MONDINI MATTEO - COSTA PATRICK - ANDREA NUGARA

An interdisciplinary analysis of AI research: examining terminology evolution, authorship patterns, and content similarity across academic disciplines and years

In this Presentation

Overview

Introduction

Artificial intelligence is becoming more and more crucial in everyday life, with significant investments leading to substantial advancements and outcomes. In order to deeply understand this field, it is important to analyze how scientific papers discussing AI have transformed over time as well as understand the author’s net that wrote about it.

To achieve this, we collected almost **fifteen thousand papers** from arxiv, categorized them based on academic disciplines like computer science, chemistry, physics, finance and so on. Furthermore we arranged them by year periods that was more relevant. We then created customized **models** using different techniques for each subgroup to conduct an in-depth **semantic analysis** of the most relevant words related to AI.

Besides, we carried out a **collaborators finder** to determine the degree of content similarity among different authors with the goal to detect authors that might collaborate together in future. Furthermore, we utilized the **BERT** model to retrieve relevant papers through **sentence-based queries**.

Our study employed various methodologies, including constructing yearly and academic category slices, developing models for specific cases, conducting similarity analysis, utilizing **BERT** for paper retrieval, and employing **CADE** analysis across years and categories. Ultimately, our research aims to illuminate the evolution of AI terminology and concepts also trough a **SWEAT** analysis.

Data Retrieval

arxiv api: iteratively we got up to 15.000 papers with total of 770.000.000 words and stored them in a .csv file with relevant data.

We then made an entire corpora in a txt file as well as corpora divided by most relevant categories and years slices in order to train different models

Data Fields

id	id of Arxiv paper
Title	name of the paper
Authors	the list of the authors of the paper
Published At	the moment (datetime) of the publication
Updated At	the moment (datetime) of the last update
Summary	the abstract of the paper
Categories	different categories to which the paper refers
full-text	the complete text of the pdf paper cleaned and preprocessed

Word2Vec

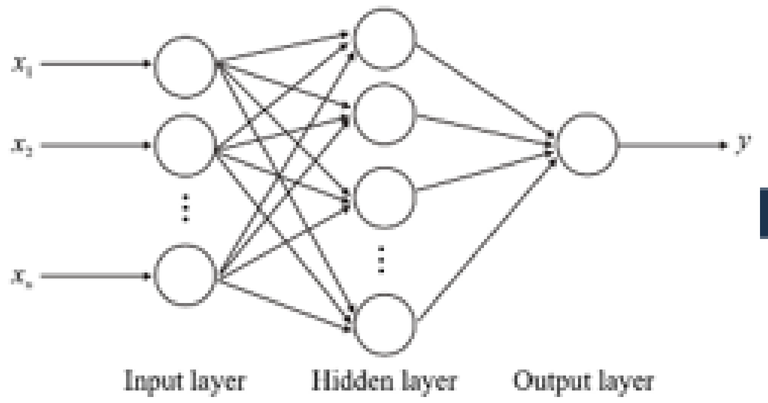
Definition: Word2Vec (word to vector) is a technique used to convert words to vectors, thereby capturing their meaning, semantic similarity, and relationship with surrounding text.

This method helps computers learn the **context** and connotation of expressions and keywords from large text collections such as news articles and books. The basic idea behind **Word2Vec** is to represent each word as a **multi-dimensional vector**, where the position of the vector in that high-dimensional space captures the meaning of the word.

Use: The purpose of the analysis is to observe how the **semantics** of certain terms can vary across different fields and time periods. Each model has been trained on specific categories or time slices.

Let's find the vectors for the words in the Sentence **'NLP is used for text analysis in machine learning'**

	Word
0	nlp
1	is
2	used
3	for
4	text
5	analysis
6	in
7	machine_learning



First 5 Numbers of Vector
[0.90100735 -1.0133158 1.3480183 0.05329623 0.38398987]
None
[0.47030523 0.05899952 0.05214813 -0.20349091 0.06618144]
None
[1.1893313 0.05017128 -0.8444921 1.6786242 -0.304366]
[-0.05181089 -1.0948385 -0.65095246 0.43532795 -0.33838072]
None
[0.38353053 -0.24118625 -0.9078417 -0.32669568 0.40940988]

In this demonstration, we utilize a specific phrase as input to generate numerical vectors for each word within the phrase. However, for simplicity and ease of interpretation, we only retain the initial five elements of each vector.

