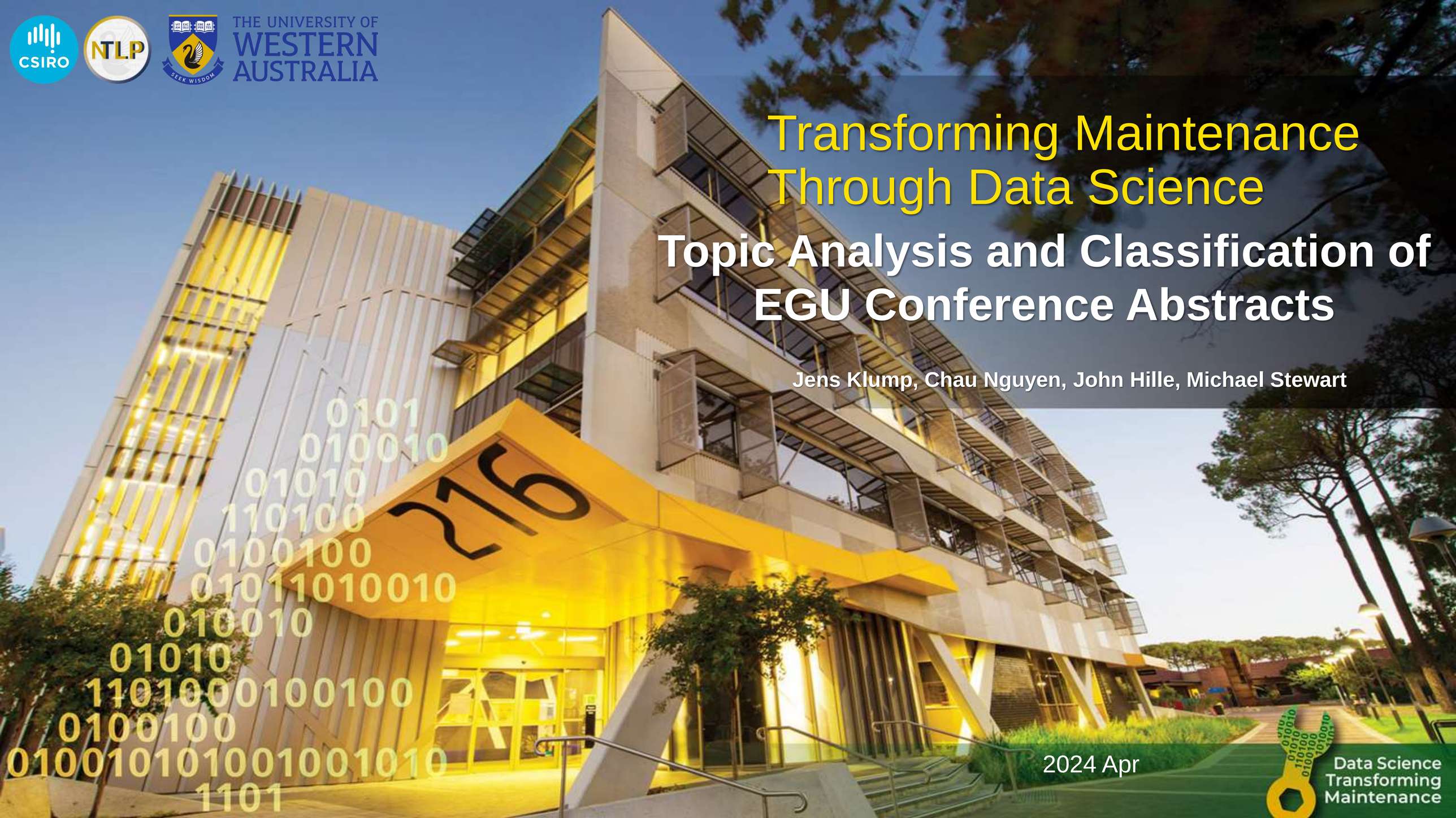

CSIRO
NTLP
THE UNIVERSITY OF WESTERN AUSTRALIA
SEEK WISDOM
Transforming Maintenance Through Data Science
Topic Analysis and Classification of EGU Conference Abstracts
Jens Klump, Chau Nguyen, John Hille, Michael Stewart
2024 Apr
Data Science Transforming Maintenance

Supported by Open Source

Python

- Abstract extraction and and matching
- General data processing
- BSD-Zero Clause

Selenium

- EGU Session/Topic extraction
- Apache 2.0

HuggingFace

- BERT/ DistilBERT
- Transformer model for prediction
- Apache 2.0

Jupyter

- Analysis and prototyping
- BSD-3 Clause

Summary

- Large corpus (2004 to 2023) of permissively licenced Geoscience Abstracts available (EGU Conference)
- Abstracts are readily available but classifying the abstracts to topics using EGU session information required considerable effort
- With topics assigned, supervised learning could be used to build an ML model that can assign the most relevant geoscience topic of a geoscience abstract
- ML model is significantly better than random model in classifying abstracts
- Accuracy of abstract to topic matching is generally successful
- Training data will be released publicly and codebase will be made open source

