Summary of the standard models outputs are shown below. The standard model was applied on the test set and then the training set for comparative purposes.
Unless specified otherwise the training set will consist of the first 1500 tweets of dataset.tsv and the test set will consist of the last 500 tweets of dataset.tsv

Classification of Test Set

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.674	0.486	0.433	0.438	82
Decision Tree Topic	0.34	0.229	0.197	0.201	86
Multinomial Bayes Sentiment	0.738	0.733	0.489	0.515	177
Multinomial Bayes Topic	0.3	0.174	0.128	0.125	129
Bernoulli Bayes Sentiment	0.728	0.827	0.426	0.428	151
Bernoulli Bayes Topic	0.184	0.028	0.054	0.022	168

Classification of Training Set

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.705	0.625	0.509	0.537	124
Decision Tree Topic	0.409	0.231	0.223	0.22	183
Multinomial Bayes Sentiment	0.941	0.961	0.84	0.884	303
Multinomial Bayes Topic	0.739	0.791	0.495	0.555	284
Bernoulli Bayes Sentiment	0.864	0.929	0.605	0.612	407
Bernoulli Bayes Topic	0.279	0.162	0.091	0.068	438

1. (1 mark) Give simple descriptive statistics showing the frequency distributions for the sentiment and topic classes across the full data set. What do you notice about the distributions?

The frequency distribution of sentiments across the full 2000 tweet data set is

Sentiment	Count	Percentage (%)
negative	1294	64.7
netural	553	27.7
positive	153	7.6

It can be seen that the majority of tweets are negative, with some neutral tweets and very few positive tweets.

The frequency distribution of topics across the full 2000 tweet data set is

Topic	Count	(%)	Topic	Count	(%)
10000	244	12.2	10010	56	2.8
10001	140	7.0	10011	13	0.7
10002	130	6.5	10012	25	1.3
10003	358	17.9	10013	104	5.2
10004	17	0.9	10014	29	1.5
10005	194	9.7	10015	119	5.9
10006	189	9.5	10016	59	2.9
10007	7	0.4	10017	47	2.4
10008	163	8.1	10018	38	1.9
10009	16	0.8	10019	52	2.6

Like the sentiment distribution, there is an uneven frequency distribution of topics within the tweets, skewed towards negativity. There are some tweets that comprise a very small amount of

the total data set (<3%), and so it may be difficult to attain the relevant keywords to accurately sort them.

2. (2 marks) Vary the number of words from the vocabulary used as training features for the standard methods (e.g. the top N words for N = 100, 200, etc.). Show metrics calculated on both the training set and the test set. Explain any difference in performance of the models between training and test set, and comment on metrics and run times in relation to the number of features.

Classifiers on test set

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Bernouli Bayes Sentiment	0.718	0.581	0.51	0.531	134
N=200					
Bernouli Bayes Sentiment	0.708	0.565	0.502	0.519	136
N=300					
Bernouli Bayes Sentiment	0.736	0.685	0.568	0.599	137
N=400					
Bernouli Bayes Sentiment	0.75	0.667	0.573	0.598	140
N=500					
Bernouli Bayes Sentiment	0.742	0.703	0.552	0.58	142
N=600					
Bernouli Bayes Sentiment	0.738	0.693	0.543	0.568	147

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Bernouli Bayes Topic	0.312	0.198	0.17	0.175	142
N=200					
Bernouli Bayes Topic	0.34	0.196	0.184	0.181	138
N=300					
Bernouli Bayes Topic	0.372	0.239	0.206	0.209	141
N=400					
Bernouli Bayes Topic	0.378	0.209	0.193	0.191	145
N=500					
Bernouli Bayes Topic	0.38	0.215	0.195	0.193	145
N=600					
Bernouli Bayes Topic	0.382	0.223	0.196	0.193	154

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Multinomial Bayes Sentiment	0.714	0.585	0.485	0.508	163
N=200					
Multinomial Bayes Sentiment	0.724	0.637	0.509	0.534	136
N=300					
Multinomial Bayes Sentiment	0.742	0.694	0.585	0.62	136
N=400					
Multinomial Bayes Sentiment	0.752	0.705	0.595	0.631	140
N=500					
Multinomial Bayes Sentiment	0.736	0.683	0.573	0.607	138
N=600					
Multinomial Bayes Sentiment	0.736	0.69	0.569	0.603	138

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Multinomial Bayes Topic	0.288	0.185	0.162	0.166	134
N=200					
Multinomial Bayes Topic	0.338	0.2	0.189	0.188	135
N=300					
Multinomial Bayes Topic	0.366	0.229	0.217	0.217	138
N=400					
Multinomial Bayes Topic	0.39	0.24	0.227	0.226	139
N=500					
Multinomial Bayes Topic	0.404	0.247	0.235	0.234	142
N=600					
Multinomial Bayes Topic	0.396	0.221	0.221	0.216	140

