**Measuring the quality of the automated parsing**

We evaluated the automated parsing system, composed of Python scripts we wrote for this research, to gauge the possible mechanical errors. The related files are available in the provided GitHub repository. We tested the performance in the following critical tasks in comparison with the human standard: (1) identifying the correct number of relevant sentences (sentences within speeches) and attributing them to the correct political parties, and (2) finding all the sentences with a religion reference, attributing them to the correct religion (i.e., using the appropriate keyword class), and adequately ignoring all the sentences without such a reference.

We carried out the test with a sample: a fabricated test document with 1000 relevant sentences (i.e., sentences within speeches). The document was composed by manually selecting authentic sentences from the data corpus, with minor edits. This was done so that all the parties and keyword classes were present in a balanced way. There were precisely 500 relevant sentences related to religion and 500 sentences unrelated to religion (according to human assessment). The test document was finalized by conforming it to the textual conventions of the actual data (e.g., by starting the document with the meeting details and nesting the sentences within speeches by various MPs). This finalization causes the total number of sentences to be 1008. The test document is in the GitHub repository's "data" folder.

We observe the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the automated parsing system by comparing its output with the correct, human-assigned positive (P) and negative (N) values. We calculate the Matthews correlation coefficient (MCC) to assess each task's performance separately. As MCC measures binary classification performance, we evaluate the system individually with each political party and keyword class (i.e., religion). As a relatively strict measure, MCC produces a high score only if good results are gained in all four categories (TP, TN, FP, FN). The results are presented below.

*Table: The performance test*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Task target** | **Correct values** | | **Automated parsing results** | | | |  |
| **Relevant sentences** | **P** | **N** | **TP** | **TN** | **FP** | **FN** | **MCC** |
| Christian Democrats | 112 | 896 | 112 | 896 | 0 | 0 | 1 |
| Centre Party | 111 | 897 | 110 | 873 | 23 | 1 | 0,89 |
| National Coalition Party | 111 | 897 | 98 | 884 | 0 | 13 | 0,93 |
| Finns Party | 111 | 897 | 104 | 889 | 1 | 7 | 0,96 |
| Swedish People’s Party of Finland | 111 | 897 | 77 | 863 | 0 | 34 | 0,82 |
| Social Democratic Party of Finland | 111 | 897 | 109 | 895 | 0 | 2 | 0,99 |
| Blue Reform | 111 | 897 | 89 | 861 | 14 | 22 | 0,81 |
| Left Alliance | 111 | 897 | 108 | 894 | 0 | 3 | 0,98 |
| Green League | 111 | 897 | 110 | 888 | 8 | 1 | 0,96 |
| **Relevant sentences with a religion reference** |  |  |  |  |  |  |  |
| ALL KEYWORDS | 500 | 508 | 471 | 508 | 0 | 29 | 0,94 |
| General | 150 | 858 | 134 | 842 | 1 | 15 | 0,94 |
| Christianity | 150 | 858 | 145 | 853 | 0 | 5 | 0,98 |
| Islam | 150 | 858 | 142 | 851 | 0 | 7 | 0,97 |
| Judaism | 40 | 968 | 40 | 968 | 0 | 0 | 1 |
| Other minority religions | 10 | 998 | 10 | 968 | 0 | 0 | 1 |

The MCC is a version of the Pearson Correlation Coefficient transferred to a binary classification and performance testing context.[[1]](#footnote-1) Thus, its output value (from -1 to 1) is interpreted the same as Pearson Correlation Coefficient, where anything above 0.75 suggests a very high positive relationship. As reported in Table 1, the results range from 0,81 to 1, generally indicating a good performance across the board.

1. See Powers, David M. W. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* 2 (1): 37–63. [↑](#footnote-ref-1)