Data Modelling

Data Collection

Abstracts Extraction

- PDF to Text and XML source files

Topics Extraction

- EGU Websites with Selenium Web Driver

Abstracts to Topics Matching

- Matched on EGU identifiers

Data Modelling

Training

- Supervised learning
- BERT and DistilBERT

Evaluation

- Accuracy of single true label vs multiple true label

Geoscience Abstract Topic Classification Model

Navigation

1. Abstract Extraction

2. Topic Extraction

3. Abstract To Topic Matching

4. Training and Evaluation

5. Model Accuracy

BERT vs DistilBERT?

6. Potential Applications

Data Collection – Abstracts extraction – Stats

Year	Original Data	Converted Format
2015 - 2023	XML	JSON
2000 - 2014	PDFs	JSON

Year	Num Abstracts	Topics
2015	14K	N/A
2014	15K	N/A
2013	13K	N/A
…	…	…
2004	7K	N/A

Unavailable Topics

Data Collection – PDF Abstracts extraction

Others

Geophysical Research Abstracts
Vol. 17, EGU2015-150, 2015
EGU General Assembly 2015

Title

Root and microbial respiration from urban, agricultural and natural soils within the Moscow megapolis

Authors

Viacheslav Vasenev (1,2), Simona Castaldi (3), Marya Vizirskaya (1), Nadezhda Ananyeva (4), Kristina Ivashchenko (4), Riccardo Valentini (), and Ivan Vasenev ()

Affiliations

(1) Russian State Agrarian University, Moscow, Russian Federation (vasenyov@mail.ru), (2) Peoples' Friendship University of Russia, (3) Seconda Università di Napoli, (4) Institute of Physico-chemical and Biological Problems in Soil Science

Abstract

Urbanization is an important process of land-use change, which is increasing with the growth of population and abandonment of rural areas. Urbanization alters profoundly soil features and functions, among which soil respiration, which is one of the main carbon fluxes to the atmosphere. Soil respiration is the result of heterotrophic and autotrophic components, which are driven by biotic and abiotic factors. Little is known about soil respiration and its components in urban environments, which represent highly variable systems, characterized by different functional zones, types and intensities of urban management. In the present study we analyzed the spatial variability and temporal dynamics of total soil respiration (Rs) and its components, autotrophic (Ra) and heterotrophic respiration (Rh), from soils of different environments included in the Moscow megalopolis area. In particular we compared highly impacted areas urban green lawns with less anthropized ecosystems within the Moscow city: arable lands and urban forest sites. Experiments were set after snow melt and respiration fluxes were analyzed during the whole summer period till the beginning of the autumn. Data showed that Rs was significantly higher in the most disturbed sites, the green lawns, and showed the highest variability among the three analyzed land use types. Rh was the dominant component of soil respiration in all sites and did not vary significantly during the study period. However, significant differences was shown for the metabolic quotient qCO_2, estimated as heterotrophic respiration ratio to microbial carbon (Rh/Cmic). The most disturbed sites showed the highest qCO_2 within the lawn land use, followed by arable sites and forest sites, characterized by the lowest qCO_2. Ra contributed to total Rs only at a minor extent (26%) and increased in all study sites along the season following the phenological cycle of the plant communities. Ra absolute values and relative contribution to Rs did not change significantly among land use types. Overall, the high observed fluxes of CO_2 in urban lawns seemed to be driven by land management and disturbance impact on the microbial community

TEXT

JSON

Sharing is encouraged

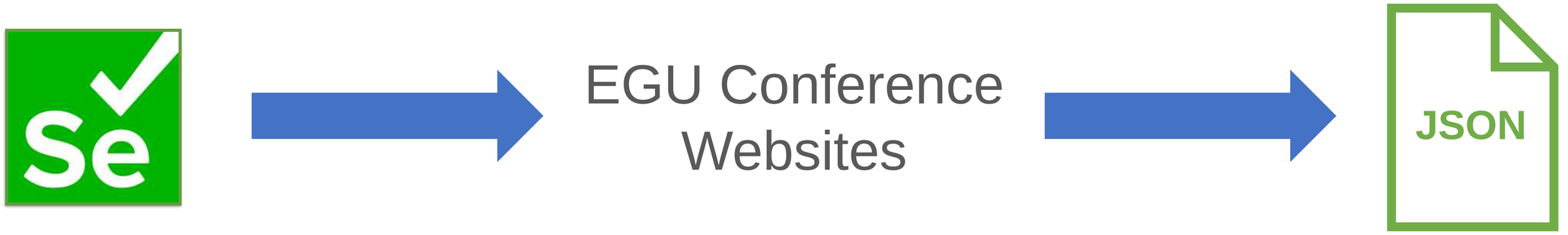
CSIRO NTLP THE UNIVERSITY OF WESTERN AUSTRALIA