The model employed for this task is the 'statistics_model', which has been trained on a corpus of academic papers pertaining to the field of statistics. It's important to note that the actual dimensionality of these vectors is 100, meaning that in their full form, each vector consists of 100 elements.

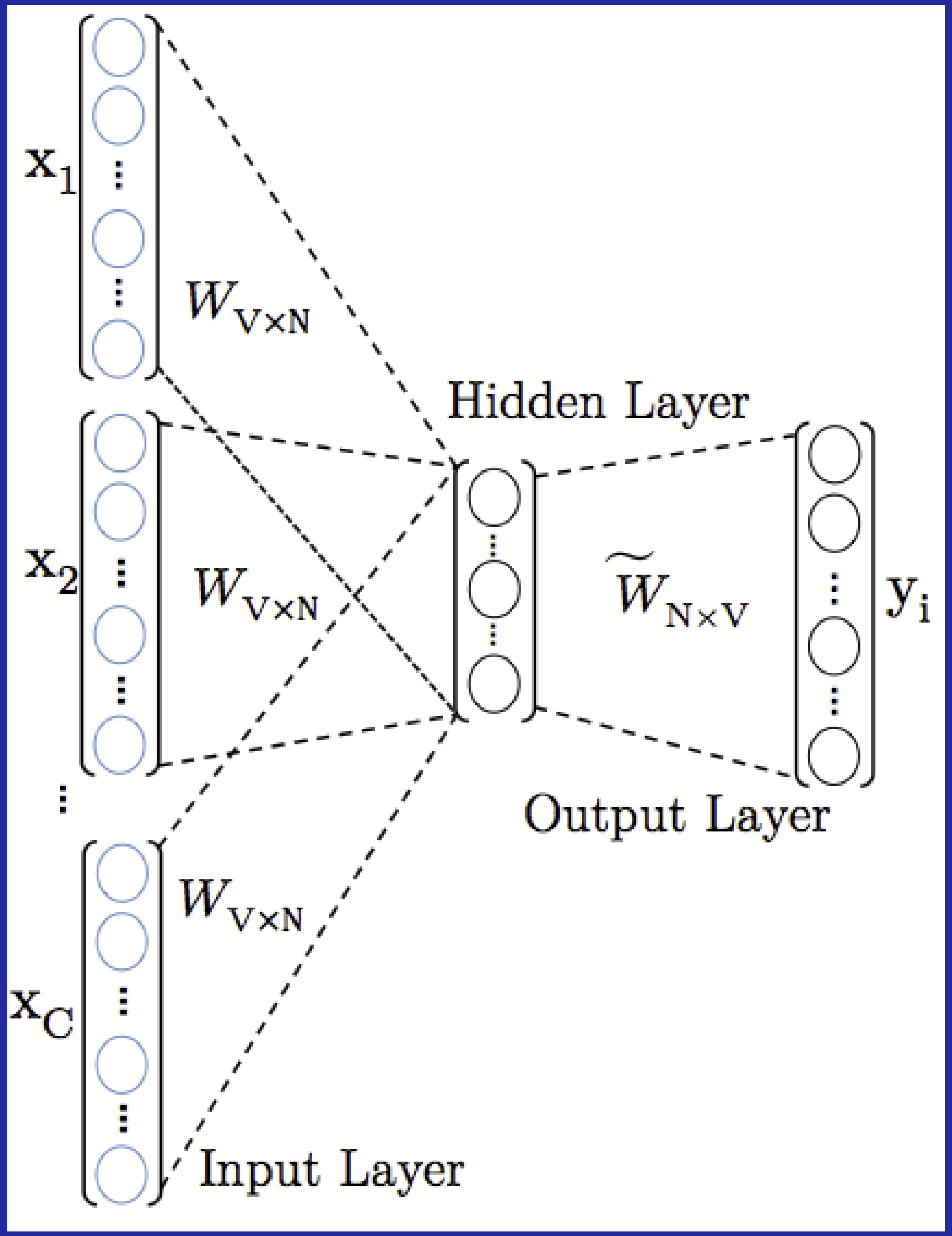
CBOW - Continuous Bag Of Words

In the context of **Word2Vec**, the Continuous Bag of Words (**CBOW**) model is one of two predictive models used for learning word embeddings, the other being the Skip-Gram model.

