0.4762752550510102,
0.10450090018003601, 0.62138427685537112, 0.12448489697939588, 0.75965193038607726,
0.23592718543708743, 0.80880176035207041 ]

The values of **min samples leaf** ranged between 0.1 and 1. Here is the list of min samples leaf values tried:

[ 0.41430286057211452, 0.21906381276255255, 0.43926785357071418, 0.22738547709541912,
0.12728545709141831, 0.26755351070214045, 0.46647329465893184, 0.14656931386277255,
0.30164032806561314, 0.15545109021804362, 0.28787757551510307, 0.19569913982796561,
0.11848369673934787, 0.26123224644928988, 0.19097819563912782, 0.24202840568113626,
0.45239047809561916, 0.22082416483296663, 0.18953790758151634, 0.34044808961792361,
0.31236247249449894, 0.46039207841568319, 0.2080216043208642, 0.11248249649929987,
0.34116823364672938, 0.48055611122224451, 0.18953790758151634, 0.1938587717543509,
0.19497899579915984, 0.14112822564512903, 0.23898779755951194, 0.20322064412882579,
0.39005801160232056, 0.12624524904980997, 0.39669933986797368, 0.18881776355271057,
0.44374874974995004, 0.1049609921984397, 0.1073614722944589, 0.29547909581916387,

0.10936187237447489, 0.35269053810762152, 0.39701940388077617, 0.12168433686737348, 0.43806761352270462, 0.41086217243448697, 0.16665333066613325, 0.19529905981196241, 0.19489897979595922, 0.16193238647729546 ]

The best f1 score obtained was **0.421** with:
Max depth set to **1.20464092819**.
Min samples split set to **0.808801760352**.
Min samples leaf is set to set to **0.161932386477**.

**Linear SVM**

The values of **C** ranged between 0.1 and 10. Here is the list of C values tried:

[6.8650330066013199, 9.6752150430086026, 2.175455091018204, 6.1382276455291054, 3.741948389677936, 8.8058211642328459, 9.5227245449089821, 1.2288257651530308, 7.623524704940988, 0.69609921984396883, 5.9322664532906577, 7.3640928185637122, 0.78125625125025, 9.6791758351670332, 8.6473894778955795, 5.4906381276255249, 5.5500500100019998, 2.5180636127225449, 1.163472694538908, 1.010982196439288, 3.4151830366073215, 2.4329065813162636, 5.1183236647329462, 9.5940188037607523, 5.1915983196639326, 3.4607321464292862, 5.3322064412882577, 6.5065813162632526, 9.5524304860972187, 0.24060812162432488, 6.2709141828365675, 0.56539307861572319, 5.9976195239047811, 3.741948389677936, 2.9359271854370874, 9.2830966193238655, 1.0882176435287059, 4.6846169233846764, 1.2486297259451891, 4.942068413682736, 0.97335467093418682, 5.1321864372874577, 9.9346469293858775, 7.8572114422884578, 4.385577115423084, 6.9957391478295659, 3.7696739347869577, 2.7992798559711942, 3.0250450090018006, 8.5760952190438093]

The best f1 score is: 0.465
The c is set to: 0.240608121624

## Part d)

With **Naive Bayes**
The f1 score on training data: 0.747.
The f1 score on validation data: 0.421.
The f1 score on testing data: 0.432.

With **Decision Tree**
The f1 score on training data: 0.396142857143.
The f1 score on validation data: 0.385.
The f1 score on testing data: 0.385.

With **SVM**.
The f1 score on training data: 0.985142857143.
The f1 score on validation data: 0.465.
The f1 score on testing data: 0.463.

### Part e)

The linear SVM performed the best on this task. Its accuracy on the testing dataset is 46.3% which is 3.1% more than the naive bayes classifier at 43.2%. We can also see that the f1 score on the validation and the testing dataset is similar for all classifiers on this task. One more observation that we can make is that all these classifiers are much more accurate than the random classifier and the majority-class classifier.
The decision tree probably performed badly because of the low variance in the data. This would explain why the f1 score is low on all the datasets. The Naive Bayes did not performed well because of the high number of feature. The SVM classifier, without surprises, performed incredibly well on the training data, but had the largest f1 score on the testing data. This is explained by the variance in the datasets and the number of features involved.
The role of C was regularize by how much is the classifier let misclassified elements to happens on the training set. As we can see, the f1 score on the training data is near 98% compared to 46.5% on the validation set, which means that there is still place for optimization.

