Naive Bayes Classification

Index	Class	Recall	Precision	F-Measure
0	Positive	0.688644688645	0.926108374384	1.147058824
0	Negative	0.857142857143	0.835714285714	0.987341772
1	Positive	0.630036630037	0.891191709845	1.17167382
1	Negative	0.846153846154	0.783050847458	0.961267606
2	Positive	0.677655677656	0.876777251185	1.128099174
2	Negative	0.81684981685	0.796428571429	0.987341772
3	Positive	0.721611721612	0.895454545455	1.107505071
3	Negative	0.846153846154	0.799307958478	0.971530249
4	Positive	0.85347985348	0.978991596639	1.068493151
4	Negative	0.937728937729	0.924187725632	0.992727273
5	Positive	0.857142857143	0.962962962963	1.058139535
5	Negative	0.882783882784	0.92337164751	1.02247191
6	Positive	0.85347985348	0.962809917355	1.060194175
6	Negative	0.893772893773	0.931297709924	1.020560748
7	Positive	0.846153846154	0.946721311475	1.056092843
7	Negative	0.864468864469	0.921875	1.032136106
8	Positive	0.871794871795	0.9296875	1.032136106
8	Negative	0.864468864469	0.940239043825	1.041984733
9	Positive	0.860805860806	0.93625498008	1.041984733
9	Negative	0.813186813187	0.90612244898	1.054054054

Original System	Positive Average	Negative Average
Recall	0.786080586	0.862271062
Precision	0.930696015	0.876159524
F-Measure	1.087137743	1.007141622

Improved Bayes Classification (taking into account number of words in the file)

Index	**Class**	**Recall**	**Precision**	**F-Measure**
0	Positive	0.615384615	0.949152542	1.213333333
0	Negative	0.578754579	0.869731801	1.200883644
1	Positive	0.58974359	0.913294798	1.215264767
1	Negative	0.688644689	0.842911877	1.100725753
2	Positive	0.816849817	0.884615385	1.039827772
2	Negative	0.802197802	0.824701195	1.013832078
3	Positive	0.816849817	0.926108374	1.062685702
3	Negative	0.802197802	0.82527881	1.014182083
4	Positive	0.831501832	0.991111111	1.087571681
4	Negative	0.805860806	0.950570342	1.082388391
5	Positive	0.831501832	0.977678571	1.080797216
5	Negative	0.805860806	0.9437751	1.078824568
6	Positive	0.758241758	0.982378855	1.128768495
6	Negative	0.813186813	0.9437751	1.074326191
7	Positive	0.915750916	0.956331878	1.021676906
7	Negative	0.860805861	0.941666667	1.044861048

8	Positive	0.860805861	0.941908714	1.044989292
8	Negative	0.827838828	0.94893617	1.068155699
9	Positive	0.816849817	0.948275862	1.074457047
9	Negative	0.783882784	0.938596491	1.089820359

Improved System	Positive Average	Negative Average
Recall	0.785347985	0.776923077
Precision	0.947085609	0.902994355
F-Measure	1.096937221	1.076799981

Additional Notes/ Analysis:
In terms of precision, our improved system using the number of words in a file is better than our original system. In terms of recall, our improved system is worse than our original system. In terms of f-measure, our improved system is better than our original system. Looking at the f-measure, I think the improved system performed better than the original because the more words a document has, the more the conditional probabilities for positive and negative documents will be different from each other. In other words, if there are more words in a file, the likelihood of them being neutral words increase. Therefore, we increased the threshold for a "neutral" classification for longer documents. We can further improve performance in the future by including a dictionary of neutral words such as "the", "me", "and" to exclude in our calculation of conditional probabilities.