For our project, we chose to employ the CBOW model. This decision was primarily guided by the model's computational efficiency. The CBOW model is known for its faster training times and lower resource consumption compared to other models, making it an optimal choice for our specific use case.

The **CBOW** model functions by predicting a target word from its surrounding context words. In simpler terms, it uses the words around a particular word to predict that word.

The '**context**' is defined as a window of words around the target word, and the size of this window is a parameter that can be adjusted during the training process. This approach allows the model to capture the semantic and syntactic relationships between words, thereby creating meaningful word embeddings.

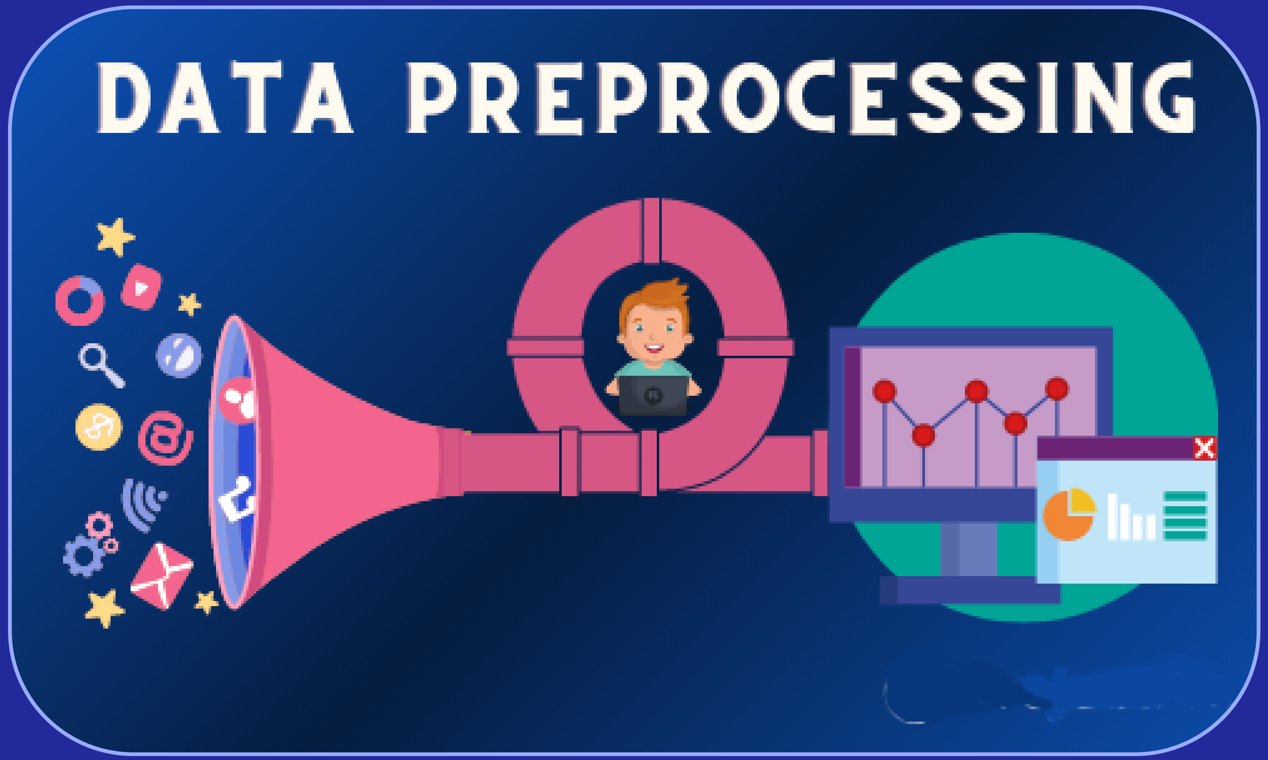


Text Lowercasing: All text was transformed to lower case to avoid duplication based on casing

Number Removal: Numerical values were removed from the text

Punctuation Removal: Punctuation marks and symbols were stripped from the text

Accent and Diacritic Removal: Accent marks and other diacritics were removed to standardize the text



Stop-word Removal: Common words (stop-words) that don't provide significant meaning were removed.