# Section 3

## Part a)

See Jupyter notebook for this section.

## Part b)

### Naive Bayes

The values of **var smoothing** ranged between 1e-10 and 1e-1. Here is the list of var smoothing values tried:

[ 0.043808761808541714, 0.045509101874854976, 0.029245849240588123, 0.015903180720224046, 0.040248049669673942, 0.046749349923224652, 0.067213442721324262, 0.052170434134646934, 0.02678535714462893, 0.089617923595099028, 0.020384076894978998, 0.067573514735367074, 0.055291058256351278, 0.063832766589477896, 0.014502900665613125, 0.017843568795899184, 0.061872374513022611, 0.068173634758771751, 0.02306461299951991, 0.091638327673894779, 0.036967393541728351, 0.079395879196439298, 0.0084416884292258451, 0.035187037472294465, 0.037187437550310065, 0.031466293327185445, 0.099699939988297667, 0.029505901250730149, 0.057971594360892185, 0.011422284545469095, 0.043568713799179842, 0.067113422717423488, 0.060652130465433093, 0.038847769615063019, 0.032866573381796362, 0.037867573576835373, 0.035167033471514307, 0.013562712628945791, 0.033266653397399489, 0.084236847385237054, 0.0781156231465093, 0.078655731167573517, 0.00068013612652530523, 0.066193238681536315, 0.027785557183636733, 0.0084416884292258451, 0.029265853241368277, 0.089717943598999803, 0.080436087237007403, 0.022244448967533512 ]

The best f1 score obtained was **0.267** with var smoothing set to **0.000680136126525**.

### Decision Trees

The values of **max depth** ranged between 1 and 32. Here is the list of max depth values tried:

[13.774554910982197, 3.0712142428485696, 11.356071214242847, 6.1036207241448288, 2.1286257251450289, 10.605721144228845, 26.24524904980996, 18.723144628925784, 28.96759351870374, 9.3716743348669738, 12.428885777155431, 24.242248449689939, 11.263052610522104, 1.0372074414882977, 2.0046009201840365, 1.2046409281856372, 21.352470494098821, 11.796359271854371, 3.4122824564912984, 15.324864972994598, 8.0012002400480107, 1.3782756551310262, 6.835367073414683, 30.827965593118623, 9.2352470494098817, 13.452090418083616, 16.044208841768352, 5.5641128225645131, 27.479295859171835, 1.8495699139827966, 1.700740148029606, 3.5425085017003402, 15.269053810762152, 12.360672134426885, 4.4850970194038808, 7.3438687737547506, 7.2384476895379075, 29.686937387477496, 17.482896579315863, 24.68873774754951, 12.478495699139827, 28.911782356471292, 19.585117023404681, 22.865573114622922, 21.451690338067614, 15.690738147629526, 18.462692538507699, 29.637327465493097, 24.515103020604119, 27.039007801560309]

The values of **min sample split** ranged between 1 and 32. Here is the list of min sample split values tried:

[0.92438487697539506, 0.31478295659131827, 0.52776555311062212, 0.97137427485497108,
0.25699139827965595, 0.51408281656331267, 0.48185637127425485, 0.32648529705941187,
0.2845369073814763, 0.7225645129025805, 0.72652530506101221, 0.5815963192638528,
0.16985397079415884, 0.11224244848969794, 0.70528105621124226, 0.60356071214242846,
0.87271454290858175, 0.73246649329865976, 0.92492498499699938, 0.97479495899179835,
0.74830966193238646, 0.54486897379475896, 0.36231246249249849, 0.19793958791758354,
0.17615523104620925, 0.91304260852170438, 0.57997599519903986, 0.52398479695939193,
0.58087617523504709, 0.54684936987397481, 0.92348469693938795, 0.79169833966793357,
0.8044808961792359, 0.18407681536307263, 0.50238047609521908, 0.14878975795159033,
0.79745949189837972, 0.83454690938187637, 0.3196439287857572, 0.32270454090818168,
0.41704340868173639, 0.7382276455291058, 0.1187237447489498, 0.99297859571914382,
0.23466693338667735, 0.80880176035207041, 0.41038207641528313, 0.23538707741548312,
0.33224644928985803, 0.30632126425285056]

