MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction

Tanishq Gupta[1], Mohd Zaki[2], N. M. Anoop Krishnan[2,3,*], Mausam[3,4,*]

[1]*Deparment of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India 110016*

[2]*Deparment of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India 110016*

[3]*School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016*

[4]*Deparment of Computer Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India 110016*

[*]Corresponding authors: N. M. A. Krishnan (krishnan@iitd.ac.in), Mausam (mausam@iitd.ac.in)

Supplementary Information

Supplementary Discussion

Supplementary Table 1 shows the details about the number of papers and words for each family of materials in the Material Science Corpus (MSC).

Corpus details	# (Papers)	# (Words)
Inorganic glasses and ceramics	42,186	112,560,687
Bulk metallic glasses	21,093	59,338,072
Alloys	21,093	55,683,291
Cement and concrete	69,606	56,936,694
Total	153,978	284,518,744

Table 1 Numbers of papers and tokens taken for training the language model

Supplementary Table 2 represents the word count of important strings relevant to the several sub-fields in material science. It should be noted that the corpus encompasses manuscripts from several important sub-fields such as thermoelectric, nanomaterials, polymers, and biomaterials.

String	Training corpus	Validation corpus
thermoelectric	5,104	1,902
nanomaterial	5,114	873
polymer	173,738	31,234
biomaterial	5,259	887

Table 2 Frequency of words

We pre-trained MatSciBERT for 30 epochs with 2,595 training steps per epoch and a mini-batch size of 256. This is in accordance with BERT and BioBERT. BERT was pre-trained for 40 epochs and BioBERT v1.1 was trained for 1M steps with a batch size of 192 over PubMed abstracts having 4.5B words. 1M training steps with a batch size of 192 correspond to 750K steps with a batch size of 256. And since MSC is ~6.3% of the size of PubMED abstracts, we should pre-train for 750K * 0.063 = 47,250 steps. This implies training for 47250/2595 ~ 19 epochs. We chose the mid-way of BERT and BioBERT and hence train MatSciBERT for 30 epochs.

Supplementary Figure 1 shows various MLM evaluation metrics on the validation set with respect to the number of days of training for MatSciBERT. Further pretraining of MatSciBERT is not possible because we are using a triangular learning rate, wherein, the learning rate decays to 0 at the end of 15 days.

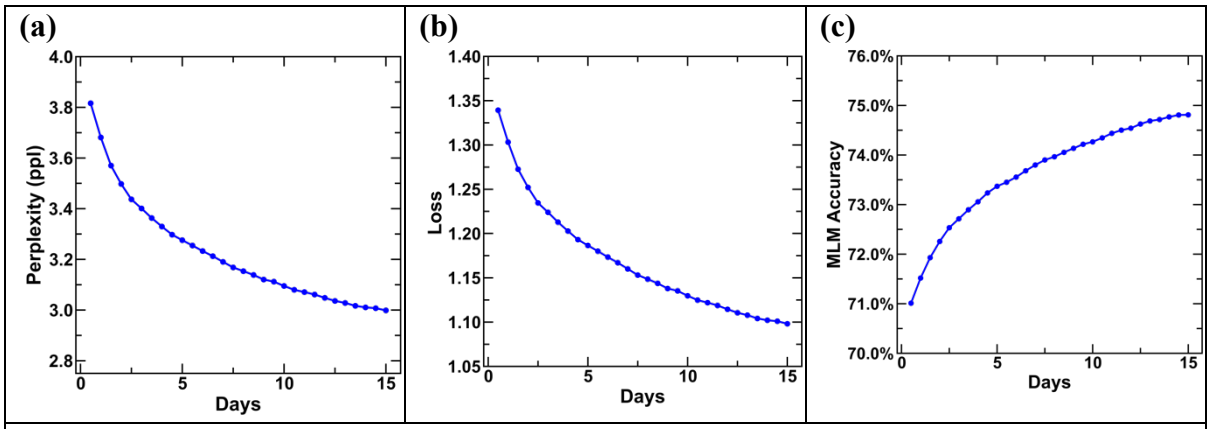

Fig. 1 Visualising MLM performance on the validation set as a function of number of days elapsed for MatSciBERT training (a) Perplexity, (b) Loss, and (c) Accuracy

Supplementary Figure 2 (a) shows the test set Macro-F1 (averaged across 3 seeds) obtained by fine-tuning on the first cross-validation split of the SOFC-Slot dataset. The increasing trend of the graph shows the importance of pre-training for longer durations. Fig 2 (b) shows the similar graph for Abstract Classification task.

