# Appendix

**Table A1. Selection of target articles using the MeSH tree number for sP and fP.**

|  |  |  |  |
| --- | --- | --- | --- |
| Flag | MeSH Tree Number | small PubMed abstracts (sP) | Focused PubMed abstracts (fP) |
| **Included** | [C] Diseases | ✓ | ✓ |
| **Excluded** | [A13] Animal Structures | ✓ |  |
| [A16] Embryonic Structures | ✓ |  |
| [A18] Plant Structures | ✓ | ✓ |
| [A19] Fungal Structures | ✓ | ✓ |
| [A20] Bacterial Structures | ✓ | ✓ |
| [A21] Viral Structures | ✓ | ✓ |
| [B01.650] Plants | ✓ | ✓ |
| [B02] Archaea | ✓ | ✓ |
| [B03] Bacteria | ✓ |  |
| [B04] Viruses | ✓ |  |
| [B05] Organism Forms | ✓ | ✓ |
| [C22] Animal Diseases | ✓ |  |
| [D26] Pharmaceutical Preparations | ✓ |  |
| [E03] Anesthesia and Analgesia | ✓ |  |
| [E05] Investigative Techniques | ✓ | ✓ |
| [E06] Dentistry | ✓ |  |
| [E07] Equipment and Supplies | ✓ | ✓ |
| [F02] Psychological Phenomena | ✓ | ✓ |
| [F04] Behavioral Disciplines and Activities | ✓ | ✓ |
| [G] Phenomena and Processes | ✓ |  |
| [G17] Mathematical Concepts |  | ✓ |
| [H01] Natural Science Disciplines | ✓ | ✓ |
| [I] Anthropology, Education, Sociology, and Social Phenomena | ✓ | ✓ |
| [J] Technology, Industry, and Agriculture | ✓ | ✓ |
| [K] Humanities | ✓ | ✓ |
| [L] Information Science | ✓ | ✓ |
| [N] Named Groups | ✓ | ✓ |
| [Z] Geographicals | ✓ | ✓ |

**If an article has an “Included” tree number and does not have any “Excluded” ones, it is considered a target article. Furthermore, to adjust the corpus size of sP, we sampled articles published after 2010.**

**Table A2.  Performance of ouBioBERT on the BLUE task in detail.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Total | MedSTS | BIOSSES | BC5CDR-disease | BC5CDR-chemical | ShARe/CLEFE | DDI | ChemProt | i2b2 2010 | HoC | MedNLI |
| BioBERT (W + BC/ P: BaseVocab) | 82.8(0.1) | 84.9(0.5) | 89.3(1.7) | 85.7(0.4) | 93.3(0.1) | 78.0(0.8) | 80.4(0.4) | 73.3(0.4) | 74.5(0.6) | 85.8(0.6) | 82.9(0.7) |
| BlueBERT (W + BC/ P: BaseVocab) | 82.9(0.1) | 84.8(0.5) | 90.3(2.0) | 86.2(0.4) | 93.3(0.3) | 78.3(0.4) | 80.7(0.6) | 73.5(0.5) | 73.9(0.8) | **86.3**(0.7) | 82.1(0.8) |
| BlueBERT (W + BC/ P/ M: BaseVocab) | 81.6(0.5) | 84.6(0.8) | 82.0(5.1) | 84.7(0.3) | 92.3(0.1) | **79.9**(0.4) | 78.8(0.8) | 68.6(0.5) | **75.8**(0.3) | 85.0(0.4) | **83.9**(0.8) |
| ouBioBERT (prop fP + oP: AmpVocab) |  |  |  |  |  |  |  |  |  |  |  |
| *SimPT* | *AmpV* |  |  |  |  |  |  |  |  |  |  |  |
| ✓ | ✓ | **83.9**(0.2) | 84.5(0.5) | 93.1(0.5) | 87.3(0.3) | 93.9(0.1) | 79.8(0.3) | 80.7(0.7) | **75.5**(0.3) | 74.7(0.5) | 86.2(0.6) | 83.2(0.4) |
| ✓ |  | 83.9(0.3) | 85.0(0.2) | **93.3**(1.4) | **87.4**(0.2) | 93.9(0.4) | 79.4(0.4) | **81.3**(0.9) | 74.8(0.3) | 74.6(0.5) | 85.4(0.8) | 83.6(0.4) |
|  |  | 83.6(0.1) | **85.0**(0.3) | 92.6(0.6) | 87.1(0.3) | **94.1**(0.3) | 78.9(0.6) | 81.1(0.7) | 74.1(0.3) | 74.1(0.6) | 85.9(0.4) | 82.7(0.5) |

**Notes: The numbers are mean (standard deviation) on five different random seeds. The best scores are in bold, and the second best are underlined. SimPT: simultaneous pre-training; AmpV: Amplified vocabulary.**