The values of **min sample leaf** ranged between 1 and 32. Here is the list of min sample leaf values tried:

[0.34476895379075823, 0.24722944588917786, 0.34492898579715947, 0.28883776755351076,
0.41486297259451899, 0.34196839367873577, 0.44958991798359682, 0.35317063412682537,
0.45151030206041209, 0.46575315063012612, 0.23066613322664536, 0.12176435287057412,
0.33748749749949991, 0.41046209241848375, 0.37893578715743148, 0.28907781556311263,
0.42486497299459902, 0.30700140028005607, 0.40494098819763957, 0.41174234846969404,
0.28227645529105827, 0.49295859171834377, 0.37133426685337068, 0.23538707741548312,
0.49895979195839169, 0.3687737547509502, 0.17169433886777358, 0.35629125825165042,
0.29723944788957796, 0.26899379875975199, 0.14168833766753353, 0.38485697139427888,
0.44798959791958393, 0.26843368673734747, 0.24930986197239449, 0.14592918583716744,
0.1275255051010202, 0.2593118623724745, 0.29747949589917988, 0.14192838567713545,
0.14264852970594119, 0.16105221044208842, 0.32140428085617123, 0.3440488097619524,
0.17257451490298059, 0.47959591918383682, 0.18545709141828368, 0.1879375875175035,
0.32100420084016806, 0.41318263652730547]

The best f1 score is: 0.39
The max depth is set to: 1.37827565513
The min samples split is set to: 0.544868973795
The min samples leaf is set to: 0.492958591718

**Linear SVM**

The values of **C** ranged between 1 and 32. Here is the list of C values tried:

[1.1555511102220446, 3.080496099219844, 5.7302660532106415, 3.3023004600920185, 6.4293458691738348,
0.88819763952790554, 8.0235647129425889, 3.8409681936387279, 1.9516703340668136, 6.0015803160632126,
9.7504700940188034, 5.4411282256451292, 4.3063612722544509, 7.936427285457091, 1.7160032006401282,
7.9661332266453293, 0.91988397679535905, 2.9775155031006202, 6.8214642928585718, 4.3835967193438687,
0.12178435687137429, 8.2433886777355472, 7.2353670734146824, 0.79511902380476096, 9.3227045409081821,
0.50202040408081616, 1.8110622124424887, 5.5599519903980799, 3.6884776955391079, 4.3954790958191641,
8.4137027405481088, 2.9497899579915985, 9.592038407681537, 5.3678535707141428, 2.0982196439287857,

0.94562912582516501, 5.2767553510702134, 5.7758151630326067, 3.6746149229845972, 4.7578915783156628, 4.9757351470294058, 6.2431886377275454, 8.6850170034006808, 0.18317663532706541, 3.120104020804161, 3.2904180836167236, 4.6628325665133028, 2.8586917383476695, 9.4633126625325072, 6.2392278455691139]

The best f1 score is: 0.501
The c is set to: 7.23536707341

### Part c)

With **Naive Bayes**
The f1 score on training data: 0.487571428571
The f1 score on validation data: 0.267
The f1 score on testing data: 0.2285

With **Decision Tree**
The f1 score on training data: 0.389571428571
The f1 score on validation data: 0.39
The f1 score on testing data: 0.3785

With **SVM**.
The f1 score on training data: 0.636857142857
The f1 score on validation data: 0.501
The f1 score on testing data: 0.502

### Part d)

The naive Bayes classifier did terribly in this task. Its results are close to what the random classifier achieve and the naive Bayes approach is worse than the majority-class classifier. The large number of features is probably the cause of this small f1 score. The decision tree did not perform well either and its f1 score is close to the majority-class classifier. As for the SVM, we can tell that the model has learned.
The main reasons why the SVM is the only model that performed well is because there are lots of features (10́000) and because there is lots of variability in the datasets.
The best f1 score was achieve on the SVM with a C value of 7.23536707341 which is a high value but allowed the model to generalize well and not be over trained. We can tell that the model is not too much over trained because the f1 score on the training data is close to the testing (and validation) f1 score.

### Part e)

The overall f1 score of the BBoW and the FBoW are comparable except for the Naive Bayes classifiers. We can observe a difference of near 21% between the 2 vector representations. This is due to the fact that Gaussian naive Bayes is expecting a normal distribution of the words. On the other hand, the decision tree classifier performed approximately equally in both case. Finally, the SVM performed best with the vector representation FBoW. This is probably due to the fact that there is more information contained in the FBoW representation, and the SVM model is the only algorithm capable of taking all this information in consideration.