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Decision Tree Sentiment	0.692	0.505	0.447	0.453	77
N=200					
Decision Tree Sentiment	0.674	0.486	0.433	0.438	82
N=300					
Decision Tree Sentiment	0.674	0.486	0.433	0.438	93
N=400					
Decision Tree Sentiment	0.674	0.486	0.433	0.438	95
N=500					
Decision Tree Sentiment	0.674	0.486	0.433	0.438	107
N=600					
Decision Tree Sentiment	0.674	0.486	0.433	0.438	132

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Decision Tree Topic	0.336	0.235	0.201	0.206	79
N=200					
Decision Tree Topic	0.34	0.229	0.197	0.201	92
N=300					
Decision Tree Topic	0.34	0.229	0.197	0.201	118
N=400					
Decision Tree Topic	0.34	0.229	0.197	0.201	115
N=500					
Decision Tree Topic	0.34	0.229	0.197	0.201	123
N=600					
Decision Tree Topic	0.34	0.229	0.197	0.201	209

Classifiers on training set

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Bernoulli Bayes Sentiment	0.726	0.649	0.592	0.613	150
N=200					
Bernoulli Bayes Sentiment	0.757	0.698	0.654	0.673	142
N=300					
Bernoulli Bayes Sentiment	0.765	0.719	0.655	0.681	147
N=400					
Bernoulli Bayes Sentiment	0.787	0.748	0.679	0.706	151
N=500					
Bernoulli Bayes Sentiment	0.817	0.802	0.713	0.747	157
N=600					
Bernoulli Bayes Sentiment	0.819	0.799	0.713	0.746	158

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Bernoulli Bayes Topic	0.437	0.4	0.291	0.319	149
N=200					
Bernoulli Bayes Topic	0.523	0.474	0.34	0.365	146
N=300					
Bernoulli Bayes Topic	0.561	0.487	0.345	0.363	151
N=400					
Bernoulli Bayes Topic	0.582	0.514	0.355	0.374	154
N=500					
Bernoulli Bayes Topic	0.591	0.465	0.345	0.358	162
N=600					
Bernoulli Bayes Topic	0.596	0.482	0.34	0.349	163

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Multinomial Bayes Sentiment	0.723	0.669	0.576	0.607	153
N=200					
Multinomial Bayes Sentiment	0.754	0.695	0.635	0.66	142
N=300					
Multinomial Bayes Sentiment	0.771	0.731	0.667	0.694	199
N=400					
Multinomial Bayes Sentiment	0.8	0.764	0.696	0.724	197
N=500					
Multinomial Bayes Sentiment	0.817	0.797	0.723	0.754	199
N=600					
Multinomial Bayes Sentiment	0.824	0.795	0.732	0.759	199

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Multinomial Bayes Topic	0.437	0.452	0.292	0.323	193
N=200					
Multinomial Bayes Topic	0.544	0.537	0.403	0.437	192
N=300					
Multinomial Bayes Topic	0.587	0.608	0.451	0.479	146
N=400					
Multinomial Bayes Topic	0.627	0.651	0.477	0.515	148
N=500					
Multinomial Bayes Topic	0.657	0.67	0.495	0.531	151
N=600					
Multinomial Bayes Topic	0.685	0.688	0.514	0.552	149

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Decision Tree Sentiment	0.705	0.626	0.508	0.536	109
N=200					
Decision Tree Sentiment	0.705	0.625	0.509	0.537	121
N=300					
Decision Tree Sentiment	0.705	0.625	0.509	0.537	127
N=400					
Decision Tree Sentiment	0.705	0.625	0.509	0.537	126
N=500					
Decision Tree Sentiment	0.705	0.625	0.509	0.537	139
N=600					
Decision Tree Sentiment	0.705	0.625	0.509	0.537	164

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
N=100					
Decision Tree Topic	0.401	0.235	0.225	0.224	108
N=200					
Decision Tree Topic	0.409	0.231	0.223	0.22	141
N=300					
Decision Tree Topic	0.409	0.231	0.223	0.22	157
N=400					
Decision Tree Topic	0.409	0.231	0.223	0.22	147
N=500					
Decision Tree Topic	0.409	0.231	0.223	0.22	272
N=600					
Decision Tree Topic	0.409	0.231	0.223	0.22	179

Increasing the number of N (top number of most frequent words) increase accuracy and macro average parameters for multinomial bayes and binomial bayes models. The decision tree sentiment classifer on the other hand, was the only classifer to become more accurate with a lower number of N, however not by much.The naive bayes models stopped increasing in accuracy after 400 features for the test set, indicated potential local or global maxima that will be further examined. It can be seen that all the models yielded more accurate results when making predictions on the training set, which is to be expected since the models' statistical inferences were derived from the labelled training set. It is for this reason that there is notable improvements in accuracy when the number of features used were increased for the training set. Increasing the number of maximum features increased computation time. Increasing the number of elements to be analysed (1500 vs 500 tweets) also increased computation time. This increase was more notable when using the decision tree model.

