Appendix A: Datasets

Dataset	Source	Docs	Words	Ratio	Cutoff	Classes	Class Sizes (sorted)
cora36	whizbang.com	1800	5171	3	3	36	50 (each)
fbis	TREC	2463	2000	1	10	17	38 43 46 46 46 48 65 92 94 119 121 125 139 190 358 387 506
la1	TREC	3204	31472	10	1	6	273 341 354 555 738 943
la2	TREC	3075	31472	10	1	6	248 301 375 487 759 905
oh0	OHSUMED	1003	3182	3	3	10	51 56 57 66 71 76 115 136 181 194
oh5	OHSUMED	918	3012	3	3	10	59 61 61 72 74 85 93 120 144 149
oh10	OHSUMED	1050	3238	3	3	10	52 60 61 70 87 116 126 148 165 165
oh15	OHSUMED	913	3100	3	3	10	53 56 56 66 69 98 98 106 154 157
ohscal	OHSUMED	11162	11465	1	3	10	709 764 864 1001 1037 1159 1260 1297 1450 1621
re0	Reuters-21578	1504	2886	2	3	13	11 15 16 20 37 38 39 42 60 80 219 319 608
re1	Reuters-21578	1657	3758	2	3	25	10 13 15 17 18 18 19 19 20 20 27 31 31 32 37 42 48 50 60 87 99 106 137 330 371
tr11	TREC	414	6429	16	3	9	6 11 20 21 29 52 69 74 132
tr12	TREC	313	5804	19	3	8	9 29 29 30 34 35 54 93
tr21	TREC	336	7902	24	3	6	4 9 16 35 41 231
tr23	TREC	204	5832	29	3	6	6 11 15 36 45 91
tr31	TREC	927	10128	11	3	7	2 21 63 111 151 227 352
tr41	TREC	878	7454	8	3	10	9 18 26 33 35 83 95 162 174 243
tr45	TREC	690	8261	12	3	10	14 18 36 47 63 67 75 82 128 160
wap	WebACE	1560	8460	5	3	20	5 11 13 15 18 33 35 37 40 44 54 65 76 91 91 97 130 168 196 341

The ‘ratio’ is the number of words divided by the number of documents, and is directly influenced by the rare-word cutoff used, shown in the following column.

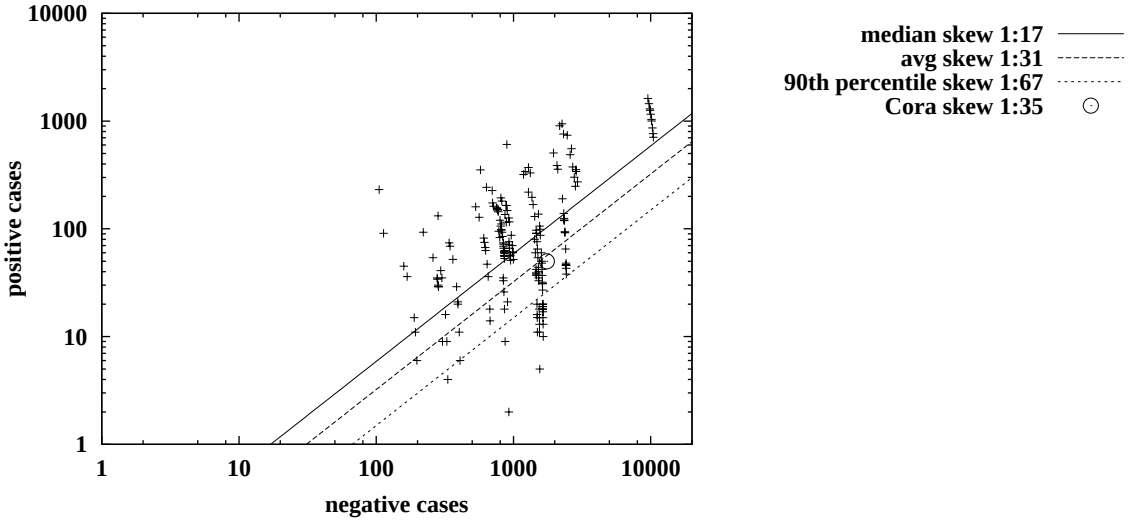

Figure 12. Sizes of positive and negative classes for each of the 229 binary classification tasks. Note the Cora dataset has 36 data points overlaid at (1750,35).