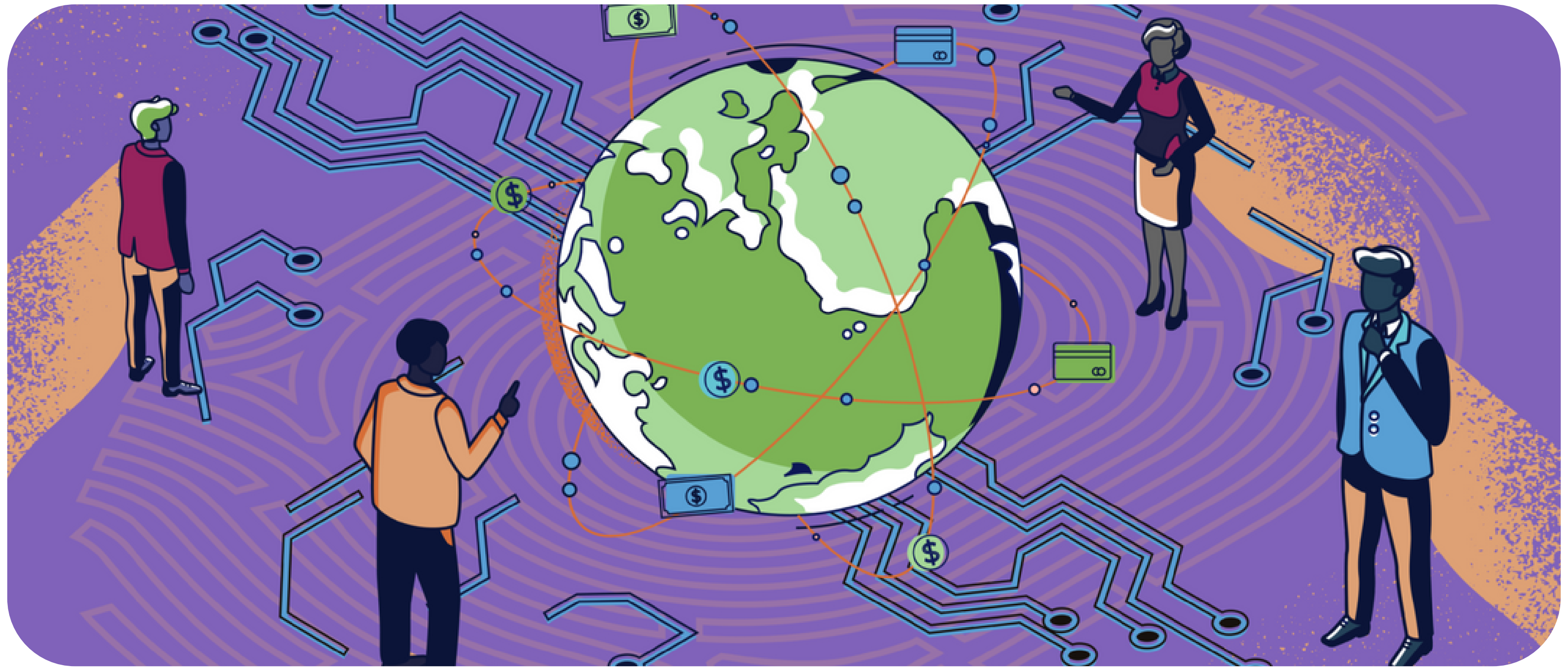
Demain specific stop words removal ("table", "figure", , "section", "reference", "abstract", "introduction" etc..)

Lemmatization: Words were lemmatized to their root form using WordNet's lemmatizer and part-of-speech tagging.

Bigram Creation: Based on the processed text, bigrams (pair of consecutive written units such as words) were identified and created.

Cross-category Analysis

Do models trained on corpora related to different topics have different outcomes on the same input?