Data Science Transforming Maintenance

Data Collection – XML Abstracts extraction

Data Collection –Topics extraction – Stats

Web Scraping with Selenium



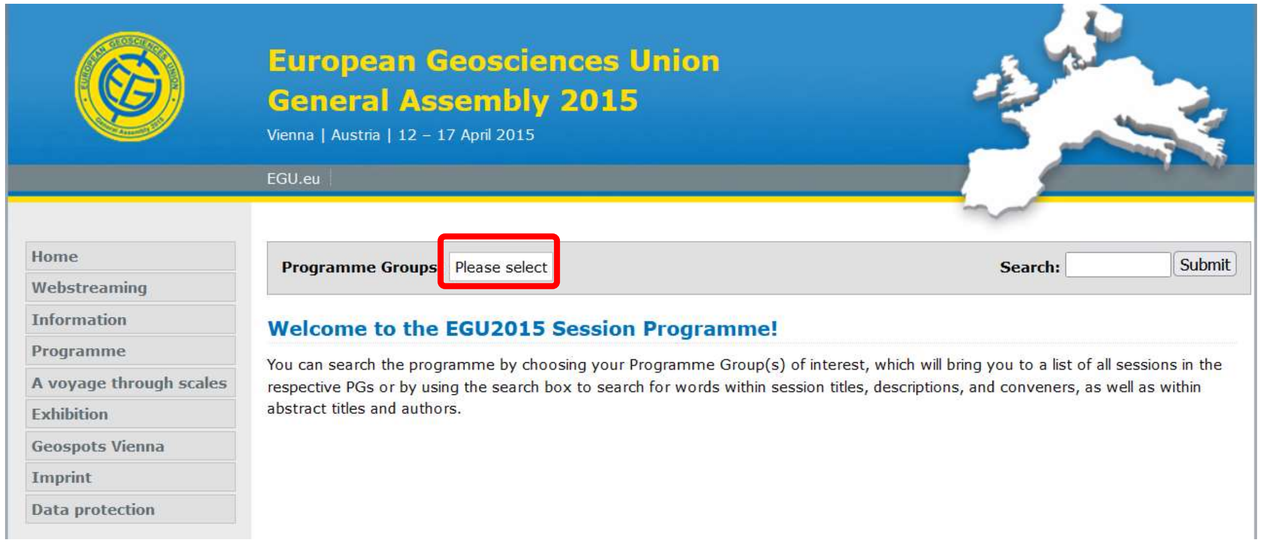
Year	Num Topics	EGU Conference Links
2015	22	https://meetingorganizer.copernicus.org/egu2015/sessionprogramme
2014	23	https://meetingorganizer.copernicus.org/egu2014/sessionprogramme
2013	23	https://meetingorganizer.copernicus.org/egu2013/sessionprogramme

Available Topics

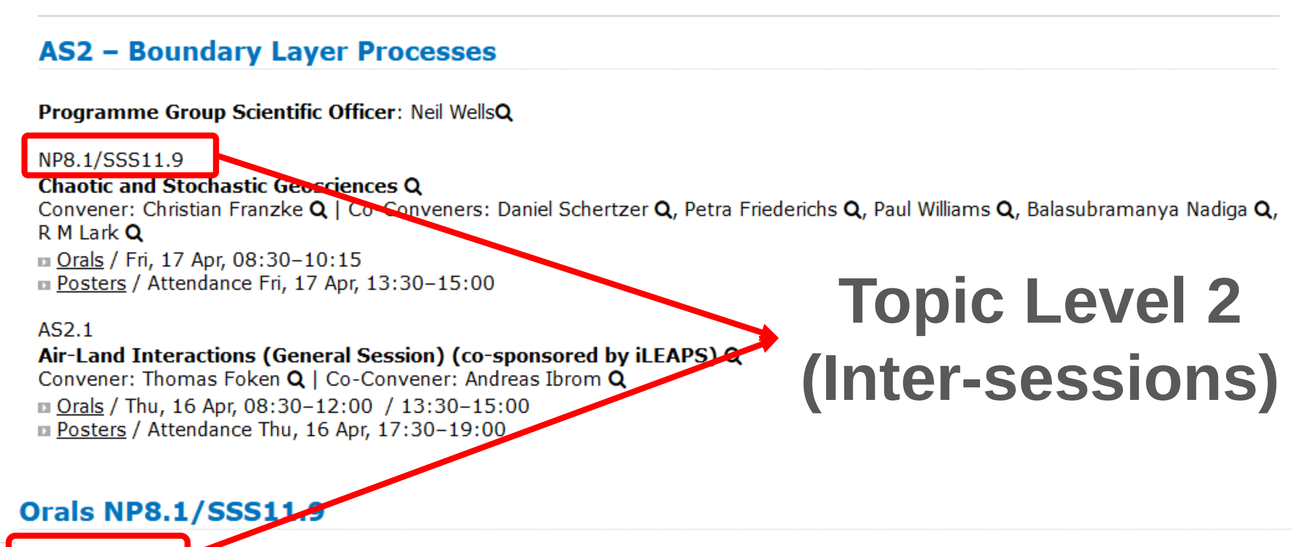
- 2015: exclude IS
- 2013 - 2014: exclude IS and PSD