3. (2 marks) Evaluate the standard models with respect to baseline predictors (VADER for sentiment analysis, majority class for both classifiers). Comment on the performance of the baselines and of the methods relative to the baselines.

Classifers on Test Set

Standard Models

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.674	0.486	0.433	0.438	82
Decision Tree Topic	0.34	0.229	0.197	0.201	86
Multinomial Bayes Sentiment	0.738	0.733	0.489	0.515	177
Multinomial Bayes Topic	0.3	0.174	0.128	0.125	129
Bernoulli Bayes Sentiment	0.728	0.827	0.426	0.428	151
Bernoulli Bayes Topic	0.184	0.028	0.054	0.022	168

Baselines

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Vader Sentiment	0.434	0.404	0.455	0.364	114
Majority Sentiment	0.67	0.223	0.333	0.267	n/a
Majority Topic	0.174	0.009	0.05	0.015	n/a

Classifiers on Training set

Standard Models

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.705	0.625	0.509	0.537	124
Decision Tree Topic	0.409	0.231	0.223	0.22	183
Multinomial Bayes Sentiment	0.941	0.961	0.84	0.884	303
Multinomial Bayes Topic	0.739	0.791	0.495	0.555	284
Bernoulli Bayes Sentiment	0.864	0.929	0.605	0.612	407
Bernoulli Bayes Topic	0.279	0.162	0.091	0.068	438

Baselines

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Vader Sentiment	0.443	0.423	0.472	0.382	388
Majority Sentiment	0.639	0.213	0.333	0.26	n/a
Majority Topic	0.181	0.009	0.05	0.015	n/a

For the test set, the decision tree was just as accurate at predicting sentiment as the majority sentiment classifer. The decision tree however was much better at predicting topics than the majority topic classifer (twice as accurate).The multinomial bayes classifer improved on accuracy in comparison to all baseline predictors for both sentiment and topics. The binomial bayes classifer has better accuracy compared to the baseline predictors, however its ability to classify topics was just as bad as the majority topic classifer. There is little difference between the all the standard method sentiment classifiers' accuracy and the accuracy of the majority sentiment classifer. This is because most of the tweets are negative, and there are a small number of classes to choose from. The majority classifer accuracy plummets when there are multiple classes as seen in its attemp to classify topics. The vader sentiment classifer was inferior to all standard model sentiment analysis. The standard models outperformed the baseline classifers in all aspects when classifying the test set. Not much inference can be obtained comparing the standard models and baselines on the training set.

4. (2 marks) Evaluate the effect that preprocessing the input features, in particular stop word removal plus Porter stemming as implemented in NLTK, has on classifier performance, for the three standard methods for both sentiment and topic classification. Compare results with and without pre-processing on training and test sets and comment on any similarities and differences.

Classifers on Test Set

Standard Models

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.674	0.486	0.433	0.438	82
Decision Tree Topic	0.34	0.229	0.197	0.201	86
Multinomial Bayes Sentiment	0.738	0.733	0.489	0.515	177
Multinomial Bayes Topic	0.3	0.174	0.128	0.125	129
Bernoulli Bayes Sentiment	0.728	0.827	0.426	0.428	151
Bernoulli Bayes Topic	0.184	0.028	0.054	0.022	168

Standard Models with stop word filtration and stemming

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.69	0.578	0.413	0.42	508
Decision Tree Topic	0.366	0.227	0.221	0.215	484
Multinomial Bayes Sentiment	0.744	0.787	0.521	0.548	732
Multinomial Bayes Topic	0.392	0.286	0.203	0.21	594
Bernoulli Bayes Sentiment	0.732	0.801	0.447	0.45	572
Bernoulli Bayes Topic	0.224	0.086	0.073	0.046	602

Standard Models with no pre-processing

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.67	0.489	0.444	0.447	57
Decision Tree Topic	0.336	0.214	0.186	0.188	95
Multinomial Bayes Sentiment	0.738	0.79	0.475	0.497	141
Multinomial Bayes Topic	0.28	0.155	0.115	0.11	105
Bernoulli Bayes Sentiment	0.718	0.841	0.408	0.403	133
Bernoulli Bayes Topic	0.178	0.02	0.052	0.018	132

Classifers on Training Set

Standard Models

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.705	0.625	0.509	0.537	124
Decision Tree Topic	0.409	0.231	0.223	0.22	183
Multinomial Bayes Sentiment	0.941	0.961	0.84	0.884	303
Multinomial Bayes Topic	0.739	0.791	0.495	0.555	284
Bernoulli Bayes Sentiment	0.864	0.929	0.605	0.612	407
Bernoulli Bayes Topic	0.279	0.162	0.091	0.068	438

