Preliminary experiment

We started our experiments with a preliminary phase where we evaluated three types of division proposed on the PubMed 200k RCT dataset for 1000 texts. We consider in the following three tables, the baseline as the Bert classifier without data increase.

Symmetrical division 1 evaluation: We sought to assess the influence of the size of the portion of text removed at the ends, which we varied as follows 5%,15%,25%,35%,45%. The results are presented in the table 1. The best accuracy is obtained for the smallest deletions, which suggests that the increase works best when the modifications to the original text are the least significant.

Table 1: Symmetrical division 1: study according to the proportion of texts deleted per sample at the ends.

Division size(%)	0.05	0.15	0.25	0.35	0.45
Baseline without DA	0.6513	0.6513	0.6513	**0.6513**	**0.6513**
DA - symmetrical division 1	**0.6666**	**0.6543**	**0.6546**	0.6506	0.6469

Division 2 assessment by sliding window: We evaluated for the same window size, 90% of the initial text, an offset of 1 to 5 words. The results are presented in the table 2. The best results are obtained with the largest stride of 5 words.

Table 2: Division 2 per sliding window: study according to the stride used to slide the window

stride (word)	1	2	3	4	5
Baseline without DA	0.6513	0.6513	0.6513	0.6513	0.6513
DA - division 2 window	**0.6586**	**0.651**	**0.6533**	**0.663**	**0.662**

Evaluation of division 3 in equal parts: We evaluated the impact of the number of divisions in 2,3,4,5 and 6 parts. The results are presented in the table 3. We find that dividing the document in half gives the best results.

Table 3: Division 3 by equal division

Number of parts	2	3	4	5	6
Baseline without DA	0.6513	0.6513	0.6513	0.6513	0.6513
DA - pyramid division 3	**0.677**	**0.659**	**0.6513**	**0.6553**	0.6283

We retain from these preliminary experiments that division 1 with the smallest percentage of division as well as division 3 into 2 portions of texts.

In the rest of our experiments, we will use a combination of these two cuts that we will call pyramid .