Literature self Automated Report

Author: Denitsa Panova

Date: 2024-04-05

Disclaimer: The objective of this report is to present the outcomes generated by automated literature research. A specialized interpretation is essential to derive accurate conclusions regarding correct articles to be selfed.

Data Overview

Query "bee acoustic machine learning" has been executed in researchgate.com and all available results are scraped - in total 970. Note, that in the context of this report, a result is a search result, it can be an article, presentation, etc.

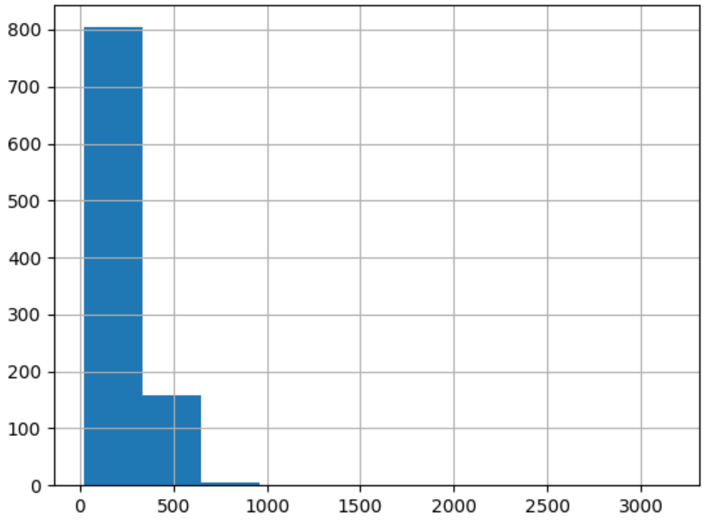
Time Distribution

Firstly, we will investigate how the results's distribution over the years.

Year	count
2024	553
2023	338
2022	40
2021	21
2019	9
2020	9

Word Distribution

Then, we will investigate what is the average count of words per abstract. Here the goal is to see if the default transformer model is still an adequate solution.The default model is all-mpnet-base-v2 and if the word count is above 384, it truncates the text and we wouldn't have full results in the encoding stage. We have 0.92 of the results consenting the criteria.



Result type

Investigate what type of results are extracted. This will help in the correct choice for clustering the results. The default types for clustering are: Article and Conference Paper.

Text Type	count
Article	707
Preprint	105
Conference Paper	51
Chapter	27
Literature Review	26
Book	11
Research	9
Research Proposal	9
Poster	7
Thesis	7
Technical Report	5
Data	2
Presentation	2
Patent	2

Language

Investigate what languages the results are in. We are considering only English languages for the purposes of this research After removing non-English results, we lose 0.01 of the data.

Language	count
en	990
id	5
es	1
ru	1
it	1
tr	1

Optimal ML Path

Traveling Salesmen Problem (TSP)

