Appendix A - Tables

Appendix A1 - Respective chart: B1

Sample measure: Cross Validation score - Each column from #1 to #10 is the averaged result of N measures (first column)

Number of measures	AVG #1	AVG #2	AVG #3	AVG #4	AVG #5	AVG #6	AVG #7	AVG #8	AVG #9	AVG #10	Variance (MAX - MIN)	Average	Error % (Variance / Average * 100)
1	0.576	0.57	0.552	0.562	0.539	0.572	0.588	0.599	0.501	0.522	0.098	0.5581	17.56
5	0.581	0.591	0.549	0.535	0.563	0.56	0.536	0.547	0.542	0.521	0.070	0.5525	12.67
10	0.536	0.558	0.563	0.57	0.558	0.524	0.57	0.51	0.533	0.546	0.060	0.5468	10.97
25	0.529	0.553	0.55	0.564	0.558	0.518	0.557	0.558	0.557	0.547	0.046	0.5491	8.38
50	0.53	0.552	0.551	0.55	0.533	0.54	0.542	0.534	0.55	0.536	0.022	0.5418	4.06
100	0.563	0.562	0.57	0.567	0.569	0.562	0.56	0.568	0.564	0.562	0.010	0.5647	1.77

Appendix A2 - Respective chart: B2

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val variance (+/-)	Recall	Precision	F1-score	Testing time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	5	400	400	1,000	0.441	0.068	0.563	0.756	0.639	38 minutes	~ 440,000	0.600	12,305	8.13
Multinomial NB	5	400	400	2,500	0.571	0.068	0.804	0.769	0.781	40 minutes	~ 440,000	0.731	12,305	20.32
Multinomial NB	5	400	400	3,500	0.550	0.057	0.860	0.735	0.780	40 minutes	~ 440,000	0.731	12,305	28.44
Multinomial NB	5	400	400	5,000	0.605	0.058	0.853	0.767	0.804	50 minutes	~ 440,000	0.757	12,305	40.63
Multinomial NB	5	400	400	7,500	0.594	0.069	0.901	0.754	0.815	50 minutes	~ 440,000	0.766	12,305	60.95
Multinomial NB	5	400	400	10,000	0.582	0.070	0.926	0.749	0.821	0.9 hours	~ 440,000	0.770	12,305	81.27
Multinomial NB	5	400	400	12,305	0.574	0.064	0.924	0.740	0.814	1.1 hours	~ 440,000	0.763	12,305	100.00

Appendix A3 - Respective chart: B3

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val variance (+/-)	Recall	Precision	F1-score	Testing time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	5	400	/	1,000	0.469	0.037	0.694	0.701	0.676	1.20 hours	~ 1,280,000	0.635	18,616	5.37
Multinomial NB	5	400	/	2,500	0.539	0.033	0.795	0.734	0.759	1.40 hours	~ 1,280,000	0.707	18,616	13.43
Multinomial NB	5	400	/	3,500	0.544	0.038	0.801	0.737	0.724	1.5 hours	~ 1,280,000	0.702	18,616	18.80
Multinomial NB	5	400	/	7,500	0.583	0.036	0.908	0.744	0.816	1.9 hours	~ 1,280,000	0.763	18,616	40.29
Multinomial NB	5	400	/	10,000	0.593	0.035	0.923	0.750	0.825	3 hours	~ 1,280,000	0.773	18,616	53.72
Multinomial NB	5	400	/	18,616	0.590	0.034	0.937	0.748	0.830	3.6 hours	~ 1,280,000	0.776	18,616	100.00

Appendix A4 - Respective chart: B4

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val variance (+/-)	Recall	Precision	F1-score	Testing time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	5	400	400	1,000	0.384	0.068	0.691	0.664	0.663	38 minutes	~ 440,000	0.601	12,305	8.13
Bernoulli NB	5	400	400	2,500	0.412	0.065	0.828	0.677	0.727	40 minutes	~ 440,000	0.661	12,305	20.32
Bernoulli NB	5	400	400	3,500	0.392	0.060	0.841	0.649	0.720	40 minutes	~ 440,000	0.651	12,305	28.44
Bernoulli NB	5	400	400	5,000	0.404	0.061	0.850	0.662	0.724	50 minutes	~ 440,000	0.660	12,305	40.63
Bernoulli NB	5	400	400	7,500	0.406	0.065	0.872	0.661	0.732	50 minutes	~ 440,000	0.668	12,305	60.95
Bernoulli NB	5	400	400	10,000	0.410	0.070	0.879	0.656	0.732	0.8 hours	~ 440,000	0.669	12,305	81.27
Bernoulli NB	5	400	400	12,305	0.404	0.065	0.876	0.651	0.729	1.1 hours	~ 440,000	0.665	12,305	100.00