Standard Models with stop word filtration and stemming

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.695	0.675	0.444	0.46	746
Decision Tree Topic	0.403	0.223	0.226	0.217	714
Multinomial Bayes Sentiment	0.924	0.936	0.828	0.869	827
Multinomial Bayes Topic	0.777	0.839	0.567	0.63	833
Bernoulli Bayes Sentiment	0.862	0.92	0.604	0.609	865
Bernoulli Bayes Topic	0.333	0.249	0.119	0.106	899

Standard Models with no pre-processing

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.707	0.625	0.518	0.545	73
Decision Tree Topic	0.406	0.22	0.216	0.211	77
Multinomial Bayes Sentiment	0.943	0.962	0.831	0.877	143
Multinomial Bayes Topic	0.724	0.794	0.456	0.508	190
Bernoulli Bayes Sentiment	0.869	0.933	0.61	0.617	206
Bernoulli Bayes Topic	0.265	0.132	0.085	0.059	198

For classifying the Test Set:
Implementing the pre processing specified for the standard models (removal of characters and http links) only improved accuracy of the multinomial bayes topics and bernoulli bayes sentiments by 1-2%. All other accuracy figures were similar. All classifiers had an improvement in their macro average 1-3%. Implying a greater ability to categorise the smaller classes more accurately. It is

clear that this increased performance induced greater computational costs as seen in increased calculation time.

Implementing stop words filtration and stemming produced more accurate predicions to the standard models(approx 2-4% more). There was even more improvement to the f1-macro average, implying even greater ability to classify the less frequent classes. This improvement came with drastically increased calculation time per classifer (approx 4-5 times as long)

For classifying the Training Set:
All round increases in performance for the each classifer with the exception of computation time.

5. (2 marks) Sentiment classification of neutral tweets is notoriously difficult. Repeat the experiments of items 2 (with N = 200), 3 and 4 for sentiment analysis with the standard models using only the positive and negative tweets (i.e. removing neutral tweets from both training and test sets). Compare these results to the previous results. Is there any difference in the metrics for either of the classes (i.e. consider positive and negative classes individually)?

Classifers on Test Set

Comparisons to Item 2

Bernoulli Sentiments
60 2

	precision	recall	f1-score	support
negative	0.93	0.97	0.95	335
positive	0.58	0.38	0.45	40

Bernoulli Sentiment
75 2

	precision	recall	f1-score	support
negative	0.78	0.86	0.82	335
neutral	0.53	0.47	0.50	125
positive	0.39	0.17	0.24	40

Decision Tree Sentiments
60 2

	precision	recall	f1-score	support
negative	0.90	0.96	0.93	335
positive	0.30	0.15	0.20	40

Decision Tree Sentiment
86 2

	precision	recall	f1-score	support
negative	0.75	0.86	0.80	335
neutral	0.44	0.36	0.39	125
positive	0.27	0.07	0.12	40

Molinomial Sentiments
79 2

	precision	recall	f1-score	support
negative	0.92	0.96	0.94	335
positive	0.52	0.33	0.40	40

Multinomial Sentiment
72 2

	precision	recall	f1-score	support
negative	0.78	0.89	0.83	335
neutral	0.55	0.46	0.50	125
positive	0.58	0.17	0.27	40

Comparisons to Item 3

Majority_sentiment

	precision	recall	f1-score	support
negative	0.89	1.00	0.94	335
positive	0.00	0.00	0.00	40

Vader_sentiment

	precision	recall	f1-score	support
negative	0.95	0.47	0.63	335
neutral	0.00	0.00	0.00	0
positive	0.19	0.62	0.29	40

Vader_sentiment

	precision	recall	f1-score	support
negative	0.78	0.47	0.59	335
neutral	0.30	0.26	0.28	125
positive	0.14	0.62	0.22	40

Comparisons to Item 4

Bernoulli Sentiments
411 ms

	precision	recall	f1-score	support
negative	0.90	1.00	0.94	335
positive	1.00	0.03	0.05	40

Bernoulli Sentiment
639 2

	precision	recall	f1-score	support
negative	0.74	0.96	0.84	335
neutral	0.66	0.36	0.47	125
positive	1.00	0.03	0.05	40

Decision Tree Sentiments
355 ms

	precision	recall	f1-score	support
negative	0.89	1.00	0.94	335
positive	0.00	0.00	0.00	40

Decision Tree Sentiment
536 2

	precision	recall	f1-score	support
negative	0.71	0.95	0.81	335
neutral	0.52	0.19	0.28	125
positive	0.50	0.10	0.17	40

Multinomial Sentiments
444 ms

	precision	recall	f1-score	support
negative	0.91	0.99	0.95	335
positive	0.69	0.23	0.34	40