The corpus was partitioned into several **categories** for the purpose of this study. For each category, a distinct model was trained, resulting in a range of models reflective of the different categories.

These models were then used to identify words most similar to specified input terms. The goal was to observe and analyze the variance in outputs across the diverse models.

This approach enabled the exploration of the nuanced contexts and meanings that specific key terms may have within different categories or fields of study.

Available categories

Category	Arxiv code	Number of papers	Number of tokens
Electrical Engineering and Systems Science	eess	1478	11.049.271
Statistics	stat	1097	10.686.482
Mathematics	math	723	3.554.971
Quantitative Biology	q-bio	415	3.198.048
Astrophysics	astro-ph	309	2.979.418
Quantitative Finance	q-fin	221	2.025.403

Results for word '**machine_learning**' differences in different models:

- Astrophisics
- Biology
- Electrical engineering
- Finance
- Math
- Statistics

Word: machine_learning

	Model	machine_learning Most Similar Word 1	machine_learning Similarity Score 1	machine_learning Most Similar Word 2	machine_learning Similarity Score 2	machine_learning Most Similar Word 3	machine_learning Similarity Score 3	machine_learning Most Similar Word 4	machine_learning Similarity Score 4
0	astro_model	semi_supervized	0.800523	algorithms	0.787236	methodology	0.776255	anomaly_identification	0.775642
1	biology_model	application	0.716028	mlbased	0.703069	ecology	0.693592	subfield	0.680797
2	electrical_model	anomaly_detection	0.791501	applied	0.784325	algorithms	0.782831	transformer_fault	0.778101
3	finance_model	algorithms	0.860040	stateoftheart	0.847113	technique	0.845234	traditional	0.831891
4	math_model	privacy_preserving	0.812577	encryption	0.797731	reinforcement_learning	0.791501	integrates	0.780914
5	statistics_model	application	0.750528	advance	0.731617	artificial_intelligence	0.719807	technique	0.704393

Results for word '**deep_learning**' differences in different models:

- Astrophisics
- Biology
- Electrical engineering
- Finance
- Math
- Statistics

Word: deep_learning

	Model	deep_learning Most Similar Word 1	deep_learning Similarity Score 1	deep_learning Most Similar Word 2	deep_learning Similarity Score 2	deep_learning Most Similar Word 3	deep_learning Similarity Score 3	deep_learning Most Similar Word 4	deep_learning Similarity Score 4
0	astro_model	forecast	0.767314	heliophysics_community	0.730957	realtime	0.725210	physics_inspired	0.717860
1	biology_model	neural_network	0.715699	lack_interpretability	0.709462	endtoend	0.690175	imagenet_classification	0.689758
2	electrical_model	state_art	0.886761	csmri	0.877590	novel	0.863843	deep_learningbased	0.855180
3	finance_model	traditional	0.846411	machine_learning	0.773912	modelling	0.767287	stateoftheart	0.756759
4	math_model	modulation_classification	0.748187	generative_adversarial	0.747389	diagnosis_covid	0.744835	model_driven	0.744261
5	statistics_model	learn	0.883352	neural_networks	0.799444	technique	0.711405	novel	0.706513

Results for word '**artificial_intelligence**' differences in different models:

- Astrophisics
- Biology
- Electrical engineering
- Finance
- Math
- Statistics

Word: artificial_intelligence

	Model	artificial_intelligence Most Similar Word 1	artificial_intelligence Similarity Score 1	artificial_intelligence Most Similar Word 2	artificial_intelligence Similarity Score 2	artificial_intelligence Most Similar Word 3	artificial_intelligence Similarity Score 3	artificial_intelligence Most Similar Word 4	artificial_intellige Similarity Scor
0	astro_model	methodology	0.792946	advance	0.779492	innovative	0.774213	multimessenger_astrophysics	0.765322
1	biology_model	bioimage_analysis	0.692844	fight_covid	0.682709	collaboration	0.664424	increasingly	0.640167
2	electrical_model	trustworthy	0.900462	future_direction	0.854412	taxonomy	0.853014	lesson	0.852492
3	finance_model	development	0.769001	debate	0.738649	responsible	0.737711	innovative	0.729098
4	math_model	automation	0.695160	trustworthy	0.629925	blockchain	0.611709	rigorous	0.598542
5	statistics_model	technology	0.723312	machine_learning	0.719808	science	0.662396	intelligent	0.662156

