APPENDIX: ScilitBERT a pre-trained model for academic domain language understanding

Jean-Baptiste de la Broise
MDPI
Basel, Switzerland
jeanbaptiste.delabroise@mdpi.com

Nolwenn Bernard
MDPI
Basel, Switzerland
bernard@mdpi.com

Jean-Philippe Dubuc
MDPI
Basel, Switzerland
dubuc@mdpi.com

Andrea Perlato
MDPI
Basel, Switzerland
andrea.perlato@mdpi.com

Bastien Latard
MDPI
Basel, Switzerland
latard@mdpi.com

A. fine-tuning hyper-parameters

Selected hyper-parameters are generally recommended in a fine-tuning sequence classification task. There is no hyper-parameters tuning here.

TABLE I
HYPER-PARAMETERS FOR FINE-TUNING ON JOURNAL FINDER TASK

h-param	ScilitBERT	RoBERTa	SciBERT
learning rate	3e-5		
beta 1	0.9		
beta 2	0.999		
batch size	16		10[a]
epochs	3		
weight decay	0.01		
fp16	True		

[a] A batch contains less samples for SciBERT as each input weighs more due to BERT architecture requiring an input_type_ids field, which field is not required when training a model based on RoBERTa.

TABLE II
HYPER-PARAMETERS FOR FINE-TUNING ON THE WEB OF SCIENCE TASK

h-param	value
learning rate	3e-5
beta 1	0.9
beta 2	0.999
batch size	10
epochs	3
weight decay	0.01
fp16	True

B. Categories

The categories used to represent the distribution of pre-training data are detailed in Table IV.

TABLE III
LIST OF ABBREVIATIONS FOUND IN TABLE IV AND PAPER FIG. 2

Original	Abbreviated
AGRICULTURE	AGRI.
APPLICATIONS	APPs.
APPLIED.	APPd.
ARCHITECTURE	ARCHITECT.
AUSTRALIAN	AU.
BIODIVERSITY.	BIODIV.
BIOLOGY	BIO.
CANADIAN	CA.
CARDIOVASCULAR	CARDIOVASC.
CHEMISTRY	CHEM.
COMPLEMENTARY	COMP.
COMPUTER SCIENCE	C.S.
DISEASES	DISs.
EDUCATION	EDU.
ENGINEERING	ENG.
EXPERIMENTAL	EXP.
GASTROENTEROLOGY	GASTROENT.
HARDWARE	HW.
HEPATOLOGY	HEPATO.
INSTRUMENTS	INSTRU.
LABORATORY	LABO.
MATHEMATICAL	MATH.
MEDICINE	MED.
PHILOSOPHY	PHILO.
PSYCHOLOGY	PSYCHO.
SCIENCE	SCI.
SUSTAINABLE	SUSTAIN.
SYSTEM	SYST.
SYSTEMS	SYSs.
TECHNOLOGY	TECH.

