Validation of Text Similarity Methods

Matching of Grant Proposals to Reviewers: PostDoc Mobility 2021

Gabriel Okasa & Anne Jorstad

Validation Summary

We validate several transformer models with various hyperparameter settings and data inputs for the task of semantic textual similarity.

We do not impose any restrictions for the similarity search and restrict the applications and referees' texts to English texts only. We keep only referees that have at least English 10 publications.

Given that the context window for transformers is limited to 512 tokens, we truncate the texts exceeding this length from the end of the text sequence.

For matching the referees, we take into account the similarity average of 20 percent most similar publication of a given referee.

The research area distribution for the validated applications is as follows:

- LS: 123 (0.38)
- MINT: 147 (0.46)
- SSH: 50 (0.16)

We validate the following models:

- allenai/scibert_scivocab_uncased; allenai/specter2_base; bert-base-uncased

and extract the text embeddings using:

- cls_token; mean_pooling

for the following type of texts:

- abstract; title; title_abstract

based on the following years of publications:

- 5; 10

For each validation scenario, we extract the text embeddings and compute the similarity between the applications and each publication of a referee based on the cosine similarity.

We measure the performance of the methods based on the mean average precision at $K = 2$ and $K = 5$.

Validation Results

Table 1: Validation results: Mean Average Precision

model	embedding	years	text	map_at_2	map_at_5
specter2_base	mean_pooling	10	abstract	0.3703	0.4687
specter2_base	mean_pooling	10	title_abstract	0.3680	0.4619
specter2_base	cls_token	10	abstract	0.3648	0.4605
specter2_base	cls_token	10	title_abstract	0.3656	0.4576
specter2_base	cls_token	5	abstract	0.3578	0.4554
specter2_base	mean_pooling	5	title_abstract	0.3602	0.4536
specter2_base	cls_token	5	title_abstract	0.3586	0.4520
specter2_base	mean_pooling	5	abstract	0.3562	0.4518
scibert_scivocab_uncased	mean_pooling	10	abstract	0.3117	0.4205
scibert_scivocab_uncased	mean_pooling	10	title_abstract	0.3125	0.4184
specter2_base	cls_token	10	title	0.3078	0.4034
scibert_scivocab_uncased	mean_pooling	5	abstract	0.2938	0.3949
specter2_base	cls_token	5	title	0.2938	0.3908
scibert_scivocab_uncased	mean_pooling	5	title_abstract	0.2867	0.3905
specter2_base	mean_pooling	10	title	0.2914	0.3842
specter2_base	mean_pooling	5	title	0.2945	0.3745
bert-base-uncased	mean_pooling	10	title_abstract	0.2758	0.3696
bert-base-uncased	mean_pooling	5	abstract	0.2734	0.3684
bert-base-uncased	mean_pooling	10	abstract	0.2734	0.3675
bert-base-uncased	mean_pooling	5	title_abstract	0.2719	0.3653
scibert_scivocab_uncased	mean_pooling	10	title	0.2531	0.3316
scibert_scivocab_uncased	cls_token	10	title	0.2508	0.3298
bert-base-uncased	mean_pooling	10	title	0.2383	0.3175
scibert_scivocab_uncased	mean_pooling	5	title	0.2398	0.3167
bert-base-uncased	mean_pooling	5	title	0.2383	0.3117
scibert_scivocab_uncased	cls_token	5	title	0.2266	0.3104
bert-base-uncased	cls_token	10	title_abstract	0.2070	0.2917
bert-base-uncased	cls_token	5	title_abstract	0.1938	0.2719
bert-base-uncased	cls_token	10	abstract	0.1969	0.2718
bert-base-uncased	cls_token	5	abstract	0.1727	0.2456
scibert_scivocab_uncased	cls_token	10	abstract	0.1500	0.2123
scibert_scivocab_uncased	cls_token	5	abstract	0.1461	0.2001
bert-base-uncased	cls_token	10	title	0.1469	0.2000
scibert_scivocab_uncased	cls_token	5	title_abstract	0.1414	0.1941
bert-base-uncased	cls_token	5	title	0.1391	0.1937
scibert_scivocab_uncased	cls_token	10	title_abstract	0.1328	0.1908