Data Collection – Topics extraction - Example

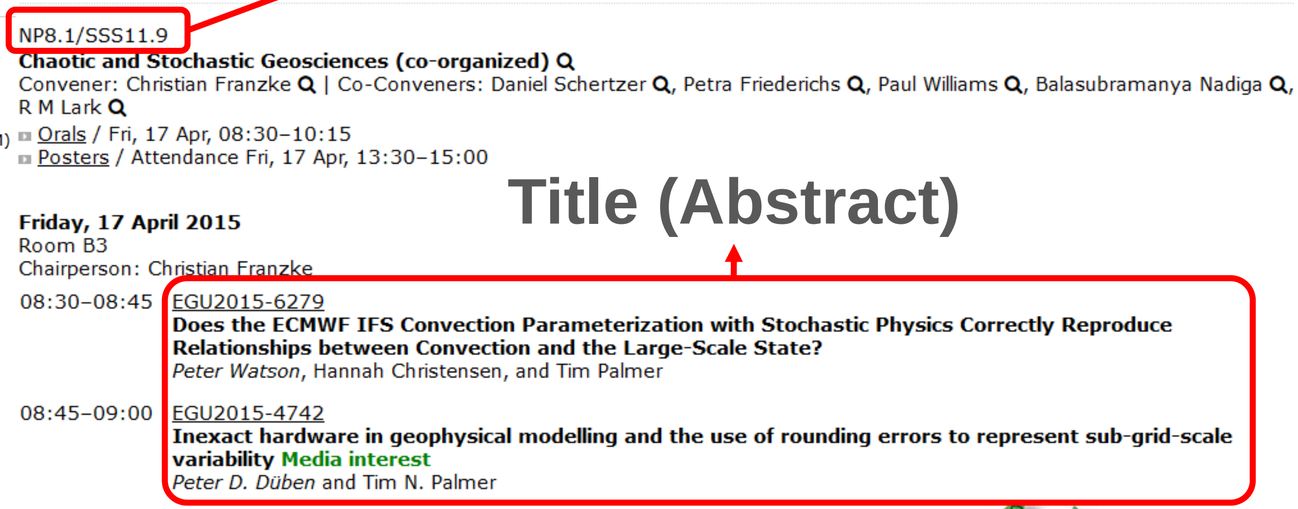
Geoscience topics available from EGU conference program



All Primary Topics



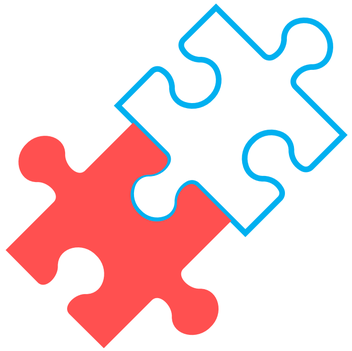
Topic Level 2 (Inter-sessions)



Title (Abstract)

