Research results

Ranking
The figures **??** & **??** show the result of the categorization task as ranking results. The rank indicates the position of the correct journal in the sorted list of matched journals. Figure **??** shows the ranking results for the different sets based on the title. Figure **??** displays the ranking results based on the abstract. Both graphs show both average and median ranks, based on the cosine-similarity between the article and journal embeddings or feature vectors.

Figure 1: Median and average title rankings

Figure 2: Median and average abstract rankings

Rank distribution
Figures **??** & **??** show the distributions of the ranks for each set. The figures plot the summed amount of articles against the ranks on a logarithmic scale. Figure **??** shows the rank distribution for the titles, figure **??** shows this for the ranks based on the abstract. These graphs give a detailed view of the ranks presented in their respective figures **??** & **??**.

Figure 3: Title rank distribution per set

Figure 4: Abstract rank distribution per set

F1-Score
Figures **??** & **??** show the precision, recall and f1 scores. These scores are calculated on journal level, and are averaged per set. Figure **??** shows the F1 score for the title and figure **??** shows the scores for the abstract. These scores indicate the performance of the sets on absolute hits/top-1.

Figure 5: Precision, recall and F1 scores based on title

Figure 6: Precision, recall and F1 scores based on abstract

Memory usage
Table **??** shows the total memory usage of each set for the *Validation set*, indicating their storage costs in gigabytes[1].

Set	Size in GB	Absolute hit percentage		Median title rank	Median abstract rank
		Title	Abstract		
tfidf 5k 5K	9.82	5.42%	10.18%	50	27
tfidf 5K 10K	11.47	6.49%	11.08%	38	15
tfidf 10K 10K	**11.61**	**6.79%**	**11.32%**	**35**	**14**
embedding	3.13	7.92%	9.24%	27	23
5k embedding	3.13	6.34%	8.36%	42	27
10k embedding	3.13	7.03%	8.76%	34	25
tfidf embedding	**3.13**	**7.89%**	**9.33%**	**27**	**22**
1k 6k embedding	3.06	5.16%	7.86%	64	31

Table 1: Memory usage and performance for each set

Journal relatedness plot
Figure **??** is the two dimensional plot of the journal embeddings. The journals are color-marked by publisher. Red is Wiley, lime is Elsevier and blue is Springer Nature. The grey points are other or not-specified. Figure **??** shows the journals, grouped by a k-means algorithm, creating 8 groups. The bottom right shows the names of the journals closest to the center of the group. The k-means algorithm ran on the 300-dimensional set, while the plot shows the 2-dimensional set.

[1] 1024 based

Figure 7: Journal plot grouped by publisher

Figure 8: Journal plot grouped k-means grouped