Validation Results by Research Area

Table 2: Validation results: Mean Average Precision by Research Area: LS

model	embedding	years	text	map_at_2	map_at_5
specter2_base	mean_pooling	5	abstract	0.3638	0.4634
specter2_base	mean_pooling	5	title_abstract	0.3598	0.4584
specter2_base	mean_pooling	10	abstract	0.3638	0.4583
specter2_base	mean_pooling	10	title_abstract	0.3516	0.4460
specter2_base	cls_token	10	abstract	0.3415	0.4430
specter2_base	cls_token	5	title_abstract	0.3455	0.4364
specter2_base	cls_token	10	title_abstract	0.3455	0.4363
specter2_base	cls_token	5	abstract	0.3435	0.4353
scibert_scivocab_uncased	mean_pooling	10	abstract	0.3313	0.4239
specter2_base	cls_token	10	title	0.3150	0.4091
scibert_scivocab_uncased	mean_pooling	10	title_abstract	0.3191	0.4054
specter2_base	mean_pooling	10	title	0.2927	0.3884
scibert_scivocab_uncased	mean_pooling	5	abstract	0.2967	0.3874
specter2_base	cls_token	5	title	0.2785	0.3840
scibert_scivocab_uncased	mean_pooling	5	title_abstract	0.2846	0.3780
specter2_base	mean_pooling	5	title	0.2927	0.3758
bert-base-uncased	mean_pooling	5	abstract	0.2663	0.3530
bert-base-uncased	mean_pooling	10	title_abstract	0.2642	0.3500
bert-base-uncased	mean_pooling	5	title_abstract	0.2561	0.3493
bert-base-uncased	mean_pooling	10	abstract	0.2622	0.3411
scibert_scivocab_uncased	cls_token	10	title	0.2195	0.2920
scibert_scivocab_uncased	mean_pooling	10	title	0.2195	0.2870
bert-base-uncased	cls_token	10	title_abstract	0.1931	0.2761
bert-base-uncased	cls_token	5	title_abstract	0.1911	0.2744
scibert_scivocab_uncased	mean_pooling	5	title	0.2134	0.2711
scibert_scivocab_uncased	cls_token	5	title	0.1850	0.2673
bert-base-uncased	mean_pooling	5	title	0.1951	0.2556
bert-base-uncased	cls_token	10	abstract	0.1829	0.2514
bert-base-uncased	mean_pooling	10	title	0.1768	0.2502
bert-base-uncased	cls_token	5	abstract	0.1585	0.2370
scibert_scivocab_uncased	cls_token	10	abstract	0.1484	0.1970
scibert_scivocab_uncased	cls_token	5	abstract	0.1341	0.1884
scibert_scivocab_uncased	cls_token	10	title_abstract	0.1220	0.1775
scibert_scivocab_uncased	cls_token	5	title_abstract	0.1179	0.1694
bert-base-uncased	cls_token	10	title	0.1179	0.1446
bert-base-uncased	cls_token	5	title	0.0915	0.1315