Data Science Transforming Maintenance

Data Collection – Abstract/Topics Matching

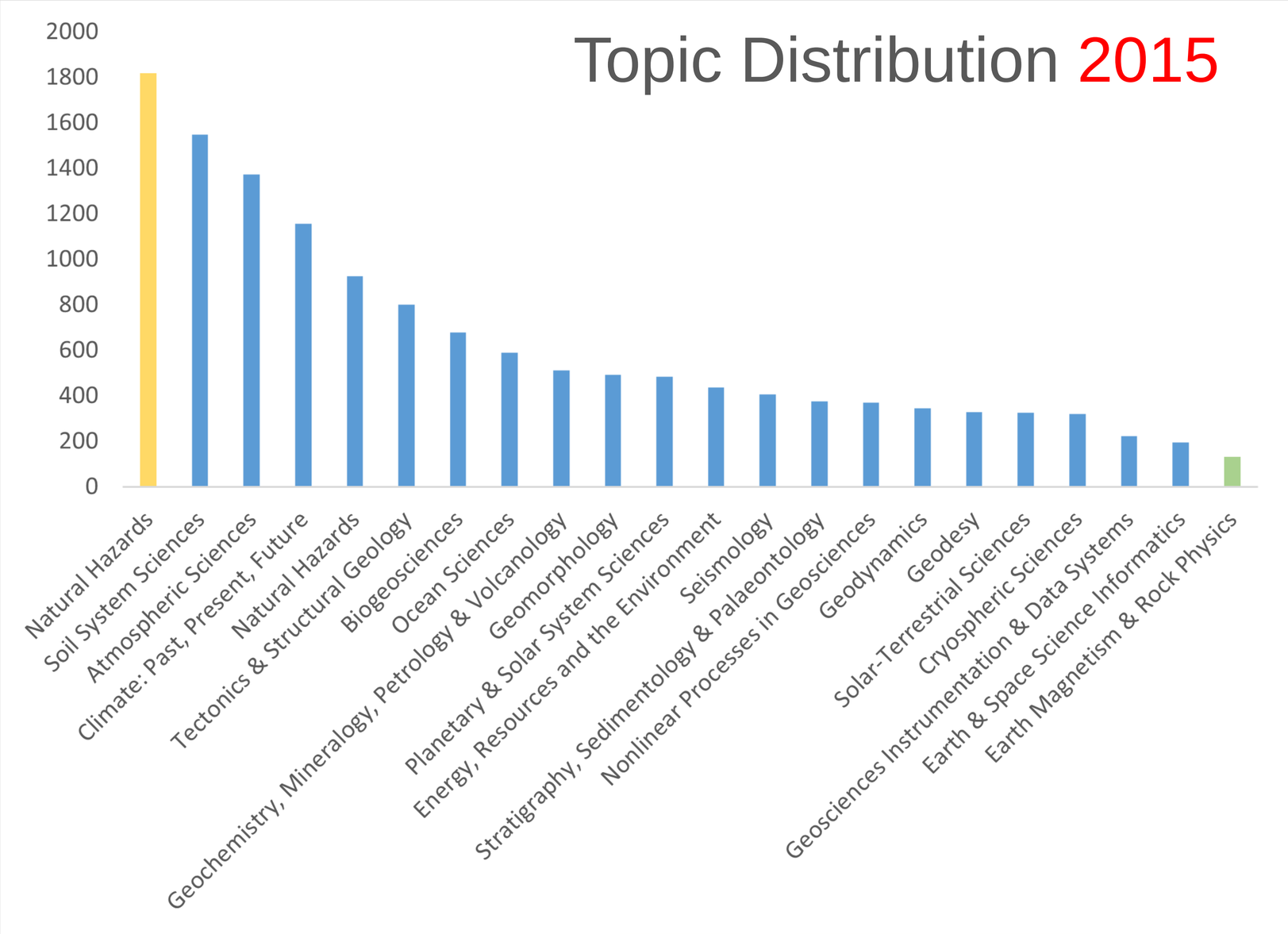
Abstracts  **Topics**

ID: EGU-****

Year	Abstracts	Primary Topics
2015	14K	22
2014	15K	23
2013	13K	23

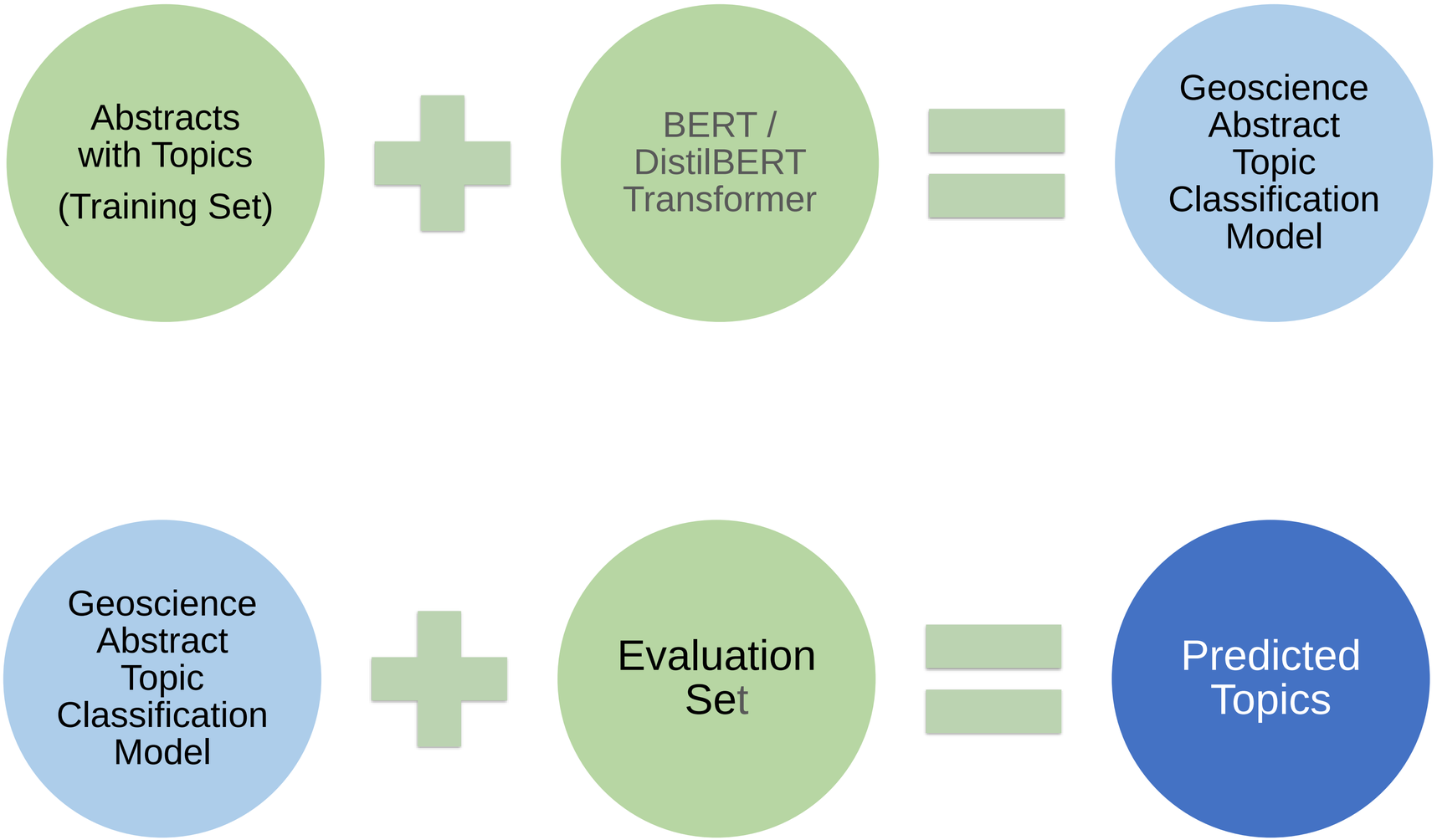
ID	Geoscience Topics 2015
AS	Atmospheric Sciences
BG	Biogeosciences
CL	Climate: Past, Present, Future
CR	Cryospheric Sciences
EMRP	Earth Magnetism & Rock Physics
ERE	Energy, Resources and the Environment
ESSI	Earth & Space Science Informatics
G	Geodesy
GD	Geodynamics
GI	Geosciences Instrumentation & Data Systems
GM	Geomorphology
GMPV	Geochemistry, Mineralogy, Petrology & Volcanology
HS	Hydrological Sciences
NH	Natural Hazards
NP	Nonlinear Processes in Geosciences
OS	Ocean Sciences
PS	Planetary & Solar System Sciences
SM	Seismology
SSP	Stratigraphy, Sedimentology & Palaeontology
SSS	Soil System Sciences
ST	Solar-Terrestrial Sciences
TS	Tectonics & Structural Geology

