



Semi-automatic literature reviews

Sam Dornan

What exactly is a semi-automatic literature review?

Using machine learning to infer the relationship between words in a corpus

- Can review larger volumes of text
- Find relationships between disparate disciplines and fields of study
- Already used across material sciences
- Can it be applied to the social sciences?

Unsupervised word embeddings capture latent knowledge from materials science literature

[Vahe Tshitoyan](#) ✉, [John Dagdelen](#), [Leigh Weston](#), [Alexander Dunn](#), [Ziqin Rong](#), [Olga Kononova](#), [Kristin A. Persson](#), [Gerbrand Ceder](#) ✉ & [Anubhav Jain](#) ✉

[Nature](#) **571**, 95–98 (2019) | [Cite this article](#)

69k Accesses | **381** Citations | **1899** Altmetric | [Metrics](#)

How would it be done?

Social sciences is a part of the world as much as the world is a part of the social sciences



Large language models such as GPT and T-5 can “understand” the world.



Open-source programmes allow models to be tailored to a specific task.



Tailor a language model to the social sciences

The dataset

Easy to access

Easy to work with

Small-fits in memory

Tailored

Wisdom of the crowds

The screenshot displays the Web of Science interface. At the top, the Clarivate logo is on the left, and 'English' and 'Products' are on the right. The main header shows 'Web of Science' and a search bar. Below the search bar, the results are for 'Sociology (Topic)' with 170 results from the Web of Science Core Collection. The results are refined by 'Highly Cited Papers'. A sidebar on the left contains a menu with icons for search, filters, and alerts. The main results area shows a list of publications, with the first one being 'For a Sociology of Expertise: The Social Origins of the Autism Epidemic' by Eyal, G. The article is from the American Journal of Sociology, 118 (4), pp.863-907, Jan 2013. It has 247 citations and 167 references. The interface also includes an export menu with options like EndNote online, EndNote desktop, Plain text file, RIS, BibTeX, Excel, Tab delimited file, Printable HTML file, InCites, Email, and Fast 5000. The bottom of the interface shows pagination and sorting options.

The manual part?

Garbage-in; garbage-out
Preprocessing/Post-processing

Where's the data from?
What's in the data?

What language model do we use?
Why?

The screenshot displays the Clarivate Web of Science interface. At the top, the Clarivate logo is on the left, and 'English' and 'Products' are on the right. The main header shows 'Web of Science™' and a 'Search' button. Below this, the search path is 'Search > Results for Sociology (Topic) > Results for Sociology (Topic) and Highly Cited Papers'. The main content area shows '170 results from Web of Science Core Collection for: Sociology (Topic)'. A search bar contains 'Sociology (Topic)'. Below the search bar, it says 'Refined By: Highly Cited Papers X Clear all'. There are links for 'Copy query link' and 'Publications'. A 'You may also like...' section is visible. On the left sidebar, there is a 'MENU' icon and a list of filters: 'Filter by Marked List' and 'Quick Filters'. The 'Quick Filters' list includes: 'Highly Cited Papers' (170), 'Hot Papers' (4), 'Review Article' (30), 'Open Access' (92), and 'Associated Data' (8). On the right, there is an 'Export' dropdown menu with options: 'EndNote online', 'EndNote desktop', 'Plain text file', 'RIS (other reference software)', 'BibTeX', 'Excel', 'Tab delimited file', 'Printable HTML file', 'InCites', 'Email', 'Fast 5000', and 'More Export Options'. Below the export menu, there is a 'Report' button and a 'Create Alert' button. The main results area shows a list of results. The first result is 'For a Sociology of Expertise: The Social Origins of the Autism Epidemic' by Eyal, G., published in Jan 2013 in the AMERICAN JOURNAL OF SOCIOLOGY, 118 (4), pp.863-907. It has 247 Citations and 167 References. The article abstract is: 'This article endeavors to replace the sociology of professions with the more comprehensive and timely sociology of expertise. It suggests that we need to distinguish between experts and expertise as requiring two ... Show more'. There are links for 'Free Published Article From Repository' and 'Full Text at Publisher'. The bottom right corner shows 'Related records'.

The Model

Pretrained

Small – fits in memory

Only compatible with T-5 by Raffel et al (2020).

LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale

Tim Dettmers^{λ*}

Mike Lewis[†]

Younes Belkada^{§‡}

Luke Zettlemoyer^{†λ}

University of Washington^λ

Facebook AI Research[†]

Hugging Face[§]

ENS Paris-Saclay[‡]

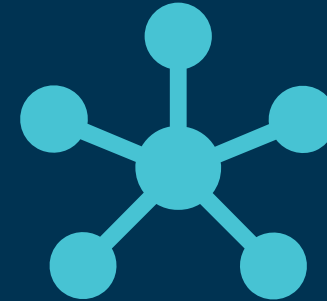
	GLUE	CNN3M	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	39.77	24.04

Table 1: Average and standard deviation of scores achieved by our baseline model and training procedure. For comparison, we also report performance when training on each task from scratch (i.e. without any pre-training) for the same number of steps used to fine-tune the baseline model. All scores in this table (and every table in our paper except Table 14) are reported on the validation sets of each data set.

The inputs and outputs - Further research



Semantic Search
Programme



Semantic Network Graph

References

DETTMERS, T., LEWIS, M., BELKADA, Y. & ZETTLEMOYER, L. 2022. LLM.int8 (): 8-bit Matrix Multiplication for Transformers at Scale. *arXiv preprint arXiv:2208.07339*.

RAFFEL, C., SHAZEER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W. & LIU, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21, 1-67.

TSHITOYAN, V., DAGDELEN, J., WESTON, L., DUNN, A., RONG, Z., KONONOVA, O., PERSSON, K. A., CEDER, G. & JAIN, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571, 95-98.

YOSHITAKE, M., SATO, F., KAWANO, H. & TERAOKA, H. 2022. MaterialBERT for natural language processing of materials science texts. *Science and Technology of Advanced Materials: Methods*, 2, 372-380.