Conclusions

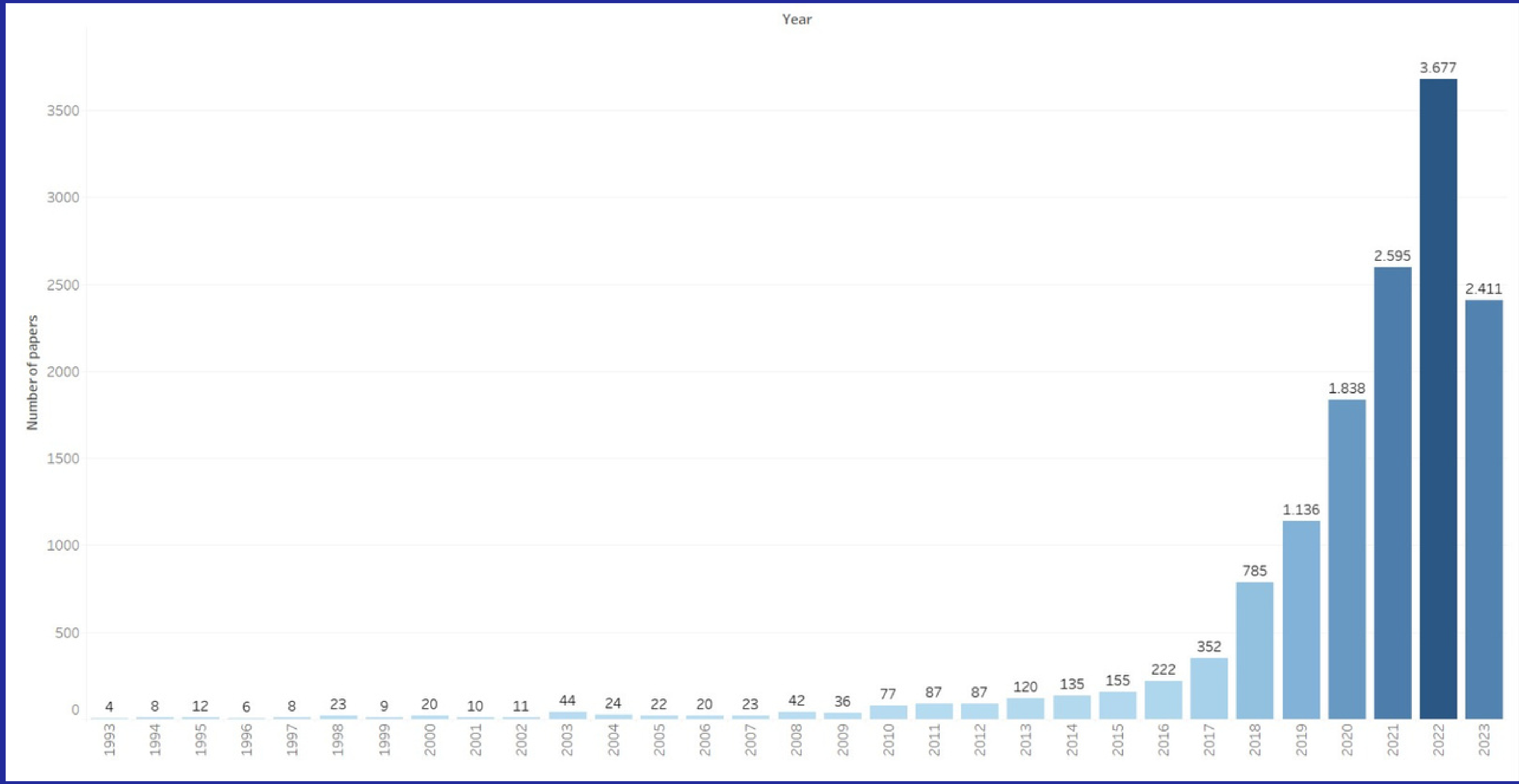
Each category slice shows diverse outputs based on the model, reflecting how the connotation of "machine_learning" , "deep learning" "artificial_intelligence" adapts to the specifics of each field. This highlights these technologies' versatility across a wide spectrum, from astrophysics to finance.

