



Smarter Literature Reviews: Modernizing Scientific Knowledge with Ai Based Automation

Priya Bhayani (225017)

Pooja Kuntinamadunagaraju (225270)

Agenda

1. Introduction
2. Program Workflow
3. Evaluation & Results
4. Conclusion

Overview

This project develops an AI-powered pipeline to automate essential steps in scientific literature reviews. Unlike traditional approaches based on TF-IDF, it focuses on enhancing the critical filtering and clustering phase through semantic embeddings.

The goal is to improve the relevance and coherence when grouping research papers.

Specifically, we aim to:

- Replace outdated keyword-based filtering with deep learning-driven semantic understanding.
- Enable meaningful clustering of related research works based on context, not just word overlap.
- Develop a scalable toolchain that can be adapted to a variety of research fields.

Introduction

What is the Current Problem

- Information Overload: The sheer volume of new research papers is overwhelming, making it hard to stay current.
- Outdated Methods: Traditional literature reviews are slow and manual, taking too much time.
- Semantic Gap: Old methods like TF-IDF don't understand the true meaning of words, leading to:
 - Poorly organized topics.
 - Missed connections between related papers.
 - A lot of extra manual work for researchers.

Introduction

Our Solution: Smart Automation with Embeddings

- Core Idea: We replace old methods with advanced AI models called "transformer-based sentence encoders."

Key Embedding Models:

- E5-base: Great at understanding the full meaning and context of sentences.
- All-MiniLM-L6-v2: Powerful and efficient, providing deep understanding in a compact form.

Benefits:

- Generates "semantic embeddings" that truly capture meaning.
- Allows for precise filtering and intelligent grouping of papers.

Program Workflow

Workflow: An End-to-End Automated Pipeline :

Step 1. Data Collection — Findpapers (PubMed, Scopus, arXiv)

Step 2. Converting data from JSON to Excel

Step 3. Clean and Preprocess Abstract Texts

Step 4. Generate Sentence Embeddings (E5 & All-MiniLM-L6-v2)

Step 5. Filter via Cosine Similarity (≥ 0.80)

Program Workflow

Step 1. Data Collection (Findpapers Library):

- **Findpapers** is a Python library that allows you to automatically search academic papers from multiple databases.
- enter a search query (e.g. "VR AND Mental Health"), specify the date range, and select databases.
- It fetches paper details like **title**, **authors**, **publication date**, **abstract**, **DOI**, **journal name** and The results are saved in a **JSON file**.

The screenshot shows the Findpapers web interface. At the top, there is a search bar with the text "search_rhif_all". Below it, there is a slider for "NUM:" ranging from 0 to 1000. Further down, there are date pickers for "SINCE:" (01/01/2013) and "UNTIL:" (01/05/2025). At the bottom, there is a section "Select Journals:" with three checkboxes: "pubmed:" (checked), "arxiv:" (checked), and "scopus:" (checked).



```

"abstract": "Conversations utilizing video-conferencing (VTC) as virtual rehabilitation have gained popularity, yet their efficacy has not been fully established. This study analyses  

used in conversational studies that the past decade that employed VTC for social communication. The findings reveal a moderate positive effect (d=0.40) on improving social  

the reducing repetitive behaviors, and increasing functional independence in individuals with autism. Given the findings and the implications, autism, and related anxiety, notably  

social deficits. While participants indicated the need, followed by abstractness, duration, and content. This literature may be relevant to the publishing and responses of studies in terms of  

VTC and related. Individuals with autism spectrum disorder (ASD) and comorbid anxiety have been the focus of VTC in the literature and the literature. Identifying the most  

effective factors in VTC. Furthermore, research findings indicate the potential of VTC, but to the variability among individuals with ASD and the complexity of assessment measures.  

This analysis is needed to identify the underlying factors influencing the effectiveness of these conversations.",
"authors": [
  "Taheri, Amir",
  "Taheri, Amir",
  "Gholami, Amir",
  "Shahmoradian, Amir",
  "Taheri, Amir"
],
"journal": "Journal of Autism and Developmental Disorders",
"pubmed_id": "35111111",
"arxiv_id": "2101.00000",
"scopus_id": "2-s2.0-35111111111",
"title": "The Effect of Virtual Reality on Social Communication Skills in Children with Autism Spectrum Disorder: A Systematic Review",
"doi": "10.1007/s10804-021-03584-4",
"year": 2021,
"month": 01,
"day": 01,
"volume": 51,
"issue": 1,
"pages": "1-15",
"abstract": "The purpose of this study was to investigate the effectiveness of virtual reality (VR) in improving social communication skills in children with autism spectrum disorder (ASD). The study included 10 studies that met the inclusion criteria. The results showed that VR had a positive effect on social communication skills in children with ASD. The effect size was moderate to large. The study also found that VR had a positive effect on the quality of life of children with ASD. The study concluded that VR is an effective intervention for improving social communication skills in children with ASD."