Appendix A5 - Respective chart: B5

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val variance (+/-)	Recall	Precision	F1-score	Testing time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	5	400	/	1,000	0.382	0.034	0.780	0.655	0.701	1.20 hours	~ 1,280,000	0.630	18,616	5.37
Bernoulli NB	5	400	/	2,500	0.427	0.035	0.817	0.662	0.728	1.40 hours	~ 1,280,000	0.659	18,616	13.43
Bernoulli NB	5	400	/	3,500	0.394	0.034	0.827	0.649	0.765	1.5 hours	~ 1,280,000	0.659	18,616	18.80
Bernoulli NB	5	400	/	7,500	0.438	0.036	0.878	0.667	0.755	1.9 hours	~ 1,280,000	0.685	18,616	40.29
Bernoulli NB	5	400	/	10,000	0.396	0.040	0.888	0.654	0.748	3 hours	~ 1,280,000	0.672	18,616	53.72
Bernoulli NB	5	400	/	18,616	0.415	0.040	0.893	0.661	0.893	3.6 hours	~ 1,280,000	0.716	18,616	100.00

Appendix A6 - Respective chart: B6

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	10	200	200	1,000	0.245	0.056	0.232	0.577	0.306	40 minutes	~ 700,000	0.340	19,848	5.04
Multinomial NB	10	200	200	2,500	0.455	0.063	0.467	0.741	0.550	50 minutes	~ 700,000	0.553	19,848	12.60
Multinomial NB	10	200	200	3,500	0.473	0.060	0.522	0.761	0.597	1 hour	~ 700,000	0.588	19,848	17.63
Multinomial NB	10	200	200	7,500	0.542	0.068	0.505	0.822	0.595	1.2 hours	~ 700,000	0.616	19,848	37.79
Multinomial NB	10	200	200	10,000	0.565	0.066	0.588	0.827	0.670	1.35 hours	~ 700,000	0.663	19,848	50.38
Multinomial NB	10	200	200	15,000	0.550	0.063	0.594	0.806	0.666	1.4 hours	~ 700,000	0.654	19,848	75.57
Multinomial NB	10	200	200	19,848	0.618	0.058	0.663	0.851	0.731	1.6 hours	~ 700,000	0.716	19,848	100.00

Appendix A7 - Respective chart: B7

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	10	200	/	1,000	0.356	0.032	0.507	0.586	0.498	2.20 hours	~ 1,900,000	0.487	29345	3.41
Multinomial NB	10	200	/	2,500	0.466	0.030	0.681	0.689	0.673	2.40 hours	~ 1,900,000	0.627	29345	8.52
Multinomial NB	10	200	/	3,500	0.468	0.033	0.702	0.691	0.685	2.40 hours	~ 1,900,000	0.637	29345	11.93
Multinomial NB	10	200	/	7,500	0.521	0.033	0.795	0.714	0.741	2.8 hours	~ 1,900,000	0.693	29345	25.56
Multinomial NB	10	200	/	10,000	0.489	0.034	0.843	0.693	0.747	3.10 hours	~ 1,900,000	0.693	29345	34.08
Multinomial NB	10	200	/	15,000	0.515	0.033	0.868	0.704	0.765	4.5 hours	~ 1,900,000	0.713	29345	51.12
Multinomial NB	10	200	/	20,000	0.493	0.033	0.878	0.694	0.763	5 hours	~ 1,900,000	0.707	29345	68.15

Appendix A8 - Respective chart: B8

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	10	200	200	1,000	0.210	0.054	0.258	0.410	0.287	40 minutes	~ 700,000	0.291	19841	5.04
Bernoulli NB	10	200	200	2,500	0.327	0.059	0.462	0.552	0.470	50 minutes	~ 700,000	0.453	19841	12.60
Bernoulli NB	10	200	200	3,500	0.319	0.056	0.497	0.582	0.499	1 hour	~ 700,000	0.474	19841	17.64
Bernoulli NB	10	200	200	7,500	0.311	0.068	0.461	0.580	0.476	1.2 hours	~ 700,000	0.457	19841	37.80
Bernoulli NB	10	200	200	10,000	0.309	0.059	0.494	0.569	0.491	1.35 hours	~ 700,000	0.466	19841	50.40
Bernoulli NB	10	200	200	15,000	0.308	0.060	0.489	0.577	0.498	1.4 hours	~ 700,000	0.468	19841	75.60
Bernoulli NB	10	200	200	19841	0.314	0.062	0.513	0.577	0.505	1.6 hours	~ 700,000	0.477	19841	100.00