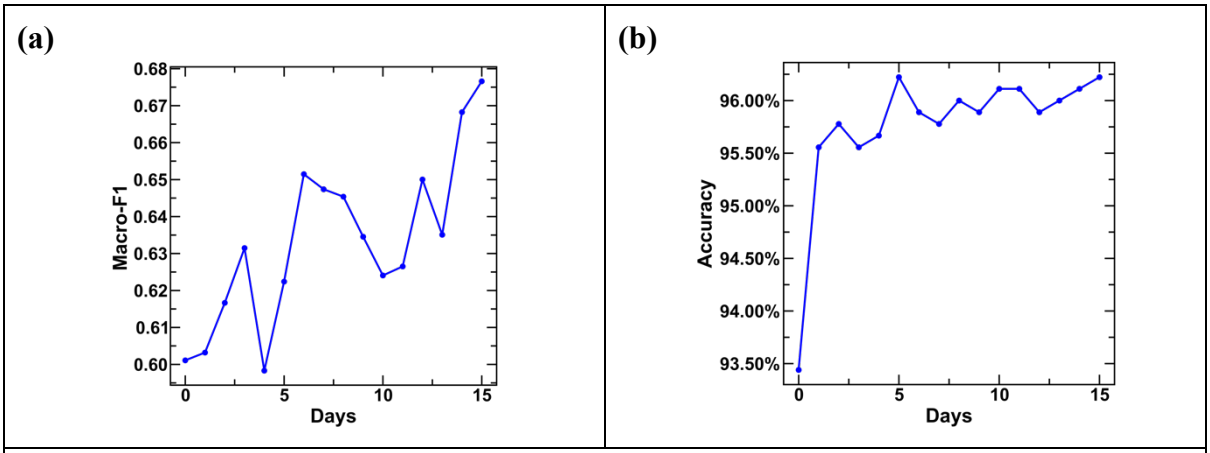

Fig. 2 (a) Macro-F1 (test set) (b) Accuracy (test set) vs number of days elapsed for MatSciBERT training for NER and Abstract Classification tasks respectively

We report the Micro-F1 scores for the SOFC-Slot and SOFC datasets in Supplementary Table 3. The research group which achieved the state-of-the-art results for Macro-F1 did not report the Micro-F1 scores, so we omit that. Supplementary Table 4 shows results for the Matscholar dataset.

Dataset	Architecture	LM = MatSciBERT	LM = SciBERT	LM = BERT
SOFC-Slot	LM - Linear	71.30 ± 1.40 (69.74 ± 4.88)	67.85 ± 1.24 (67.58 ± 3.31)	66.31 ± 2.51 (65.31 ± 5.35)
	LM-CRF	72.85 ± 1.11 (72.33 ± 3.06)	69.68 ± 2.47 (70.15 ± 4.18)	69.03 ± 1.14 (69.00 ± 5.29)
	LM-BiLSTM-CRF	72.42 ± 1.41 (72.15 ± 4.52)	70.66 ± 1.33 (70.56 ± 5.27)	67.67 ± 1.76 (68.20 ± 5.09)
SOFC	LM - Linear	84.42 ± 1.20 (82.24 ± 3.62)	81.93 ± 1.33 (81.61 ± 3.47)	78.74 ± 2.32 (80.27 ± 3.58)
	LM-CRF	84.77 ± 1.02 (83.42 ± 2.86)	83.46 ± 0.76 (83.03 ± 3.22)	81.19 ± 1.46 (81.94 ± 2.91)
	LM-BiLSTM-CRF	84.61 ± 0.90 (83.39 ± 3.39)	82.48 ± 0.90 (82.78 ± 3.10)	79.87 ± 1.65 (81.59 ± 3.14)