Table 3: Validation results: Mean Average Precision by Research Area: MINT

model	embedding	years	text	map_at_2	map_at_5
specter2_base	cls_token	5	title_abstract	0.4116	0.5153
specter2_base	cls_token	5	abstract	0.4048	0.5140
specter2_base	cls_token	10	title_abstract	0.4082	0.5123
specter2_base	cls_token	10	abstract	0.4031	0.5033
specter2_base	mean_pooling	10	title_abstract	0.4031	0.4994
specter2_base	mean_pooling	10	abstract	0.3980	0.4985
specter2_base	mean_pooling	5	title_abstract	0.3827	0.4852
specter2_base	mean_pooling	5	abstract	0.3861	0.4827
scibert_scivocab_uncased	mean_pooling	10	title_abstract	0.3333	0.4563
scibert_scivocab_uncased	mean_pooling	10	abstract	0.3384	0.4536
specter2_base	cls_token	10	title	0.3418	0.4459
specter2_base	cls_token	5	title	0.3418	0.4444
scibert_scivocab_uncased	mean_pooling	5	abstract	0.3282	0.4327
scibert_scivocab_uncased	mean_pooling	5	title_abstract	0.3197	0.4324
specter2_base	mean_pooling	10	title	0.3265	0.4213
specter2_base	mean_pooling	5	title	0.3282	0.4154
bert-base-uncased	mean_pooling	10	abstract	0.3061	0.4154
bert-base-uncased	mean_pooling	10	title_abstract	0.3061	0.4126
scibert_scivocab_uncased	mean_pooling	10	title	0.3163	0.4093
bert-base-uncased	mean_pooling	5	abstract	0.2942	0.4018
bert-base-uncased	mean_pooling	5	title_abstract	0.3010	0.4018
bert-base-uncased	mean_pooling	10	title	0.3129	0.3988
scibert_scivocab_uncased	cls_token	10	title	0.3027	0.3976
scibert_scivocab_uncased	mean_pooling	5	title	0.2823	0.3852
bert-base-uncased	mean_pooling	5	title	0.2908	0.3830
scibert_scivocab_uncased	cls_token	5	title	0.2789	0.3768
bert-base-uncased	cls_token	10	title_abstract	0.2262	0.3251
bert-base-uncased	cls_token	10	abstract	0.2126	0.3051
bert-base-uncased	cls_token	5	title_abstract	0.2245	0.3048
bert-base-uncased	cls_token	5	abstract	0.2041	0.2813
bert-base-uncased	cls_token	5	title	0.1786	0.2478
bert-base-uncased	cls_token	10	title	0.1735	0.2430
scibert_scivocab_uncased	cls_token	10	abstract	0.1497	0.2273
scibert_scivocab_uncased	cls_token	5	title_abstract	0.1650	0.2217
scibert_scivocab_uncased	cls_token	10	title_abstract	0.1446	0.2106
scibert_scivocab_uncased	cls_token	5	abstract	0.1497	0.2094

Table 4: Validation results: Mean Average Precision by Research Area: SSH

model	embedding	years	text	map_at_2	map_at_5
specter2_base	mean_pooling	10	abstract	0.305	0.4070
specter2_base	mean_pooling	10	title_abstract	0.305	0.3908
specter2_base	cls_token	10	abstract	0.310	0.3775
specter2_base	cls_token	10	title_abstract	0.290	0.3493
specter2_base	mean_pooling	5	title_abstract	0.295	0.3492
scibert_scivocab_uncased	mean_pooling	10	title_abstract	0.235	0.3388
specter2_base	mean_pooling	5	abstract	0.250	0.3327
specter2_base	cls_token	5	abstract	0.255	0.3325
scibert_scivocab_uncased	mean_pooling	10	abstract	0.185	0.3150
bert-base-uncased	mean_pooling	5	abstract	0.230	0.3080
specter2_base	cls_token	5	title_abstract	0.235	0.3038
scibert_scivocab_uncased	mean_pooling	5	abstract	0.185	0.3020
scibert_scivocab_uncased	mean_pooling	5	title_abstract	0.195	0.2983
bert-base-uncased	mean_pooling	5	title_abstract	0.225	0.2973
bert-base-uncased	mean_pooling	10	abstract	0.205	0.2915
bert-base-uncased	mean_pooling	10	title_abstract	0.215	0.2913
specter2_base	mean_pooling	10	title	0.185	0.2647
specter2_base	cls_token	10	title	0.190	0.2643
specter2_base	mean_pooling	5	title	0.200	0.2513
specter2_base	cls_token	5	title	0.190	0.2502
bert-base-uncased	mean_pooling	10	title	0.170	0.2440
bert-base-uncased	mean_pooling	5	title	0.190	0.2402
bert-base-uncased	cls_token	10	title_abstract	0.185	0.2318
scibert_scivocab_uncased	mean_pooling	5	title	0.180	0.2273
bert-base-uncased	cls_token	10	abstract	0.185	0.2240
scibert_scivocab_uncased	cls_token	10	title	0.175	0.2237
scibert_scivocab_uncased	cls_token	5	title	0.175	0.2213
scibert_scivocab_uncased	mean_pooling	10	title	0.150	0.2128
bert-base-uncased	cls_token	10	title	0.140	0.2098
scibert_scivocab_uncased	cls_token	10	abstract	0.155	0.2060
scibert_scivocab_uncased	cls_token	5	abstract	0.165	0.2017
bert-base-uncased	cls_token	5	title	0.140	0.1873
scibert_scivocab_uncased	cls_token	5	title_abstract	0.130	0.1735
bert-base-uncased	cls_token	5	title_abstract	0.110	0.1690
scibert_scivocab_uncased	cls_token	10	title_abstract	0.125	0.1657
bert-base-uncased	cls_token	5	abstract	0.115	0.1617