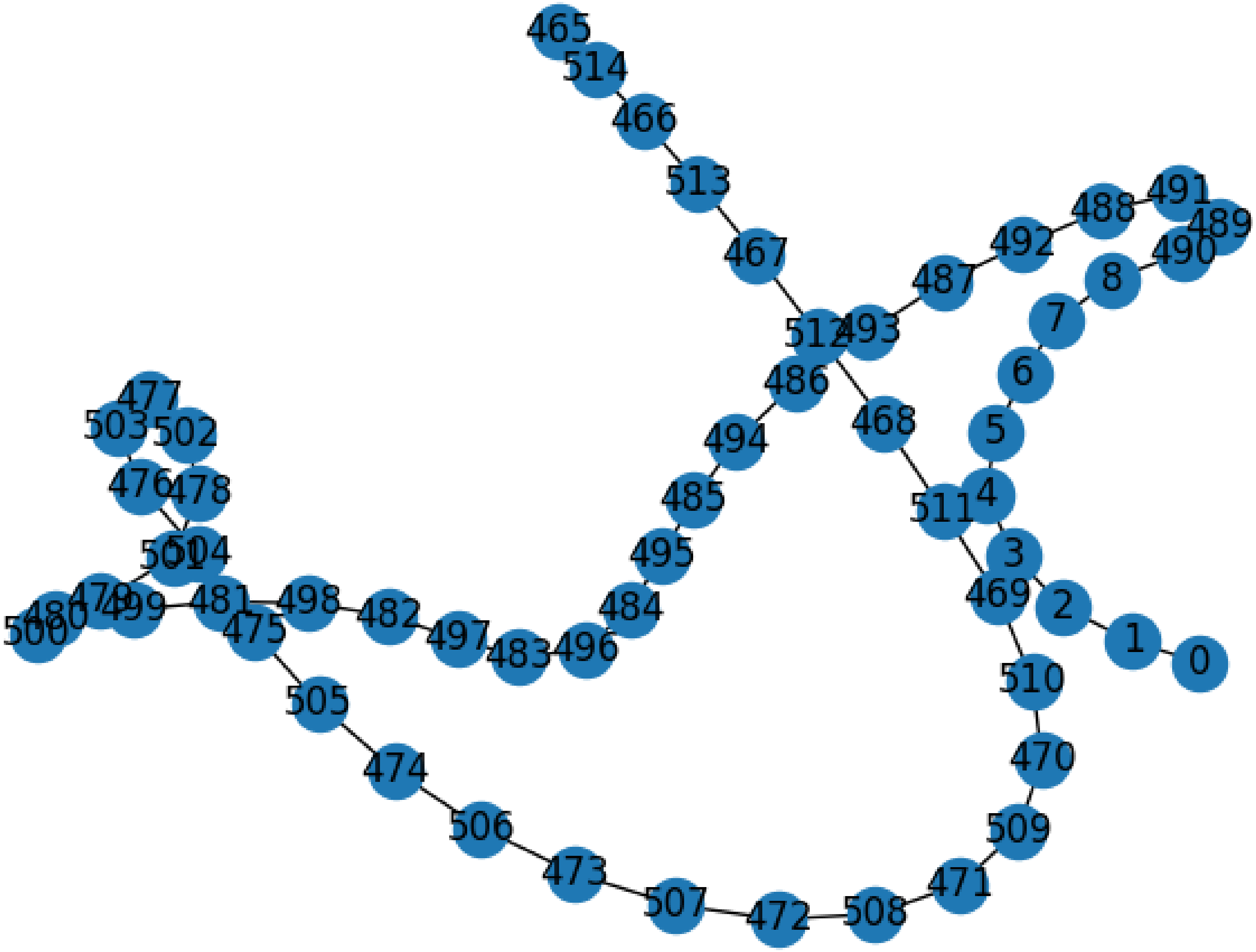
Each abstract is turned into an embedding, using the HuggingFace Transformer. The default transformer is all-mpnet-base-v2. After we calculate similarity between each scraped result with each another one, we apply TSP for finding the shortest path. The time for calculating it is 218960.53981781006 ms. Additionally, to optimize the TSP performance by artificially taking the first 9 results from the ResearchGate suggestions, then set to zero the connection between 9tha nd 10th,thus forcing the next TSP step to be there.

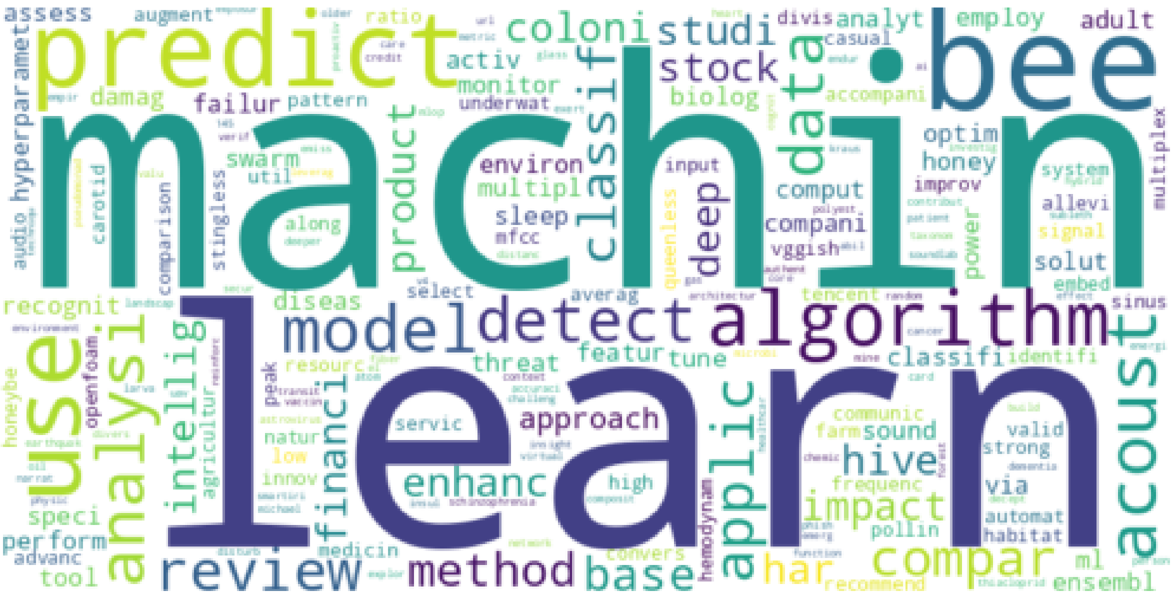
Similarity Difference

In order to investigate what is the added value of TSP, we check what is the average similarity measurein the first 50 TSP-suggested articles and the first 50 ResearchGate articles. However, we remove from calculation, the first 10 results. TSP is 0.31 and ResearchGate is 0.33.