Multinomial Sentiment
516 2

	precision	recall	f1-score	support
negative	0.79	0.90	0.84	335
neutral	0.57	0.51	0.54	125
positive	1.00	0.15	0.26	40

Classifers on Training Set
Comparisons to Item 2

Bernoulli Sentiments
117 ms

	precision	recall	f1-score	support
negative	0.95	0.96	0.96	959
positive	0.63	0.57	0.60	113

Bernoulli Sentiment
105 2

	precision	recall	f1-score	support
negative	0.81	0.86	0.83	959
neutral	0.65	0.60	0.63	428
positive	0.63	0.50	0.56	113

Decision Tree Sentiments
81 ms

	precision	recall	f1-score	support
negative	0.92	0.98	0.95	959
positive	0.64	0.27	0.38	113

Decision Tree Sentiment
112 2

	precision	recall	f1-score	support
negative	0.74	0.90	0.81	959
neutral	0.58	0.40	0.47	428
positive	0.55	0.23	0.33	113

Multinomial Sentiments
128 ms

	precision	recall	f1-score	support
negative	0.95	0.97	0.96	959
positive	0.66	0.57	0.61	113

Multinomial Sentiment
103 2

	precision	recall	f1-score	support
negative	0.80	0.87	0.83	959
neutral	0.66	0.58	0.62	428
positive	0.63	0.46	0.53	113

Comparisons to Item 3

Majority_sentiment

	precision	recall	f1-score	support
negative	0.89	1.00	0.94	959
positive	0.00	0.00	0.00	113

Vader_sentiment

	precision	recall	f1-score	support
negative	0.96	0.48	0.64	959
neutral	0.00	0.00	0.00	0
positive	0.18	0.63	0.28	113

Vader Sentiment

	precision	recall	f1-score	support
negative	0.73	0.48	0.58	959
neutral	0.40	0.30	0.35	428
positive	0.13	0.63	0.22	113

Comparisons to Item 4

Bernoulli Sentiments
729 ms

	precision	recall	f1-score	support
negative	0.90	1.00	0.95	959
positive	1.00	0.08	0.15	113

Bernoulli Bayes Sentiment
803 2

	precision	recall	f1-score	support
negative	0.84	0.99	0.91	959
neutral	0.92	0.78	0.84	428
positive	1.00	0.04	0.07	113

Decision Tree Sentiments
538 ms

	precision	recall	f1-score	support
negative	0.89	1.00	0.94	959
positive	0.00	0.00	0.00	113

Decision Tree Sentiment
776 2

	precision	recall	f1-score	support
negative	0.70	0.96	0.81	959
neutral	0.69	0.25	0.36	428
positive	0.64	0.12	0.21	113

Multinomial Sentiments
696 ms

	precision	recall	f1-score	support
negative	0.97	1.00	0.98	959
positive	0.96	0.75	0.84	113

Multinomial Bayes Sentiment
770 2

	precision	recall	f1-score	support
negative	0.93	0.98	0.95	959
neutral	0.91	0.89	0.90	428
positive	0.97	0.62	0.76	113

Removing neutral tweets has resulted in the classifiers almost always correctly classifying negative tweets (approx 90% precision and recall). Changes in the positive tweet statistics depended on the classifer and customisation to the classifer. There was a notable increase in positive precision and recall for the bernoulli and decision tree classifiers in item 2 tested with the test set. The multinomial sentiment classifer on other hand dropped in accuracy but increased its recall and overall f1 score.
There wasn't as much of an increase with the training set results with positive precision however there was still a noticeable increase in positive recall (>10%). For item 3, the vader classifier positive precision increased, however recall stayed exactly the same. With the majority sentiment classifer, precision simply increased as expected due.
For item 4, there was variation how positive classifation was affected. The bernoulli classifer stayed the same. The decision tree did not classify anything positively, a complete reduction in positive classification ability. The multinomial classifer has a decrease in precision but increase in recall.
In summary converting to binary outputs allowed som classifers to increase in recall but decrease in accuracy. Some classifers however are greatly influenced by the uneven distribution and fail to even attempt to classify the positive classes.

6. (6 marks) Describe your best method for sentiment analysis and your best method for topic classification. Give some experimental results showing how you arrived at your methods. Now provide a brief comparison of your methods in relation to the standard methods and the baselines.

Chosen sentiment classifer: Multinomial naive bayes with unrestricted number of features, stop word filtration, and word stemming
Chosen topic classifer: Multinomial naive bayes with max number of features restrictd to N = 700, stop word filtration, and word stemming

I first tested to see the effects of removing sarcastic tweets on the training set has on models. The results are shown below where negative values indicate a decrease in performance and positive an increase. Removing saracasm increased the Multinomial and bernoulli sentiment classifer's ability

distinguish between neutral and positive tweets more accurately; indicated by an increase in macro average recall (+1-2%), however accuracy and the macro averages for topics decreased, concluding that sarcastic tweets are okay to be left in for topic classifers. Changes were very small.