Data Preprocessing – Datasets Stats



ID	Geoscience Topics
AS	Atmospheric Sciences
BG	Biogeosciences
CL	Climate: Past, Present, Future
CR	Cryospheric Sciences
EMRP	Earth Magnetism & Rock Physics
ERE	Energy, Resources and the Environment
ESSI	Earth & Space Science Informatics
G	Geodesy
GD	Geodynamics
GI	Geosciences Instrumentation & Data Systems
GM	Geomorphology
GMPV	Geochemistry, Mineralogy, Petrology & Volcanology
HS	Hydrological Sciences
NH	Natural Hazards
NP	Nonlinear Processes in Geosciences
OS	Ocean Sciences
PS	Planetary & Solar System Sciences
SM	Seismology
SSP	Stratigraphy, Sedimentology & Palaeontology
SSS	Soil System Sciences
ST	Solar-Terrestrial Sciences
TS	Tectonics & Structural Geology

Model Training



Evaluation Results

For Single True Label:

- Exact match is required between predicted label and the FIRST listed session e.g. NP/SSS
- Only 'NP' is considered an accurate classification

For Multiple True Label:

- Exact match of predicted label and ANY listed session e.g. NP/SSS
- Either 'NP' or 'SSS' is an accurate classification

This recognises that the first session listed may not always be the most relevant