Table 3 Micro-F1 scores on the test set for SOFC-Slot and SOFC datasets averaged over three seeds, and five cross-validation splits. Values in the parenthesis show the results on the validation set

Architecture	LM = MatSciBERT	LM = SciBERT	LM = BERT	SOTA
LM - Linear	86.41 ± 0.21 (88.79 ± 0.47)	84.81 ± 0.35 (87.78 ± 0.12)	83.20 ± 0.28 (84.20 ± 0.31)	87.04 (87.09)
LM-CRF	87.50 ± 0.35 (89.40 ± 0.47)	86.31 ± 0.44 (88.77 ± 0.29)	85.34 ± 0.37 (85.82 ± 0.45)	
LM-BiLSTM-CRF	87.10 ± 0.22 (89.58 ± 0.42)	86.38 ± 0.25 (88.71 ± 0.28)	84.87 ± 0.13 (85.63 ± 0.17)	

Table 4 Micro-F1 scores on the test set for Matscholar averaged over three seeds. Values in the parenthesis show the results on the validation set

In Supplementary Figure 3, we show a sentence from a manuscript related to alloys. The entities are colored based on their true labels. From the predictions, it was observed that both MatSciBERT and SciBERT misclassified the phrase ‘dislocation cell-like structure’ as ‘PRO’, which was labeled as ‘O’ or ‘Other’ in the dataset. Further, ‘ND’ and ‘TD’ were labeled as ‘PRO’ by SciBERT, but MatSciBERT was able to correctly identify these entities as ‘O’. Overall, we observe that MatSciBERT is able to identify the context in the text related to materials, from which the entities are correctly tagged. The mistakes made by SciBERT are written in normal font (non-italicized) and the mistakes made by MatSciBERT are underlined.

Similarly, Supplementary Figure 4 shows the result obtained using SciBERT, where it has labeled '*YSZ'* (also called yttria-stabilized zirconia) as B-electrolyte_material that has a true label of B-interlayer_material. However, MatSciBERT was able to correctly identify the label of '*YSZ'*. Note that these are some arbitrary examples selected to demonstrate the performances of both SciBERT and MatSciBERT.

Ground truth and model predictions

By 86% cold rolling, acicular α' martensite microstructures change into extremely refined dislocation cell-like structure *with an average size of 60 nm, accompanied with the development of cold rolling texture in which the basal plane normal is tilted from the plate normal direction (*ND*) toward transverse direction (*TD*) at angles of ±49 deg. for Ti-8% V alloy and ±46 deg. for (Ti-8 mass% V)-4 mass% Sn alloy.*

Labels

Synthesis method (SMT), Symmetry/phase label (SPL), Property (PRO), Material (MAT), Descriptor (DSC)

Fig. 3 Visualizing results on the Matscholar NER dataset

Prediction of SciBERT:

While the peak power density of the cell (cell 3) with an YSZ (B-electrolyte_material) blocking layer reached approximately 35 mW / cm2, that of the single-layered GDC - based cell (cell 1) showed a much lesser power density below approximately 0.01 mW / cm2, as shown in Figure 5a, b.

Fig. 4 Visualizing prediction on SOFC slot dataset

Supplementary Tables 5 and 6 compare the topic modelling done using LDA and MatSciBERT. The topics identified using MatSciBERT are more coherent as compared to the ones obtained using LDA.