### Part f)

The best representation is FBoW because it contains more information than the BBoW. With large models or other algorithms, this information can make a large difference. In addition, all the information that BBoW has is also contained in the FBoW.

# Section 4

## Part a)

The f1 score for the **random classifier** on the IMBD reviews dataset is 0.50344.
The f1 score for the **majority classifier** on the IMBD reviews dataset is 0.5.

## Part b)

See Jupyter notebook for this section.

## Part c)

### Naive Bayes

The values of **alpha** ranged between 0 and 1. Here is the list of alpha values tried:

[0.2464492898579716, 0.50890178035607125, 0.52270454090818164, 0.77135427085417085,
0.74814962992598522, 0.20784156831366274, 0.8097619523904781, 0.57411482296459293,
0.98559711942388473, 0.54310862172434482, 0.53690738147629524, 0.85277055411082214,
0.91158231646329269, 0.083816763352670534, 0.19103820764152832, 0.80016003200640129,
0.81876375275055013, 0.80276055211042208, 0.63152630526105225, 0.93878775755151034,
0.23724744948989798, 0.93818763752750556, 0.43088617723544709, 0.24024804960992199,
0.31426285257051412, 0.375875175035007, 0.75615123024604924, 0.9541908381676335,
0.39647929585917185, 0.75135027005401078, 0.31086217243448688, 0.40108021604320865,
0.16483296659331867, 0.21684336867373474, 0.91118223644728946, 0.21364272854570915,
0.74494898979795965, 0.30786157231446287, 0.56591318263652735, 0.99739947989597921,
0.45689137827565512, 0.23704740948189637, 0.34946989397879574, 0.23764752950590118,
0.75315063012602523, 0.11762352470494099, 0.34606921384276856, 0.76035207041408281,
0.82336467293458693, 0.12722544508901781]

The best f1 score obtained was **0.8432** with alpha set to **0.216843368674**.

### Decision Trees

The values of **max depth** ranged between 1 and 32. Here is the list of max depth values tried:

[18.295259051810362, 1.0744148829765954, 23.640728145629126, 4.6835367073414682,
26.592518503700738, 4.8881776355271054, 8.1314262852570511, 30.982996599319861,
2.2712542508501699, 15.740348069613923, 24.942988597719545, 19.628525705141026,
5.0680136027205442, 8.5531106221244251, 24.583316663332667, 25.45769153830766,
23.119823964792957, 5.5207041408281654, 2.3332666533306661, 26.666933386677336,
11.858371674334867, 14.748149629925985, 6.4508901780356069, 24.366273254650931,
19.678135627125425, 22.195839167833565, 23.330666133226643, 9.6259251850370067,
6.5935187037407479, 21.457891578315664, 28.235847169433885, 9.7685537107421485,
5.8307661532306456, 15.417883576715342, 26.580116023204639, 7.3500700140028004,
30.803160632126424, 22.853170634126826, 1.6015203040608121, 22.301260252050408,
2.0790158031606323, 11.616523304660932, 11.740548109621924, 1.6759351870374075,
17.613122624524905, 21.340068013602721, 25.091818363672733, 22.772554510902179,
6.3950790158031605, 27.820364072814563]

The values of **min samples split** ranged between 0.1 and 1. Here is the list of min samples split values tried:

[0.72292458491698341, 0.45845169033806765, 0.39903980796159233, 0.62786557311462299,
0.75749149829966, 0.2886777355471094, 0.54486897379475896, 0.37761552310462099,
0.8080816163232647, 0.85129025805161029, 0.29101820364072817, 0.84444888977795562,
0.85867173434686939, 0.20388077615523106, 0.43360672134426892, 0.24312862572514504,
0.10432086417283457, 0.80466093218643731, 0.42280456091218244, 0.93122624524904984,
0.78269653930786154, 0.2649129825965193, 0.94004800960192036, 0.58735747149429884,
0.63668733746749351, 0.92474494898979798, 0.26005201040208042, 0.23196639327865576,
0.97191438287657539, 0.29227845569113825, 0.10972194438887778, 0.8320264052810562,
0.73732746549309869, 0.20406081216243249, 0.21936387277455494, 0.52524504900980196,
0.53010602120424088, 0.66459291858371672, 0.56863372674534907, 0.89665933186637325,
0.17831566313262653, 0.4917583516703341, 0.8280656131226245, 0.66621324264852977,
0.72616523304660929, 0.83580716143228651, 0.83904780956191238, 0.56701340268053613,
0.57475495099019802, 0.2649129825965193]