Training and Evaluation - Accuracy results

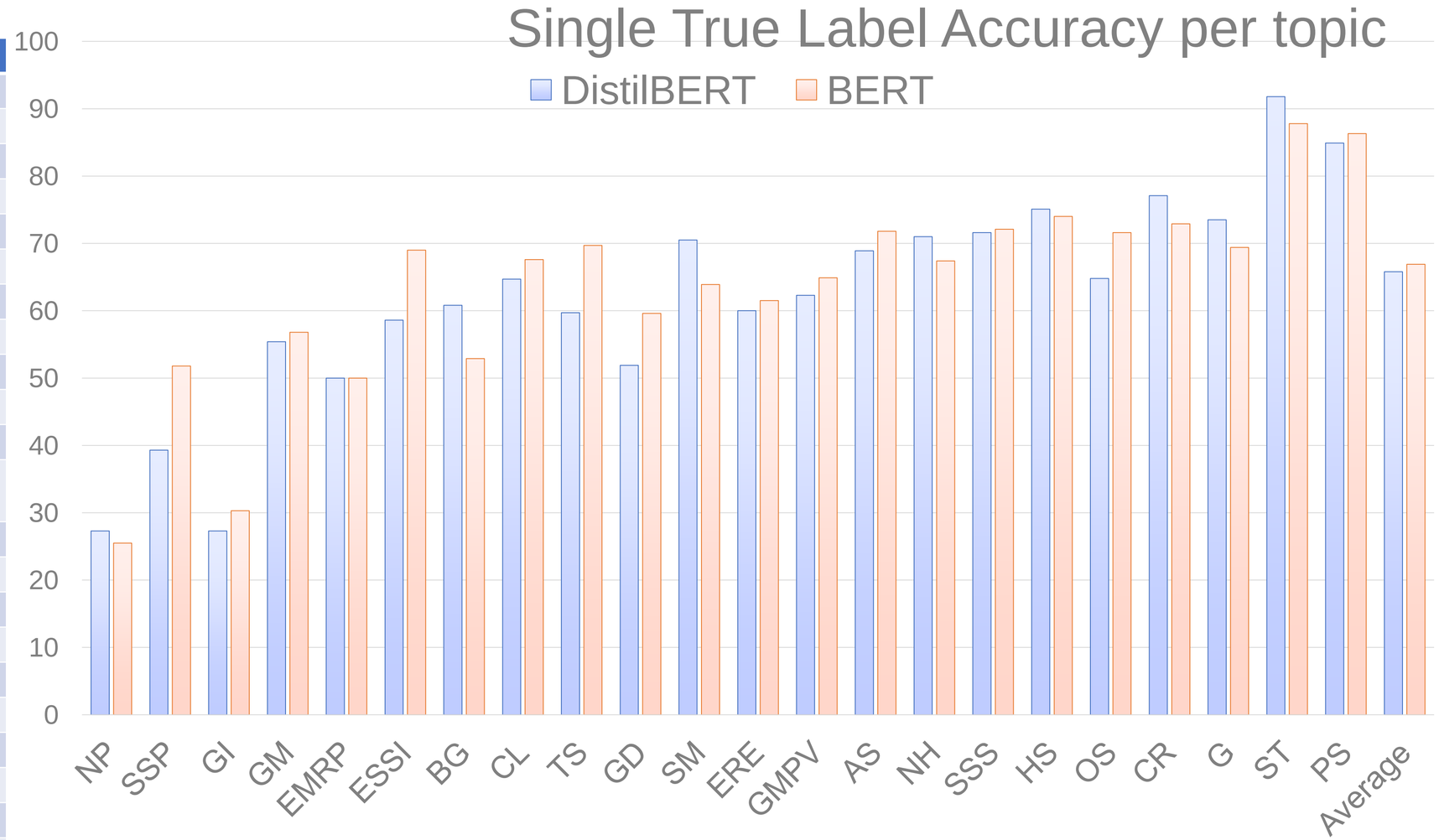
Classification Performance as percentage (%)

Year	Model	F1 (single true label)	Accuracy (single true label)	Accuracy (multiple true label)
2015	DistilBERT [1]	65.4	65.8	73.9
	BERT [2]	66.7	66.9	75.2

[1] Sanh et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://arxiv.org/pdf/1910.01108.pdf
[2] Devlin et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

Single True Label – Accuracy per Topic

ID	Geoscience Topics
AS	Atmospheric Sciences
BG	Biogeosciences
CL	Climate: Past, Present, Future
CR	Cryospheric Sciences
EMRP	Earth Magnetism & Rock Physics
ERE	Energy, Resources and the Environment
ESSI	Earth & Space Science Informatics
G	Geodesy
GD	Geodynamics
GI	Geosciences Instrumentation & Data Systems
GM	Geomorphology
GMPV	Geochemistry, Mineralogy, Petrology & Volcanology
HS	Hydrological Sciences
NH	Natural Hazards
NP	Nonlinear Processes in Geosciences
OS	Ocean Sciences
PS	Planetary & Solar System Sciences
SM	Seismology
SSP	Stratigraphy, Sedimentology & Palaeontology
SSS	Soil System Sciences
ST	Solar-Terrestrial Sciences
TS	Tectonics & Structural Geology