Appendix A9 - Respective chart: B9

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	10	200	/	1,000	0.335	0.032	0.576	0.576	0.513	2.20 hours	~ 1,900,000	0.500	29345	3.41
Bernoulli NB	10	200	/	2,500	0.268	0.030	0.766	0.546	0.627	2.40 hours	~ 1,900,000	0.552	29345	8.52
Bernoulli NB	10	200	/	3,500	0.296	0.030	0.759	0.542	0.621	2.40 hours	~ 1,900,000	0.555	29345	11.93
Bernoulli NB	10	200	/	7,500	0.285	0.032	0.773	0.527	0.616	2.8 hours	~ 1,900,000	0.550	29345	25.56
Bernoulli NB	10	200	/	10,000	0.279	0.029	0.789	0.549	0.637	3.10 hours	~ 1,900,000	0.564	29345	34.08
Bernoulli NB	10	200	/	15,000	0.263	0.028	0.803	0.547	0.637	4.5 hours	~ 1,900,000	0.563	29345	51.12
Bernoulli NB	10	200	/	20,000	0.264	0.030	0.809	0.547	0.640	5 hours	~ 1,900,000	0.565	29345	68.15

Appendix A10 - Respective chart: B10

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	20	99	99	3,500	0.288	0.067	0.260	0.431	0.312	50 minutes	~700,000	0.406	20,000	17.50
Multinomial NB	20	99	99	7,500	0.304	0.055	0.257	0.403	0.299	1.2 hours	~700,000	0.316	20,000	37.50
Multinomial NB	20	99	99	10,000	0.303	0.059	0.292	0.450	0.315	1.7 hours	~700,000	0.340	20,000	50.00
Multinomial NB	20	99	99	15,000	0.343	0.061	0.282	0.447	0.329	2 hours	~700,000	0.350	20,000	75.00
Multinomial NB	20	99	99	20,000	0.350	0.055	0.301	0.460	0.346	2.5 hours	~700,000	0.364	20,000	100.00

Appendix A11 - Respective chart: B11

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Multinomial NB	20	99	/	3,500	0.375	0.032	0.504	0.590	0.518	2.4 hours	2,374,629	0.497	37198	9.41
Multinomial NB	20	99	/	7,500	0.395	0.032	0.680	0.593	0.619	3.4 hours	2,374,629	0.572	37198	20.16
Multinomial NB	20	99	/	10,000	0.429	0.030	0.742	0.606	0.655	4 hours	2,374,629	0.608	37198	26.88
Multinomial NB	20	99	/	15,000	0.457	0.033	0.715	0.637	0.661	4.5 hours	2,374,629	0.618	37198	40.32
Multinomial NB	20	99	/	20,000	0.465	0.033	0.725	0.645	0.671	6 hours	2,374,629	0.627	37198	53.77
Multinomial NB	20	99	/	30,000	0.431	0.030	0.758	0.610	0.664	8.8 hours	2,374,629	0.616	37198	80.65

Appendix A12 - Respective chart: B12

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	20	99	99	3,500	0.106	0.038	0.125	0.158	0.133	50 minutes	~700,000	0.131	20,000	17.50
Bernoulli NB	20	99	99	7,500	0.101	0.040	0.127	0.155	0.133	1.2 hours	~700,000	0.129	20,000	37.50
Bernoulli NB	20	99	99	10,000	0.099	0.035	0.126	0.162	0.136	1.7 hours	~700,000	0.131	20,000	50.00
Bernoulli NB	20	99	99	15,000	0.113	0.037	0.135	0.154	0.137	2 hours	~700,000	0.135	20,000	75.00
Bernoulli NB	20	99	99	20000	0.102	0.038	0.123	0.153	0.129	2.5 hours	~700,000	0.127	20,000	100.00

Appendix A13 - Respective chart: B13

Type	Labels	Min docs	Max docs	Selected features	Cross val	Cross val precision	Recall	Precision	F1-score	Time	Vocabulary	Average Score	Features available	selected features / features (%)
Bernoulli NB	20	99	/	3,500	0.239	0.029	0.613	0.431	0.492	2.4 hours	2,374,629	0.444	37198	9.41
Bernoulli NB	20	99	/	7,500	0.206	0.026	0.683	0.435	0.521	3.4 hours	2,374,629	0.461	37198	20.16
Bernoulli NB	20	99	/	10,000	0.222	0.025	0.701	0.456	0.542	4 hours	2,374,629	0.480	37198	26.88
Bernoulli NB	20	99	/	15,000	0.205	0.025	0.693	0.443	0.532	4.5 hours	2,374,629	0.468	37198	40.32
Bernoulli NB	20	99	/	20,000	0.195	0.025	0.698	0.420	0.515	6 hours	2,374,629	0.457	37198	53.77
Bernoulli NB	20	99	/	30,000	0.195	0.027	0.698	0.437	0.528	8.8 hours	2,374,629	0.465	37198	80.65