```

Step 2. JSON to Excel Conversion & Data Structuring

- Parses JSON to extract title, publication date, abstract, authors, database, publisher, DOI, citation count.
- Converts this data into a **Pandas DataFrame** (a tabular format) .
- Identifies and removes near-duplicate entries based on titles.
- Saves it as an **Excel file** for easy viewing and next processing steps.

```

10  "abstracts": "Interviews utilizing focus-group interview (FGI) to explore individuals' views about popularity, yet their efficacy has not been fully elucidated. This review synthesizes
11  300 observational research from the past decade that explored both online and offline interactions. The findings revealed a numerous pattern of influences (algorithmic-based, on reporting
12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149  150  151  152  153  154  155  156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199  200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249  250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299  300  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345  346  347  348  349  350  351  352  353  354  355  356  357  358  359  360  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390  391  392  393  394  395  396  397  398  399  400  401  402  403  404  405  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435  436  437  438  439  440  441  442  443  444  445  446  447  448  449  450  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495  496  497  498  499  500  501  502  503  504  505  506  507  508  509  510  511  512  513  514  515  516  517  518  519  520  521  522  523  524  525  526  527  528  529  530  531  532  533  534  535  536  537  538  539  540  541  542  543  544  545  546  547  548  549  550  551  552  553  554  555  556  557  558  559  560  561  562  563  564  565  566  567  568  569  570  571  572  573  574  575  576  577  578  579  580  581  582  583  584  585  586  587  588  589  590  591  592  593  594  595  596  597  598  599  600  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645  646  647  648  649  650  651  652  653  654  655  656  657  658  659  660  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690  691  692  693  694  695  696  697  698  699  700  701  702  703  704  705  706  707  708  709  710  711  712  713  714  715  716  717  718  719  720  721  722  723  724  725  726  727  728  729  730  731  732  733  734  735  736  737  738  739  740  741  742  743  744  745  746  747  748  749  750  751  752  753  754  755  756  757  758  759  760  761  762  763  764  765  766  767  768  769  770  771  772  773  774  775  776  777  778  779  780  781  782  783  784  785  786  787  788  789  790  791  792  793  794  795  796  797  798  799  800  801  802  803  804  805  806  807  808  809  810  811  812  813  814  815  816  817  818  819  820  821  822  823  824  825  826  827  828  829  830  831  832
```