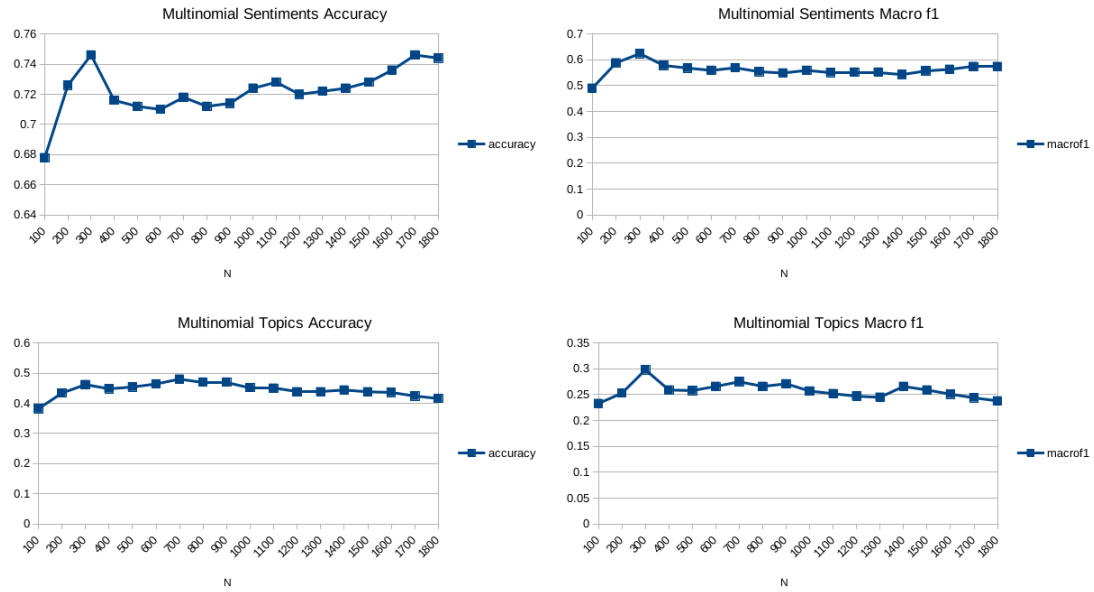
Comparison of standard models trained without sarcastic tweets vs standard models; both classifying test set

Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	**0**	-0.002	-0.007	-0.006	4
Decision Tree Topic	-0.032	0.016	-0.015	-0.011	48
Multinomial Bayes Sentiment	**0**	-0.002	0.012	0.01	17
Multinomial Bayes Topic	-0.004	0.005	**0**	0.002	25
Bernoulli Bayes Sentiment	0.008	-0.01	0.018	0.02	41
Bernoulli Bayes Topic	-0.002	-0.002	**0**	-0.001	40

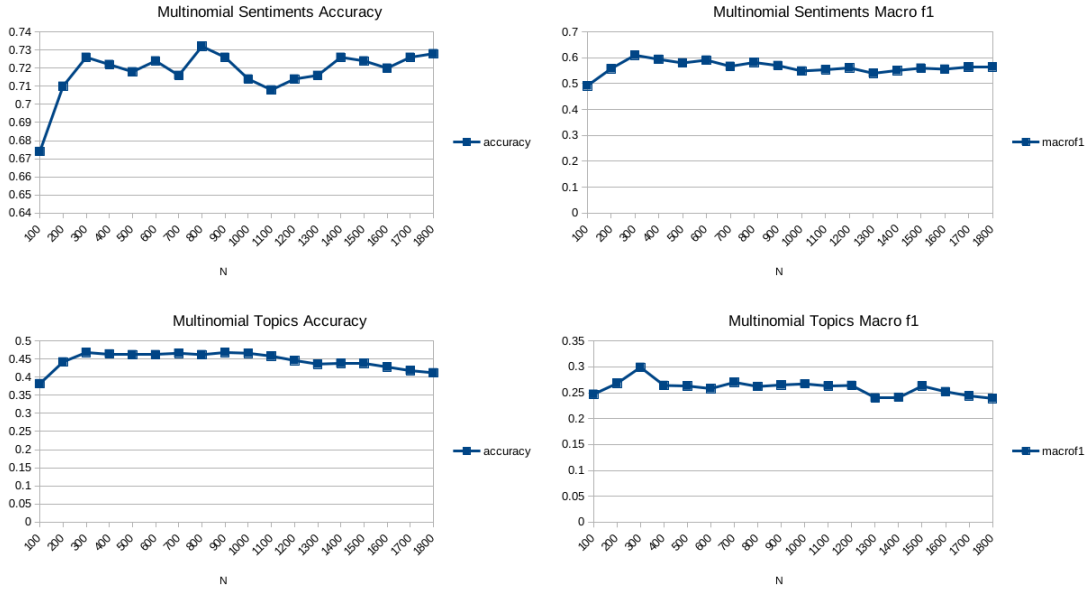
Improvements in performance = green cells, reduction in performance = red cells