The values of **min samples leaf** ranged between 0.1 and 1. Here is the list of min samples leaf values tried:

[0.46279255851170242, 0.18913782756551312, 0.33756751350270059, 0.34980996199239855,
0.21242248449689941, 0.30044008801760358, 0.23970794158831768, 0.27155431086217247,
0.16417283456691339, 0.17833566713342669, 0.14272854570914184, 0.10424084816963393,
0.2978795759151831, 0.10920184036807362, 0.37133426685337068, 0.25155031006201245,
0.18417683536707344, 0.17241448289657935, 0.17097419483896781, 0.35965193038607723,
0.44782956591318268, 0.22298459691938388, 0.35181036207241456, 0.3117223444688938,
0.44526905381076221, 0.31036207241448294, 0.28267653530706144, 0.21890378075615124,
0.3282056411282257, 0.12152430486097221, 0.10480096019203841, 0.41142228445689144,
0.45191038207641532, 0.44534906981396283, 0.33180636127225449, 0.16713342668533709,
0.27171434286857377, 0.27747549509901981, 0.20210042008401682, 0.40182036407281463,
0.3529305861172235, 0.38605721144228855, 0.12592518503700742, 0.15681136227245451,
0.37997599519903991, 0.12056411282256452, 0.34972994598919788, 0.29131826365273061,
0.12880576115223047, 0.14472894578915785]

The best f1 score obtained was **0.421** with:
Max depth set to **1.20464092819**.
Min samples split set to **0.808801760352**.
Min samples leaf is set to set to **0.161932386477**.

### Linear SVM

The values of **C** ranged between 0.1 and 10. Here is the list of C values tried:

[7.5086617323464688, 8.4592518503700731, 9.9960392078415676, 5.5579715943188637,
1.0446489297859574, 6.3659731946389275, 1.5615323064612925, 5.12624524904981,
3.9855371074214845, 4.4251850370074015, 3.955831166233247, 5.6055011002200441,

9.0216843368673736, 4.141988397679536, 6.4313262652530501, 9.4890578115623132,
8.928605721144228, 6.3481496299259854, 5.8669133826765352, 6.9957391478295659,
8.6176835367073412, 6.057031406281256, 8.4691538307661531, 5.0707941588317658,
0.44062812562512499, 7.8651330266053208, 6.8689937987597522, 0.75749149829965989,
7.6730346069213846, 5.0252450490098015, 6.813542708541708, 6.6768953790758152,
8.4354870974194842, 6.718483696739348, 0.80700140028005607, 7.1284256851370271,
9.0791158231646332, 6.7580916183236646, 5.8015603120624126, 5.6550110022004398,
9.4870774154830961, 3.7043208641728347, 7.3383476695339063, 1.7496699339867976,
0.98127625525105022, 8.4117223444688936, 1.3852770554110823, 7.9304860972194442,
4.0627725545109019, 0.45845169033806765]

The best f1 score is: 0.465
The c is set to: 0.240608121624

## Part d)

With Naive Bayes
The f1 score on training data: 0.871933333333
The f1 score on validation data: 0.8432
The f1 score on testing data: 0.83616

With Decision Tree
The f1 score on training data: 0.6472
The f1 score on validation data: 0.6499
The f1 score on testing data: 0.64996

With SVM
The f1 score on training data: 1.0
The f1 score on validation data: 0.8507
The f1 score on testing data: 0.83792

## Part e)

The Naive Bayes classifier and the SVM performed similarly. They also both performed well compared to the random classifier and the majority-class classifier. The f1 score of the decision tree is still higher than 50% (which is the approximately the random classifier and the majority-class classifier) for all training, validation and testing data sets. Since all three f1 scores for this model are close to each other, we can assume the the model is not over trained. Here again, the SVM was not confused by the variance in the datasets. As for the Decision tree, we can say that it was probably confused by the low variance. Finally, this task gave good results on the Naive Bayes even thought there are many features. The best SVM model had a C value set to 0.240608121624 which is very low, meaning that the classifier lets misclassified element pass regularly.