1	A	B	C	D	E	F	G	H	I	J
2	Title	Year	Abstract	Authors	Databases	Publisher	Journal	Keywords	DOI	Citations
3	Current Tr	2025-05-01		Gallieri G.	Scopus		Journal of Clinical Mi	10.3390/jc	0	
4	The effect 2025-05-01	Interventi	Yifu, Liu	Scopus		Springer U	Education and Infor	10.1007/s	0	
5	Examining 2025-05-01		Kuleli D.	Scopus		Elsevier M	Journal of Behaviora	10.1016/j.	0	
6	The influe 2025-05-01		Zhao M.	Scopus		Elsevier B.	Acta Psychologica	10.1016/j.	0	
7	Examining 2025-05-01		Belhan Z.	Scopus		Tissue Via	Journal of Tissue Vi	10.1016/j.	0	
8	Underlyin 2025-05-01		Schevene	Scopus		Elsevier I	n Behavior Therapy	10.1016/j.	1	
9	The Use o 2025-05-01		Laker C.	Scopus		MDPI AG	Nursing Reports	10.3390/n	0	
10	The effect 2025-05-01		Ke Z.	Scopus		Elsevier I	n General Hospital Psy	10.1016/j.	2	
11	Retail valu 2025-05-01		Rumokoy	Scopus		Elsevier LT	Journal of Retailing a	10.1016/j.	4	
12	Personaliz 2025-05-01		Namung	Scopus		JAMA Network Open	10.1001/ja	0	0	
13	Supportin 2025-05-01	PURPOSE	Hannah E	PubMed			Current ps	N Anxiety disorders/	0	
14	A new tre 2025-05-01	INTRODUC	Nicholas E	PubMed		Frontiers I	Frontiers i	emotions/ N treat	0	
15	Pathway t 2025-05-01		van t' Woi	Scopus			Brain Stimulation	10.1016/j.	0	
16	Feasibilit 2025-05-01		Kang B.-R.	Scopus		SAGE Publ	Journal of Alzheimer	10.1177/1	0	
17	Using Vitr 2025-05-01		Garcia-Gu	Scopus			Journal of Clinical Ps	10.1002/jc	2	
18	Effects of 2025-05-01		Şimşekli D	Scopus		Elsevier I	n Heart and Lung	10.1016/j.	0	
19	Use of Tec 2025-05-01		Ferreira A	Scopus			Creative Nursing	10.1177/1	0	
20	Assessing 2025-05-01		Sepanloo	Scopus			Sensors	10.3390/s	0	
21	Research (2025-05-01		Zhane B.	Scopus			Computer Animatio	10.1002/c	0	

Program Workflow

Step 3. Data Cleaning and Preprocessing:

- Raw abstracts often have noise: HTML tags, special characters, non-English text, empty values.
- Handles missing values, removes HTML tags and placeholders.
- Detects and discards non-English texts.
- Strips predefined patterns (e.g., "Objective:"), eliminates unwanted characters.
- Converts all text to lowercase for uniformity.

Program Workflow

Step 4. Generate Sentence Embeddings (E5 & All-MiniLM-L6-v2)

- Utilizes pre-trained language models from the sentence-transformers library.
- Converts cleaned abstract texts into dense numerical vectors (embeddings).
- **Two models used:**
 - E5-base → accurate, high precision
 - All-MiniLM-L6-v2 → fast and lightweight
- These embeddings are added as new columns in the DataFrame.

Program Workflow

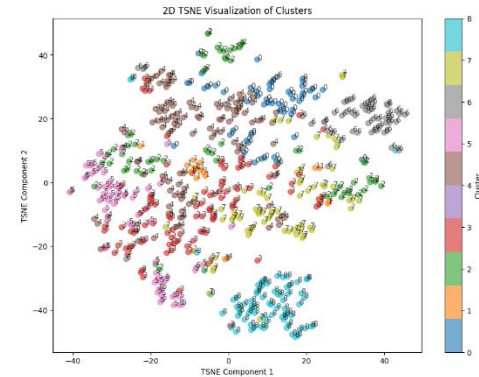
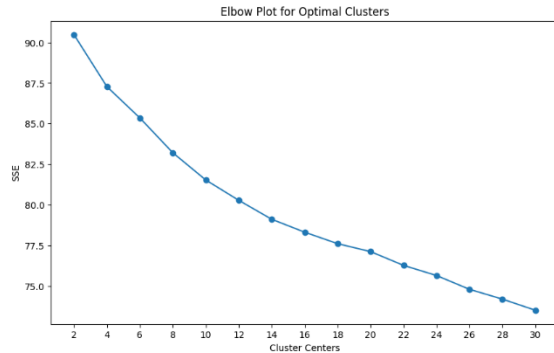
Step 5. Cosine Similarity Calculation:

- Calculates semantic closeness between document embeddings and a user-defined retrieval query.
- A custom query like "VR AND Mental Health" is converted into an embedding.
- Calculate **cosine similarity** between the query embedding and each paper's embedding.
- Applies a stringent similarity threshold (e.g., ≥ 0.80) to retain only the most relevant papers.

Program Workflow

Step 5. Clustering (MiniBatchKMeans & Elbow Method, t-SNE Visualization):

- Employs MiniBatchKMeans for thematic grouping.
- Uses the Elbow Method to determine the optimal number of clusters (e.g., $k=9$ for E5).
- Assigns cluster labels to each paper.
- Visualizes high-dimensional embeddings in 2D using t-SNE (with PCA pre-reduction) for intuitive thematic inspection.