Time Saved

Now we will calculate how much time it is saved by following the 50 articles suggested by TSP. The goal is to understand how much time is saved and at the same time wider knowledge on the topic is acquired. The metric encompasses the number of abstracts one should go over to read the same TSP-suggested results. According to a research, one can read around 150 words per minute. Therefore, the saved time by going with the TSP suggestion is 13.36 hours.





The next step is to look into the TSP results and identify those which are of interest for the research.We have identified those:

Original Index: 0

Title: Comparative Study of Machine Learning Models for Bee Colony Acoustic Pattern Classification on Low Computational Resources

Original Index: 1

Title: Automatic acoustic recognition of pollinating bee species can be highly improved by Deep Learning models accompanied by pre-training and strong data augmentation

Original Index: 2

Title: Bee detection in bee hives using selective features from acoustic data

Original Index: 3

Title: Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification

Original Index: 4

Title: Bee Swarm Activity Acoustic Classification for an IoT-Based Farm Service

Original Index: 6

Title: Identifying Queenlessness in Honeybee Hives from Audio Signals Using Machine Learning

Original Index: 494

Title: The role of hyperparameters in machine learning models and how to tune them

Original Index: 526

Title: Uncovering the important acoustic features for detecting vocal fold paralysis with explainable machine learning

Clustering Results

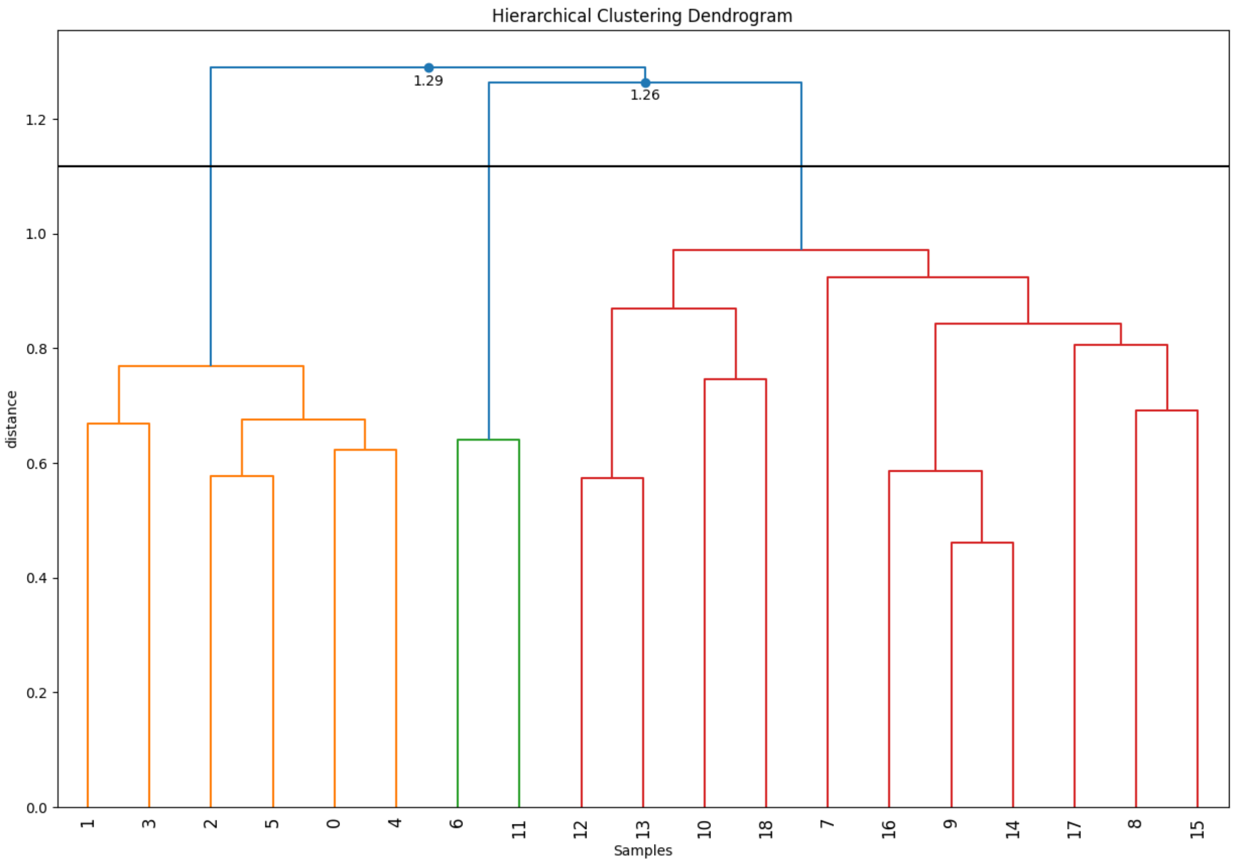
We decided to look into results with the following parameters: cosine similarity > than 0, year > than 2023, type of results Article , Conference Paper , Preprint , Patent , Thesis, number of reads > 10 and number of citations > 0.