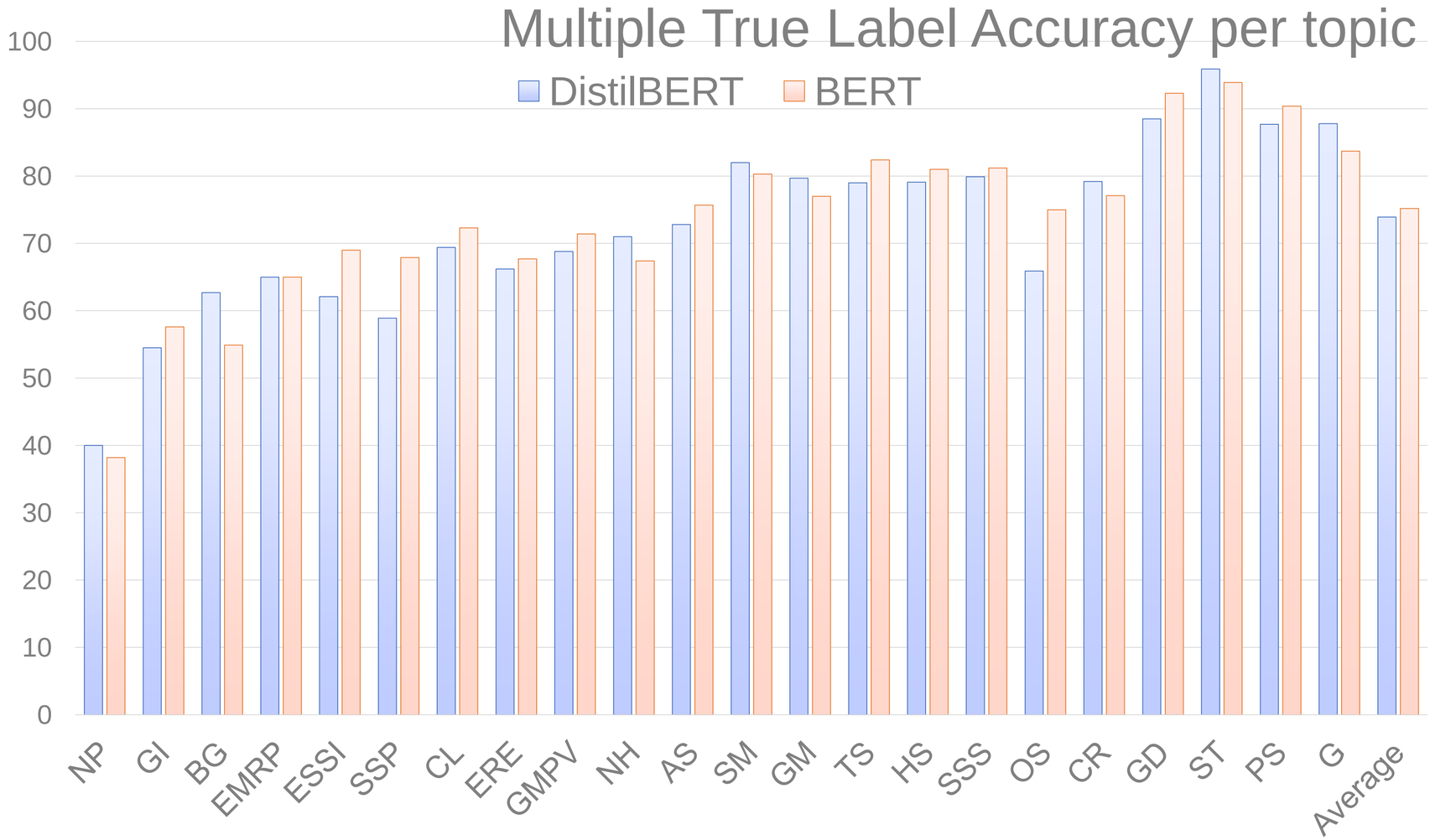


[1] Sanh et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://arxiv.org/pdf/1910.01108.pdf

[2] Devlin et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

Multiple True Label Accuracy per Topic

ID	Geoscience Topics
AS	Atmospheric Sciences
BG	Biogeosciences
CL	Climate: Past, Present, Future
CR	Cryospheric Sciences
EMRP	Earth Magnetism & Rock Physics
ERE	Energy, Resources and the Environment
ESSI	Earth & Space Science Informatics
G	Geodesy
GD	Geodynamics
GI	Geosciences Instrumentation & Data Systems
GM	Geomorphology
GMPV	Geochemistry, Mineralogy, Petrology & Volcanology
HS	Hydrological Sciences
NH	Natural Hazards
NP	Nonlinear Processes in Geosciences
OS	Ocean Sciences
PS	Planetary & Solar System Sciences
SM	Seismology
SSP	Stratigraphy, Sedimentology & Palaeontology
SSS	Soil System Sciences
ST	Solar-Terrestrial Sciences
TS	Tectonics & Structural Geology



[1] Sanh et al. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://arxiv.org/pdf/1910.01108.pdf

[2] Devlin et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

Potential Applications

Trained ML Model:

- Session recommender system for attendees
- Identifying abstracts for potential cross listing
- Assisted "binning" of abstracts into relevant sessions
- Vetting of abstracts to ensure they are relevant to sessions
- Abstract reviewer system
- Topic Distribution Analysis
- Cross listing impacts and benefits

Public Dataset Release:

- LLM Training
- Named Entity Recognition
- Knowledge Graph Extraction
- What are your ideas?

Data Science Transforming Maintenance

Acknowledgments

ARC Training Centre for Transforming Maintenance through Data Science

Jens Klump, Chau Nguyen, John Hille, Michael Stewart

CSIRO

NTLP

THE UNIVERSITY OF WESTERN AUSTRALIA

SEEK WISDOM