Evaluation & Results

Search Query Definition and Data Collection Summary

To validate the utility of this automated system, we implemented the following criteria:

- **Query** : '([Virtual Reality] OR [Augmented Reality] OR [Extended Reality] OR [Mixed Reality]) AND ([Therapy] OR [Treatment] OR [Intervention] OR [Rehabilitation] OR [Outcome]) AND ([Mental] OR [Psychological] OR [Emotional] OR [Addiction] OR [Alcohol] OR [Exposure] OR [Behavior] OR [Disorder]) AND NOT [Physical]'
- **Period**: 2013 to 2023.
- **Selected databases**: PubMed, arXiv, and Scopus
 1. PubMed : 1000
 2. arXiv : 80
 3. Scopus : 1000
- The system successfully collected 2080 records.

Evaluation & Results

1. Cosine Similarity-Based Relevance Filtering

- After converting abstracts into embeddings using E5-base and All-MiniLM-L6-v2:
- Applied a cosine similarity threshold of 0.80.
- From an initial ~2000 papers, the pipeline filtered down to 590 highly relevant papers.
- Demonstrated strong ability to isolate papers semantically aligned with the query (e.g., *VR and mental health*).

↔ Saved 613 papers with similarity ≥ 0.8 to:
/content/drive/MyDrive/DL/most_similar_papers5.xlsx

	Title	Year	Databases	
0	Adverse Effects of Virtual and Augmented Reality...	2023-05-05	PubMed	
1	Physiological Factors Based Depression Assessm...	2025-01-01	Scopus	
2	Virtual Reality Interventions and Chronic Pain...	2025-01-01	Scopus	
3	Virtual Reality as a Supplement to Traditional...	2025-01-01	Scopus	
4	The use of virtual reality and augmented reali...	2022-12-14	PubMed	
5	Virtual reality and artificial intelligence: t...	2025-03-01	Scopus	
6	Virtual reality in the diagnostic and therapy ...	2022-10-30	PubMed	
7	Efficacy of virtual reality-based training pro...	2024-12-01	Scopus	
8	Extended Reality for Mental Health Evaluation...	2024-07-24	PubMed	
9	The Efficacy of Virtual Reality on the Rehabil...	2025-04-25	PubMed	
10	Virtual Reality Interventions for Mental Health...	2023-01-01	PubMed	
11	Reversing the Ruin: Rehabilitation, Recovery, ...	2022-10-01	PubMed	
12	Advances in the use of virtual reality to trea...	2024-08-01	Scopus	
13	Virtual reality interventions for the treatmen...	2023-02-25	PubMed	
14	Innovative Approaches for the Mental Developme...	2025-01-01	Scopus	
15	Mental health providers are inexperienced but ...	2024-07-03	PubMed	
16	Immersive virtual reality in the treatment of ...	2024-03-01	PubMed	
17	Virtual Reality Mental Health Interventions in...	2024-06-17	PubMed	
18	Review on the Role of Virtual Reality in Reduc...	2024-07-08	arXiv	
19	Virtual, mixed, and augmented realities: A com...	2024-07-08	PubMed	
20	Immersive Technologies for Depression Care: Sc...	2024-04-25	PubMed	
21	The Use of Virtual Reality Interventions to Pr...	2023-07-06	PubMed	
22	Barriers to adopting therapeutic virtual reali...	2025-03-18	PubMed	
23	Digital travel using virtual reality in inpati...	2025-04-28	PubMed	
24	Use of Virtual Reality and Augmented Reality T...	2024-07-08	PubMed	
25	[The application of virtual reality in the tre...	2022-09-02	PubMed	
26	Mixed Reality Technology to Deliver Psychologi...	2023-07-26	PubMed	
27	The Use of Virtual Reality in the Rehabilitati...	2023-09-01	PubMed	
28	Effectiveness and safety of virtual reality re...	2023-09-14	PubMed	
29	'Being there together for health': A Systemati...	2024-12-06	arXiv	

	Similarity
0	0.854708
1	0.853047
2	0.852649
3	0.851417
4	0.846814
5	0.845775
6	0.845193
7	0.844850
8	0.844224
9	0.843551
10	0.843280
11	0.843059
12	0.842136
13	0.842011
14	0.842006