Topic No.	Keywords
0	['cells', 'protein', 'cell', 'activity', 'binding', 'structure', 'proteins', 'dna', 'human', 'enzyme']
1	['water', 'properties', 'concentration', 'ph', 'content', 'showed', 'results', 'temperature', 'polymer', 'different']
2	['glass', 'composite', 'fiber', 'fibers', 'properties', 'mechanical', 'materials', 'material', 'strength', 'results']
3	['based', 'use', 'new', 'paper', 'structure', 'methods', 'study', 'research', 'data', 'systems']
4	['temperature', 'phase', 'glass', 'transition', 'magnetic', 'fe', 'structure', 'properties', 'dielectric', 'samples']
5	['soil', 'study', 'patients', 'treatment', 'species', 'group', 'results', 'fluoride', 'significantly', 'rare']
6	['particles', 'size', 'particle', 'nanoparticles', 'ag', 'powder', 'surface', 'nm', 'silver', 'diameter’]
7	['mantle', 'ma', 'pb', 'high', 'sr', 'rocks', 'ree', 'ratios', 'samples', 'composition']
8	['model', 'results', 'data', 'experimental', 'flow', 'method', 'structure', 'parameters', 'time', 'transition']
9	['la', 'ce', 'et', 'da', 'ra', 'des', 'en', 'ta', 'sa', 'une']

Table 5 Top ten topics obtained based on LDA along with the top 10 keywords associated with each topic

Topic No.	Keywords
0	['materials', 'used', 'radiation', 'laser', 'model', 'process', 'paper', 'thermal', 'use', 'based', 'chapter', 'solar', 'experimental', 'applications', 'energy']
1	['dielectric', 'ceramics', 'conductivity', 'sintering', 'doping', 'phase', 'ceramic', 'sintered', 'constant', 'doped', 'grain', 'solid', 'frequency', 'electrical', 'ferroelectric']
2	['tio2', 'photocatalytic', 'light', 'visible', 'zno', 'nanoparticles', 'activity', 'efficiency', 'polymer', 'uvÃ¢', 'synthesized', 'vis', 'doped', 'film', 'crystal']
3	['laser', 'silica', 'surface', 'ion', 'radiation', 'measured', 'time', 'energy', 'induced', 'diffusion', 'irradiation', 'index', 'observed', 'used', 'glasses']
4	['complexes', 'ii', 'complex', 'ligand', 'compounds', 'fluorescence', 'synthesized', 'nmr', 'ligands', '1h', 'compound', 'crystal', 'iii', 'ir', 'h2o']
5	['luminescence', 'crystals', 'doped', 'ions', 'conductivity', 'glasses', 'excitation', 'band', 'observed', 'absorption', 'centers', 'emission', 'energy', 'crystal', 'laser']
6	['zno', 'films', 'deposited', 'film', 'gap', 'substrates', 'substrate', 'electrical', 'band', 'deposition', 'xrd', 'annealing', 'doping', 'resistivity', 'optical']
7	['adsorption', 'ph', 'fluorescence', 'protein', 'fluoride', 'water', 'phosphate', 'ii', 'binding', 'removal', 'detection', 'acid', 'concentration', 'solution', 'ions']
8	['films', 'film', 'deposition', 'deposited', 'substrates', 'si', 'substrate', 'annealing', 'optical', 'zno', 'band', 'silicon', 'thickness', 'gap', 'sputtering']
9	['waste', 'melt', 'melts', 'water', 'silicate', 'alteration', 'composition', 'data', 'compositions', 'mantle', 'dissolution', 'ash', 'elements', 'rich', 'co2']

Table 6 Top ten topics obtained based on MatSciBERT along with the top 10 keywords associated with each topic

Supplementary Note

Note that the number of words in Supplementary Table 1 also includes mathematical equations and chemical formulas. To calculate the number of words in the corpus, we split it by space, tab, and newline characters. We observe that chemical compound names like p-dimethyl-aminophenyldimethylphosphine—zinc(II) or 1,7-Dibromo-N,N'-(bicyclohexyl)-3,4:9,10-perylendiimide, and chemical formulas like [(CH3CH2)2N(CH3)2]2Ag13I15 or 20K2CO3+5Na2CO3+10LiF count as 1 word each. On the other hand, equations like n + A → (p + π -) + A count as 10 words. As such, the number of words reported in Table 1 presents a reasonable estimate of the actual number of words in our corpus.