# Section 5

## Part a)

See Jupyter Notebook for this section.

## Part b)

### Naive Bayes

The values of **var smoothing** ranged between 1e-10 and 1e-1. Here is the list of var smoothing values tried:

[0.092298459699639929, 0.098559711943828773, 0.020924184916043212, 0.088977795570134036, 0.04438887783116624, 0.010742148518943789, 0.058671734388197651, 0.071294258880476094, 0.098279655932906593, 0.011002200529085819, 0.04928985802230447, 0.0022004401858171639, 0.053470694185357076, 0.06469293862302461, 0.079695939208141636, 0.05103020609017804, 0.041308261711022212, 0.060452090457631537, 0.064732946624584925, 0.069833966823524704, 0.059891978435787169, 0.035167033471514307, 0.011442288546249251, 0.080736147248709741, 0.060812162471674341, 0.030106021274134833, 0.081076215261972395, 0.076055211066153233, 0.062072414520824175, 0.011382276543908783, 0.083276655347789566, 0.066093218677635526, 0.099199839968793765, 0.035327065477755555, 0.053070614169753956, 0.013802760638307663, 0.020344068893418686, 0.0067013403613522711, 0.028485697210942195, 0.08879775956311263, 0.009661932476815363, 0.060632126464652943, 0.069733946819623929, 0.052990598166633332, 0.040048009661872379, 0.076255251073954797, 0.069293858802460501, 0.080156031226085223, 0.06257251454032807, 0.087817563524884984]

The best f1 score obtained was **0.6794** with var smoothing set to **0.00220044018582**.

### Decision Trees

The values of **max depth** ranged between 1 and 32. Here is the list of max depth values tried:

[7.3128625725145024, 23.318263652730547, 18.506101220244048, 22.865573114622922, 17.19143828765753, 1.4030806161232245, 19.975795159031804, 10.326665333066613, 31.832566513302659, 2.8665733146629329, 22.400480096019201, 21.389677935587116, 1.2170434086817363, 19.485897179435888, 8.4414882976595322, 21.780356071214243, 20.447089417883575, 7.1702340468093615, 4.9873974794958986, 1.5519103820764153, 8.3546709341868368, 1.0372074414882977, 10.983996799359872, 10.642928585717144, 9.1050210042008395, 12.354470894178835, 2.0790158031606323, 16.11862372474495, 10.332866573314663, 10.605721144228845, 22.425285057011401, 29.25905181036207, 29.705541108221645, 4.7827565513102623, 7.5175035007001396, 20.558711742348468, 20.416083216643329, 18.586717343468692, 8.6461292258451685, 18.673534706941389, 16.323264652930586, 9.5205041008201636, 18.084416883376676, 3.5859171834366874, 31.937987597519502, 25.234446889377875, 9.3282656531306252, 6.1222244448889773, 14.785357071414282, 19.312262452490497]

The values of **min sample split** ranged between 1 and 32. Here is the list of min sample split values tried:

[0.9060212042408482, 0.89215843168633735, 0.40066013202640527, 0.20640128025605123,
0.10432086417283457, 0.54954990998199649, 0.72382476495299064, 0.97677535507101421,
0.14824964992998602, 0.82086417283456692, 0.62642528505701145, 0.18839767953590719,
0.6712542508501701, 0.70204040808161638, 0.46925385077015402, 0.57349469893978799,
0.64280856171234246, 0.96309261852370476, 0.88189637927585518, 0.14860972194438887,
0.94634926985397083, 0.70384076815363072, 0.29047809561912386, 0.80034006801360269,
0.72940588117623528, 0.54198839767953599, 0.69519903980796161, 0.88099619923984795,
0.87667533506701345, 0.40030006001200247, 0.59545909181836376, 0.36555311062212448,
0.60500100020004, 0.66909381876375273, 0.12574514902980596, 0.16067213442688538,
0.33206641328265651, 0.37779555911182239, 0.86425285057011403, 0.63956791358271659,
0.12088417683536708, 0.50220044008801767, 0.24996999399879977, 0.18641728345669134,
0.56341268253650734, 0.10576115223044609, 0.94022804560912188, 0.65325065013002603,
0.36411282256451294, 0.78809761952390478]

The values of **min sample leaf** ranged between 1 and 32. Here is the list of min sample leaf values tried:

[0.39165833166633335, 0.29923984796959396, 0.31004200840168039, 0.32460492098419691,
0.23930786157231448, 0.36237247449489907, 0.13760752150430086, 0.22426485297059415,
0.32116423284656936, 0.30660132026405285, 0.28987797559511908, 0.22490498099619927,
0.26379275855171036, 0.40166033206641338, 0.25907181436287263, 0.467993598719744,
0.48991798359671945, 0.47447489497899586, 0.27651530306061212, 0.26923384676935391,
0.26323264652930589, 0.47335467093418693, 0.4505501100220044, 0.38085617123424687,
0.34076815363072621, 0.46583316663332675, 0.10136027205441089, 0.46639327865573121,
0.14024804960992199, 0.18913782756551312, 0.42990598119623935, 0.31780356071214244,
0.13224644928985799, 0.49023804760952194, 0.13576715343068615, 0.32420484096819369,
0.11552310462092419, 0.3414082816563313, 0.10000000000000001, 0.40822164432886587,
0.10904180836167235, 0.26275255051010205, 0.43758751750350078, 0.27771554310862179,
0.39549909981996401, 0.26275255051010205, 0.21266253250650133, 0.41494298859771961,
0.4020604120824165, 0.4947189437887578]

The best f1 score is: 0.6551
The max depth is set to: 16.3232646529
The min samples split is set to: 0.120884176835
The min samples leaf is set to: 0.109041808362

**Linear SVM**

The values of **C** ranged between 1 and 32. Here is the list of C values tried:

[6.1976395279055811, 7.823544708941788, 7.2413082616523301, 1.959591918383677,
6.7085817163432688, 6.5838167633526705, 7.5601520304060807, 7.0333666733346671,
6.1916983396679335, 6.9838567713542705, 1.4446889377875576, 5.969893978795759,
9.843548709741949, 7.7661132226445284, 0.1594118823764753, 5.5896579315863173,
0.20100020004000801, 1.3575515103020606, 4.5855971194238849, 9.0177235447089412,
8.9484096819363881, 5.5183636727345471, 7.4650930186037208, 7.6552110422084416,
6.7303660732146424, 9.5979795959191847, 5.3935987197439488, 0.37527505501100222,

7.9839567913582714, 6.6155031006201241, 7.0610922184436884, 2.6903580716143232, 0.9416683336667333, 1.3674534906981397, 9.7405681136227251, 8.5840168033606723, 5.1302060412082415, 1.1139627925585118, 6.0114822964592918, 1.9378075615123027, 8.364192838567714, 2.4447889577915585, 6.3105221044208841, 2.967613522704541, 3.4547909581916385, 9.0236647329465889, 8.3582516503300663, 1.7437287457491499, 3.5934186837367474, 9.964352870574114]

The best f1 score is: 0.8576
The c is set to: 9.96435287057

### Part c)

With Naive Bayes
The f1 score on training data: 0.708266666667
The f1 score on validation data: 0.6794
The f1 score on testing data: 0.65128

With Decision Tree
The f1 score on training data: 0.6618
The f1 score on validation data: 0.6551
The f1 score on testing data: 0.6558

With SVM
The f1 score on training data: 0.8862
The f1 score on validation data: 0.8576
The f1 score on testing data: 0.85924

### Part d)

All the classifiers show that they learn since they are far from the random and majority-class classifier f1 scores. The best model is the SVM from far with an f1 score of almost 86%.
The SVM classifier probably performed better because of the large number of features involved and the low variability in the data.
The C value for the SVM classifier is equal to 9.96 which is pretty high. This allowed for more tailored outcomes.

### Part e)

The yelp dataset has lower f1 scores than the IMBD dataset for the BBoW representation. This is probably due to the fact that the yelp dataset has a wider range of classes compared to the imbd dataset, which has only 1 or 0 as possible classes.
The Naive Bayes classifier with FBoW has worse performance than with BBoW because the Gaussian Naive Bayes is expecting normal distribution of values which was not exactly the case you our datasets.
The decision tree uses threshold to determine classes. This is why we can observe very similar f1 score on both vector representation.

### Part f)

Here the FBoW is a better representation because it shows better f1 score. This is because the FBoW representation contains more information than the BBoW.

### Part g)

The IMBD dataset has better overall performance. This is because there are less classes and also because they are balanced, meaning its almost 50/50 in each classes.