Evaluation & Results

2. Clustering Analysis:

- **Methodology:** Cluster quality was assessed through the Elbow Method for optimal cluster count, analysis of cluster size distributions, manual purity labelling, and 2D semantic visualization via t-SNE.
- **Optimal Cluster Count:** The Elbow Method, plotting Sum of Squared Errors (SSE) against k (number of clusters), indicated "soft elbows" around k=9 for E5 and k=10 for All-Mini-LM6.
- **Cluster Purity:** Manual labelling revealed a mean purity of 70% for E5 clusters and 59% for All-Mini-LM6, with some clusters showing higher (e.g., E5 Cluster 7 at 83%) and lower purity.
- **2D Semantic Visualization (t-SNE):** This visualization, following PCA for dimensionality reduction, effectively illustrated well-isolated thematic clusters (e.g., autism-spectrum therapy) and areas of overlap reflecting semantic relatedness (e.g., chronic pain vs. palliative care).

Evaluation & Results

3. Retrieval Effectiveness:

- **Metrics:** Precision@K, Recall@K, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and nDCG@10 were computed for E5, All-MiniLM, and TF-IDF models using manually labeled relevance judgments.
- **Key Findings:**

Model	P@10	P@20	P@30	R@20	R@30	MAP	MRR	nDCG@10
E5	0.900	0.850	0.767	0.739	1.000	1.000	1.000	0.927
All-miniLM	0.800	0.750	0.633	0.789	1.000	1.000	1.000	0.852
TF-IDF	0.700	0.650	0.567	0.765	1.000	1.000	0.333	0.700

Evaluation & Results

4. Clustering Validity Indices:

- **Metrics:** Three standard validity metrics were applied: Silhouette Score (higher is better), Calinski-Harabasz Index (higher is better), and Davies-Bouldin Index (lower is better).
- **Findings:**
 - All-Mini-LM6 achieved the highest Calinski-Harabasz score and the lowest Davies-Bouldin index, signifying the most compact and well-separated clusters.
 - E5 followed closely with moderate cohesion.
 - TF-IDF consistently yielded the weakest cluster structure across all validity metrics.

Model	Silhouette	Calinski–Harabasz	Davies–Bouldin
E5	0.0258	10.6174	4.4173
All-miniLM	0.0221	18.0491	3.4929
TF-IDF	0.0105	5.8812	6.8061

Evaluation & Results

5. Comparison with Automated AI Tools:

- **Approach:** The pipeline's performance was benchmarked against external AI tools (Elicit, Litmaps, Semantic Scholar) using the same specialized Boolean-style query.
- **Results:** Despite indexing a wider range of sources, these tools returned only a handful of truly relevant results for our niche VR-mental health query.
- In stark contrast, our embedding-based pipeline's focused top 30 retrieval demonstrated substantially higher precision, significantly reducing the manual triage burden for researchers.

Conclusion

- We've successfully created an automated pipeline that transforms how literature reviews are done.
- The Power of Semantic Embeddings: By using E5-base and All-MiniLM-L6-v2, we've moved beyond the limitations of older methods like TF-IDF, truly understanding the meaning behind the text.
- **Through this approach, we achieved:**
 - High retrieval precision using cosine similarity filtering.
 - Accurate thematic clustering with MiniBatchKMeans, guided by the Elbow Method.
 - Clear 2D visualizations of high-dimensional data using PCA and t-SNE.
 - A systematic, structured review process aligned with the PRISMA framework.
- **The results showed that:**
 - E5-base delivered superior retrieval accuracy and early precision.
 - All-MiniLM-L6-v2 produced more cohesive and well-defined clusters.
 - Both models significantly outperformed traditional TF-IDF and common AI tools in precision, clustering quality, and efficiency.

References

- **Key References:**

- AI-Paper Miner — Open-source Project by Tamago55 (GitHub)
- Research Project by Alexander Bazhanov on automated literature review pipelines
- Bachelor Thesis by Rajakaruna on embedding-based literature review automation
- Sentence-Transformers Library — Pre-trained embedding models available via HuggingFace
- MTEB Benchmark Leaderboard — Evaluating embedding Model Performance Arcos Tasks
- PRISMA Guidelines — Standard Framework for Systematik literature Reviews and meta-Analyses
- APIs and Data Sources: Semantic Scholar API

THANK YOU!