TABLE IV

CATEGORIES FOR SOME WEB OF SCIENCE CLASSES. ABBREVIATIONS MEANING CAN BE FOUND IN TABLE III

MEDICAL	MEDICAL ETHICS	OPHTHALMOLOGY	OTORHINOLARYNGOLOGY	DERMATOLOGY	IMMUNOLOGY	TROPICAL MEDICINE
ENGINEERING	INSTRUMENTATION	NANOSCI. NANOTECHs.	CELL TISSUE ENGINEERING	ENG., AEROSPACE	ENG., BIOMEDICAL	ENGINEERING, CHEMICAL
BIOLOGY	PLANT SCIENCES	MYCOLOGY	CARDIAC CARDIOVASC. SYST.	REPRODUCTIVE BIOLOGY	MICROBIOLOGY	MARINE FRESHWATER BIO.
PHYSICS	ASTRONOMY ASTROPHYSICS	ACOUSTICS	THERMODYNAMICS	MECHANICS	NUCLEAR SCI. TECH.	PHYSICS, APPd.
LITERATURE	LITERATURE	LITERARY THEORY CRITICISM	LITERARY REVIEWS	CLASSICS	POETRY	LITERATURE, AFRICAN, AU., CA.
DISEASES	ONCOLOGY	PATHOLOGY	ALLERGY	RHEUMATOLOGY	INFECTIOUS DISs.	PERIPHERAL VASCULAR DIS.
CHEMISTRY	TOXICOLOGY	POLYMER SCIENCE	ELECTROCHEMISTRY	CHEMISTRY, ORGANIC	CHEMISTRY, PHYSICAL	CHEMISTRY, ANALYTICAL
BODY	OBSTETRICS GYNECOLOGY	SPORT SCIENCES	RESPIRATORY SYSTEM	ANDROLOGY	ANATOMY MORPHOLOGY	PHYSIOLOGY
PSYCHOLOGY	PSYCHOLOGY	PSYCHIATRY	PSYCHO., EXPERIMENTAL	PSYCHO., APPd.	PSYCHO., MATHEMATICAL	PSYCHO., BIOLOGICAL
ENVIRONMENT	SUSTAINABLE SCIENCE TECH.	FISHERIES	FORESTRY	ENVIRONMENTAL STDs.	ENVIRONMENTAL SCI.	ECOLOGY
MATERIALS	BIOMATERIALS	CERAMICS	CHARACTERIZATION TEST.	COATINGS FILMS	COMPOSITES	PAPER WOOD
COMPUTER SCIENCE	CYBERNETICS	TELECOMMS	HARDWARE ARCHITECTURE	INTERDISCIPLINARY APPs.	SOFTWARE ENG.	THEORY METHODS
VISUALIZATION	OPTICS	NEUROIMAGING	MICROSCOPY	SPECTROSCOPY	RADIOLOGY, MED. IMG.	PHOTOGRAPHIC TECH.
FOOD	FOOD SCIENCE TECHNOLOGY	HORTICULTURE	AGRICULTURAL ECO. POLICY	AGRONOMY	AGRICULTURAL ENG.	AGRI., DAIRY ANIMAL SCI.
ZOOLOGY	ZOOLOGY	ORNITHOLOGY	PALEONTOLOGY	VETERINARY SCIENCES	ENTOMOLOGY	
SOCIOLOGY	SOCIOLOGY	FAMILY STUDIES	WOMEN STUDIES	PSYCHOLOGY, SOCIAL	PUBLIC ADMINISTRATION	
MATHEMATICS	MATHEMATICS	LOGIC	STATISTICS PROBABILITY	MATHEMATICS APPd.	INTERDISCIPLINARY APPs.	
SOCIAL SCIENCES	SOCIAL SCIENCES, BIOMEDICAL	INTERDISCIPLINARY	SOC., MATH. METHODS	SOCIAL WORK	HISTORY OF SOCIAL SCI.	
GEOLOGY	GEOLOGY	SOIL SCIENCE	GEOCHEMISTRY GEOPHYSICS	CRYSTALLOGRAPHY	MINERALOGY	
ECONOMY	BUSINESS, FINANCE	DEVELOPMENT STDs.	ECONOMICS	HEALTH POLICY SERVS.	BUSINESS	
HEALTH	NUTRITION DIETETICS	SUBSTANCE USE DISORDER	PRIMARY HEALTH CARE	PUBLIC, ENV. HEALTH	HEALTH CARE SCI. SERV.	
HISTORY	HISTORY	ARCHAEOLOGY	MEDIEVAL RENAISSANCE STDs.	HISTORY PHILO. OF SCI.		
CULTURES	HOSPITALITY, LEISURE, TOURISM	ASIAN STUDIES	CULTURAL STUDIES	FOLKLORE		
ETHNOLOGY	ETHNIC STUDIES	URBAN STUDIES	RELIGION	ANTHROPOLOGY		
EDUCATION	EDUCATION, SPECIAL	EDUCATIONAL SEARCH	PSYCHOLOGY, EDUCATIONAL	EDU., SCI. DISCIPLINES		
ART	ART	DANCE	FILM, RADIO, TELEVISION	THEATER		
INDUSTRY	MINING MINERAL PROCESSING	ENERGY FUELS	METALLURGICAL ENG.	ENG., INDUSTRIAL		
GEOGRAPHY	GEOGRAPHY	GEOGRAPHY, PHYSICAL	AREA STUDIES	DEMOGRAPHY		
WATER	LIMNOLOGY	OCEANOGRAPHY	WATER RESOURCES			
MANAGEMENT	MANAGEMENT SCIENCE	MANAGEMENT	INDUS. RELATIONS LABOR			
ARCHITECTURE	ARCHITECTURE	CONSTRUCTION TECHs.	REGIONAL URBAN PLANNING			
ELOQUENCE	COMMUNICATION	POLITICAL SCIENCE	INTERNATIONAL RELATIONS			
COGNITION	BEHAVIORAL SCIENCES	MUSIC	AUDIOLOGY SPEECH-LANG.			
DATA SCIENCES	INFORMATION AND LIBRARY SCI.	ARTIFICIAL INTEL.	INFORMATION SYSTEMS			
LINGUISTICS	LINGUISTICS	LANGUAGE LINGUISTICS				
GERONTOLOGY	GERONTOLOGY	GERIATRICS GERONTO.				
TRANSPORTATION	TRANSPORTATION	TRANSPT. SCI. TECHs.				
AUTOMATION	ROBOTICS	AUTO. CONTROL SYSTs.				
BRAIN	CLINICAL NEUROLOGY	NEUROSCIENCES				
PHILOSOPHY	PHILOSOPHY	ETHICS				
EVOLUTION	GENETICS HEREDITY	EVOLUTIONARY BIO.				
CRIMINOLOGY	CRIMINOLOGY PENOLOGY					
ERGONOMICS	ERGONOMICS					
LAW	LAW					
SENSORS	REMOTE SENSING					