Let us see the distribution of the HDBSCAN clusters.

Label	count
1	11
0	6
-1	2

Note: -1 label is for the outliers

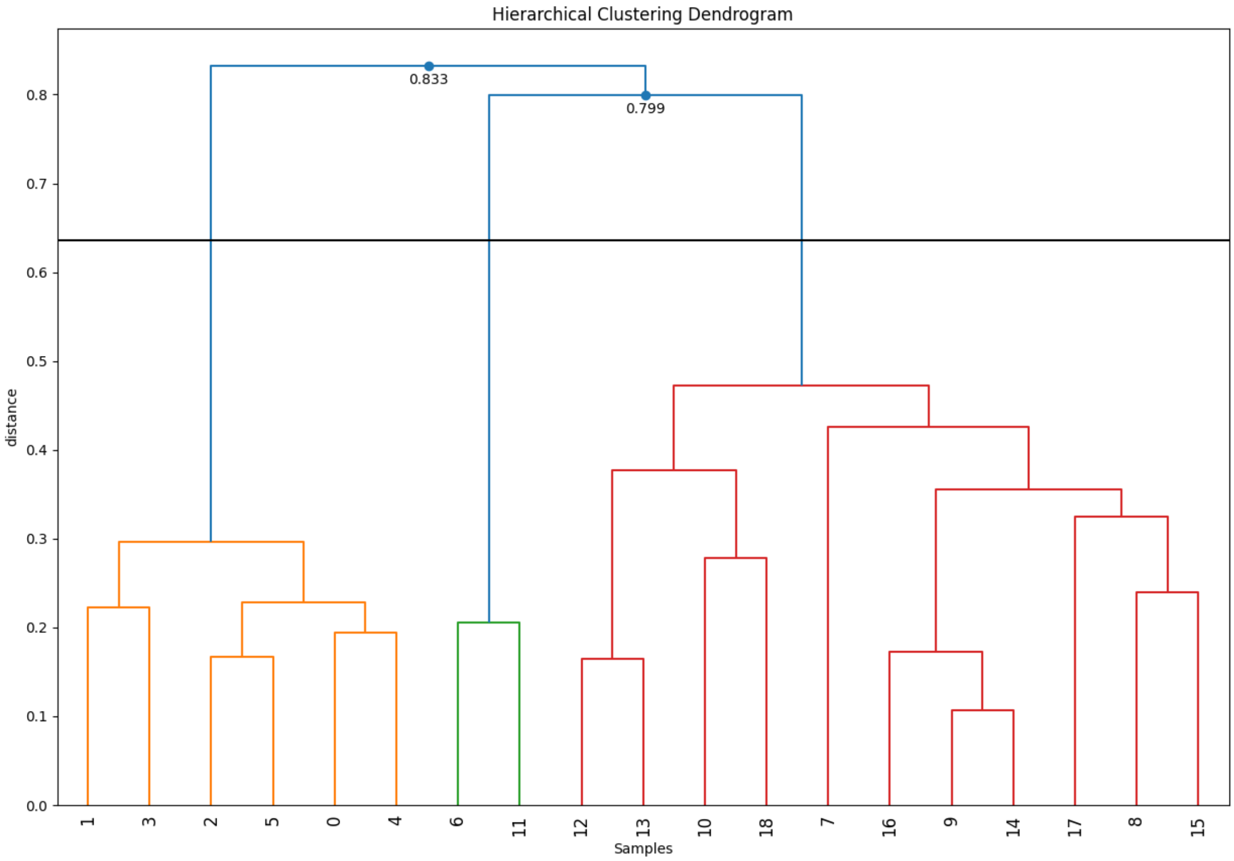
Let us see the dendogram of the HDBSCAN clusters.