I then experimented to see the effect stop word filtration and word stemming has with different maximum features (N) allowable fron N=100 to N = 1900 (examining the maxima observed in q2). I also did this with sarcasm ontop of stop word filtration and word stemming, and found that in both cases multinomial bayes was generally superior for sentiment and topic classification (validated from previous questions). The best results of sentiment and accuracy analysis with only stop word removal and word stemming occcured at N = 300 for sentiment analysis and N = 700 for topics. The best results of sentiment and accuracy analysis with stop word removal, word stemming, and sarcasm removal in the training set, occcured at N = 800 for sentiment analysis and N = 700 for topics. Differences were minor; sarcasm removal was shown to increase sentiment classifer macro values slightly for some maximum number of N, but generally decreases overall accuracy. For both type of sentiment analysis, accuracy increased as the value of N increased however average macro precision and recall peaks at a certain value of N.

Accuracy and macro f1 average of Multinomial Naive Bayes as N is increased on the test set using stop word filtration and word stemming



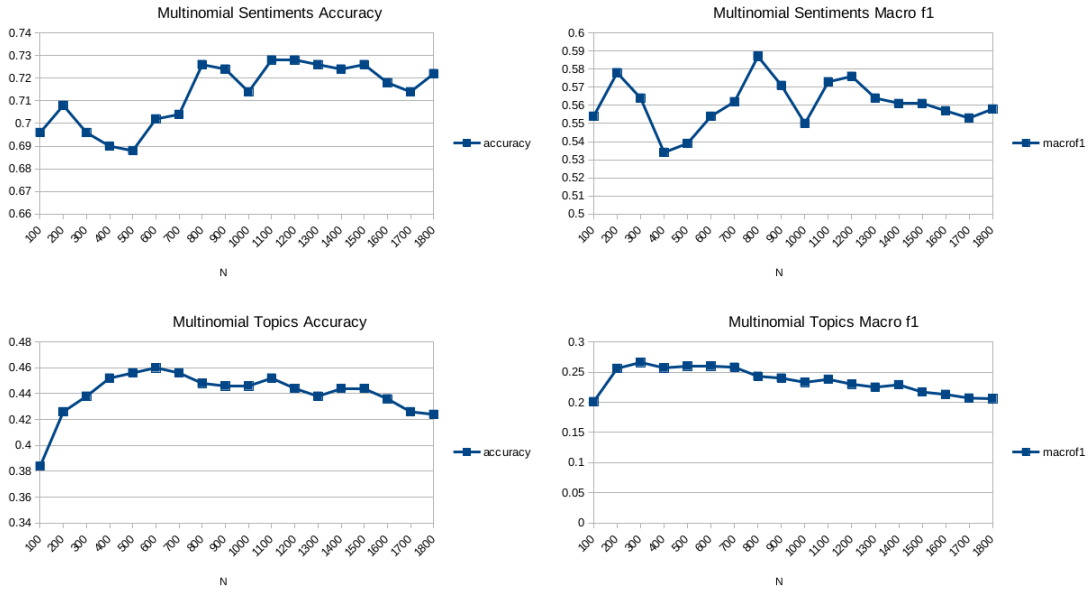
Accuracy and macro f1 average of Multinomial Naive Bayes as N is increased on the test set using stop word filtration, word stemming and removing sarcastic tweets



