Table 4.14: Topic categorization experiments datasets.

(a) Sampling and statistics of TC(EN).

| | A(20) | B(7) | C(5) | D(4) |
|---|---|---|---|---|
| alt.atheism | 480/319 | 480/319 | | |
| comp.graphics | 584/389 | 2936/1955 | 584/389 | |
| comp.os.ms-windows.misc | 591/394 | | 591/394 | |
| comp.sys.ibm.pc.hardware | 590/392 | | 590/392 | |
| comp.sys.mac.hardware | 578/385 | | 578/385 | |
| comp.windows.x | 593/395 | | 593/395 | |
| misc.forsale | 585/390 | 585/390 | | |
| rec.autos | 594/396 | 2389/1590 | | |
| rec.motorcycles | 598/398 | | | |
| rec.sport.baseball | 597/397 | | | |
| rec.sport.hockey | 600/399 | | | |
| sci.crypt | 595/396 | 2373/1579 | | |
| sci.electronics | 591/393 | | | |
| sci.med | 594/396 | | | |
| sci.space | 593/394 | | | |
| soc.religion.christian | 599/398 | 599/398 | | |
| talk.politics.guns | 546/364 | 1952/1301 | | 546/364 |
| talk.politics.mideast | 564/376 | | | 564/376 |
| talk.politics.misc | 465/310 | | | 465/310 |
| talk.religion.misc | 377/251 | | | 377/251 |
| | 11314/7532 | 11314/7532 | 2936/1955 | 1952/1301 |

(b) Sampling and statistics of TC(ZH).

| | A(20) | B(9) | C(11) |
|---|---|---|---|
| Art | 740/742 | 740/742 | |
| Literature | 33/34 | | 33/34 |
| Education | 59/61 | | 59/61 |
| Philosophy | 44/45 | | 44/45 |
| History | 466/468 | 466/468 | |
| Space | 640/642 | 640/642 | |
| Energy | 32/33 | | 32/33 |
| Electronics | 27/28 | | 27/28 |
| Communication | 25/27 | | 25/27 |
| Computer | 1357/1358 | 1357/1358 | |
| Mine | 33/34 | | 33/34 |
| Transport | 57/59 | | 57/59 |
| Environment | 1217/1218 | 1217/1218 | |
| Agriculture | 1021/1022 | 1021/1022 | |
| Economy | 1600/1601 | 1600/1601 | |
| Law | 51/52 | | 51/52 |
| Medical | 51/53 | | 51/53 |
| Military | 74/76 | | 74/76 |
| Politics | 1024/1026 | 1024/1026 | |
| Sports | 1253/1254 | 1253/1254 | |
| | 9804/9833 | 9318/9331 | 486/502 |