Let us see the distribution of the Agglomeration clusters.

Label	count
2	11
0	6
1	2

Let us see the dendogram of the Agglomeration clusters.


Hierarchical Clustering Dendrogram
0.833
0.799
distance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
3
2
5
0
4
6
11
12
13
10
18
7
16
9
14
17
8
15
Samples

Agglomeration Suggested Reads

First, we will remove the outlier results (those which are classified from HDBSCAN) and show the rest as per the Agglomeration clustering technique.

Cluster 0

Original Index: 0

Title: Comparative Study of Machine Learning Models for Bee Colony Acoustic Pattern Classification on Low Computational Resources

Click here for the article

Original Index: 1

Title: Automatic acoustic recognition of pollinating bee species can be highly improved by Deep Learning models accompanied by pre-training and strong data augmentation

Click here for the article

Original Index: 2

Title: Bee detection in bee hives using selective features from acoustic data

Click here for the article

Original Index: 3

Title: Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification

Click here for the article

Original Index: 4

Title: Bee Swarm Activity Acoustic Classification for an IoT-Based Farm Service

Click here for the article

Original Index: 6

Title: Identifying Queenlessness in Honeybee Hives from Audio Signals Using Machine Learning

Click here for the article

Cluster 2

Original Index: 270

Title: Towards A Framework for Performance Management and Machine Learning in A Higher Education Institution

Click here for the article

Original Index: 461

Title: A Review of Fake Logo Detection using Machine Learning and Deep Learning

Click here for the article

Original Index: 470

Title: Research on feature coding theory and typical application analysis in machine learning algorithms

Click here for the article

Original Index: 494

Title: The role of hyperparameters in machine learning models and how to tune them

Click here for the article

Original Index: 557

Title: Strong machine learning: A way towards human-level intelligence

Click here for the article

Original Index: 631

Title: Machine Learning and Cryptanalysis: An In-Depth Exploration of Current Practices and Future Potential

Click here for the article

Original Index: 705

Title: Examining Different Data Decoupling Techniques for Federated Machine Learning with Databases as a Service

Click here for the article

Original Index: 763

Title: Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding

Click here for the article

Original Index: 797

Title: Financial Applications of Machine Learning Using R Software

Click here for the article

Original Index: 888

Title: A Comparative Study Evaluated the Performance of Two-class Classification Algorithms in Machine Learning

Click here for the article

Original Index: 951

Title: Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms

Click here for the article

HDBSCAN Suggested Reads

Cluster 0

Original Index: 0

Title: Comparative Study of Machine Learning Models for Bee Colony Acoustic Pattern Classification on Low Computational Resources

Click here for the article

Original Index: 1

Title: Automatic acoustic recognition of pollinating bee species can be highly improved by Deep Learning models accompanied by pre-training and strong data augmentation

Click here for the article

Original Index: 2

Title: Bee detection in bee hives using selective features from acoustic data

Click here for the article

Original Index: 3

Title: Applicability of VGGish embedding in bee colony monitoring: comparison with MFCC in colony sound classification

Click here for the article

Original Index: 4

Title: Bee Swarm Activity Acoustic Classification for an IoT-Based Farm Service

Click here for the article

Original Index: 6

Title: Identifying Queenlessness in Honeybee Hives from Audio Signals Using Machine Learning

Click here for the article

Cluster 1

Original Index: 270

Title: Towards A Framework for Performance Management and Machine Learning in A Higher Education Institution

Click here for the article

Original Index: 461

Title: A Review of Fake Logo Detection using Machine Learning and Deep Learning

Click here for the article

Original Index: 470

Title: Research on feature coding theory and typical application analysis in machine learning algorithms

Click here for the article

Original Index: 494

Title: The role of hyperparameters in machine learning models and how to tune them

Click here for the article

Original Index: 557

Title: Strong machine learning: A way towards human-level intelligence

Click here for the article

Original Index: 631

Title: Machine Learning and Cryptanalysis: An In-Depth Exploration of Current Practices and Future Potential

Click here for the article

Original Index: 705

Title: Examining Different Data Decoupling Techniques for Federated Machine Learning with Databases as a Service

Click here for the article

Original Index: 763

Title: Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding

Click here for the article

Original Index: 797

Title: Financial Applications of Machine Learning Using R Software

Click here for the article

Original Index: 888

Title: A Comparative Study Evaluated the Performance of Two-class Classification Algorithms in Machine Learning

Click here for the article

Original Index: 951

Title: Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms

Click here for the article