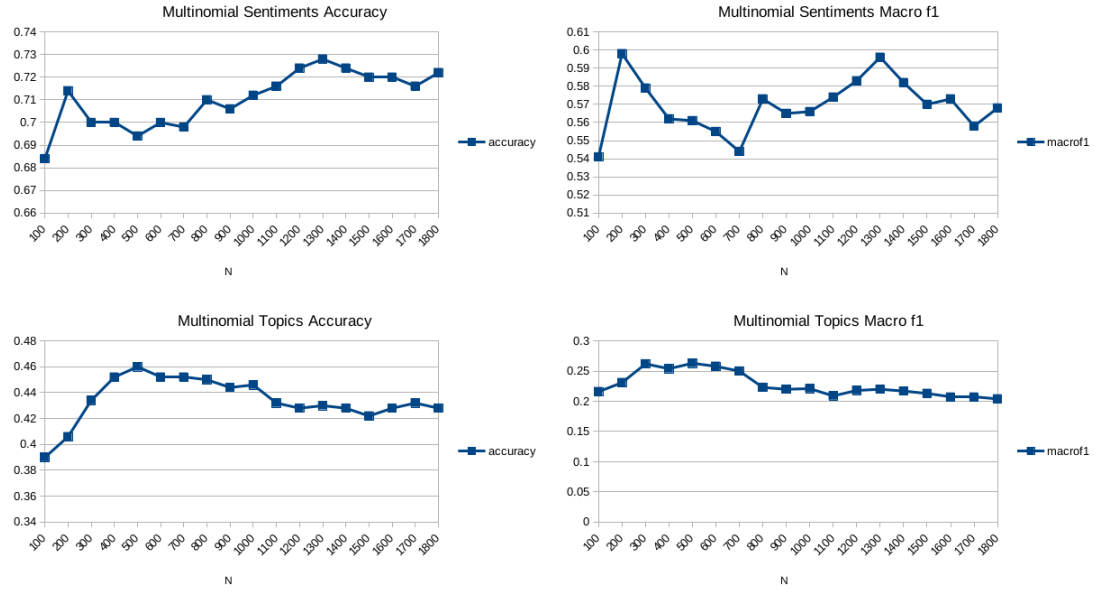
Tests were repeated using first 750 and last 750 tweets as the training set and mid 500 as the test set to see if these peaks are data specific biases, or genuine trends innate to the method of analysis. Analysis shows that the peaks in each graph was due to sample bias. Both cases however show similar graph behaviour after N = 1000.

Tests using first 750 and last 750 tweets as a training set and mid 500 tweets as test set

Accuracy and macro f1 average of Multinomial Naive Bayes as N is increased on the test set using stop word filtration and word stemming



Accuracy and macro f1 average of Multinomial Naive Bayes as N is increased on the test set using stop word filtration, word stemming and removing sarcastic tweets



Peak values from each test case

		Sentiment			Topic		
test case	sarcasm	Accuracy	Macro Avg f1	N	Accuracy	Macro Avg f1	N
Mnb_1	no	0.746	0.624	300	0.48	0.275	700
Mnb_1	yes	0.732	0.582	800	0.468	0.299	300
Mnb_2	no	0.726	0.587	800	0.46	0.26	600
Mnb_2	yes	0.724	0.582	1400	0.46	0.263	500

From the results, removing sarcasm from the training set generally reduces peak accuracy and macro average values for sentiment analysis. I've decided to go with a multinomial bayes classifer with stop word removal and word stemming with unlimited N. This is because both test cases have more stable above average values as N approaches 1900. For a topic classifer I've decided to go with a multinomial bayes classifer with stop word removal and word stemming at N = 700, where both topic models experimented with produced a local maximum at around those points.

My chosen classifiers showed overall improvement in classification ability when compared to the standard models and baselines on the test set. The only decline in performance is attributed to the increase in computation time.

Comparison of chosen classifiers to other models on test set

	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms
Chosen Sentiment Classifer	0.744	0.787	0.521	0.548	732
Chosen Topic Classifier	0.48	0.293	0.274	0.275	728

Change in values

Standard models					
Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.07	0.301	0.088	0.11	643
Decision Tree Topic	0.14	0.064	0.077	0.074	627
Multinomial Bayes Sentiment	0.006	0.054	0.032	0.033	504
Multinomial Bayes Topic	0.18	0.119	0.146	0.15	506
Bernoulli Bayes Sentiment	0.016	-0.04	0.095	0.12	454
Bernoulli Bayes Topic	0.296	0.265	0.22	0.253	437
Baselines					
Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Vader Sentiment	0.31	0.383	0.066	0.184	618
Majority Sentiment	0.074	0.564	0.188	0.281	n/a
Majority Topic	0.306	0.284	0.224	0.26	n/a

red indicates decline in performance, green indicates increase in performance

Comparison of chosen classifiers to other models on training set

	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Chosen Sentiment Classifer	0.924	0.936	0.828	0.869	827
Chosen Topic Classifier	0.751	0.722	0.592	0.627	820

Change in values

Standard models					
Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Decision Tree Sentiment	0.219	0.311	0.319	0.332	703
Decision Tree Topic	0.342	0.491	0.369	0.407	637
Multinomial Bayes Sentiment	-0.017	-0.025	-0.012	-0.015	524
Multinomial Bayes Topic	0.012	-0.069	0.097	0.072	536
Bernoulli Bayes Sentiment	0.06	0.007	0.223	0.257	420
Bernoulli Bayes Topic	0.472	0.56	0.501	0.559	382
Baselines					
Classifier	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg f1	Calc Time(ms)
Vader Sentiment	0.481	0.513	0.356	0.487	439
Majority Sentiment	0.285	0.723	0.495	0.609	n/a
Majority Topic	0.57	0.713	0.542	0.612	n/a

red indicates decline in performance, green indicates increase in performance