However, not all terms are domain-specific. Similar and general terms often recur across models, signifying a shared understanding of these technologies across various domains. This suggests that despite the unique applications in different fields, there exist common foundational principles and goals, showing the universal character of machine learning and artificial intelligence.

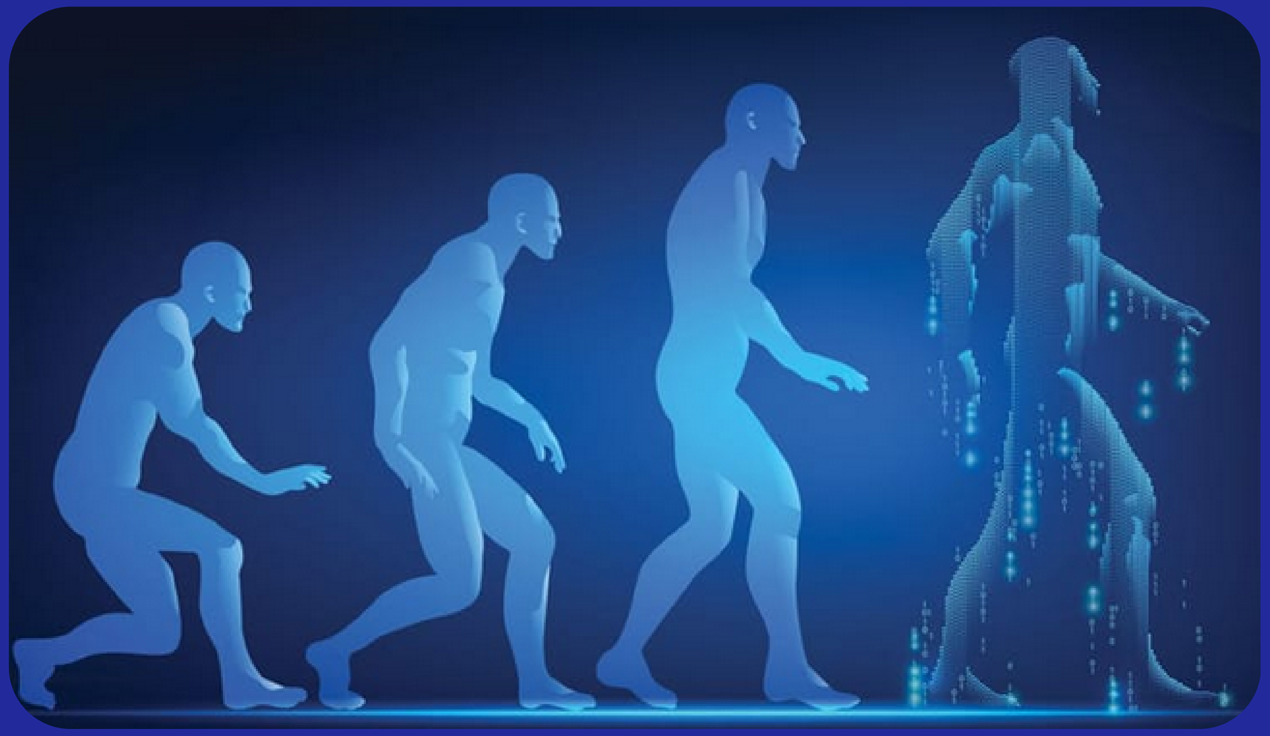
Temporal analysis

How has the concept of artificial intelligence changed over time?
Can we observe differences between the papers written before and after 2021?

To answer these questions, the models were trained in two temporal slices:
The first slice, grouping all the papers from 1993 to 2021 (included), contains 7576 papers for a total of 55,723,663 tokens.
The second slice from the beginning of 2022 to mid-2023 contains a total of 6088 papers with a total of 64,536,365 tokens.



What has been done



In the first analysis, we plotted the 10 words most similar to artificial intelligence in the two time slices on a compass graph, from which we then drew conclusions.

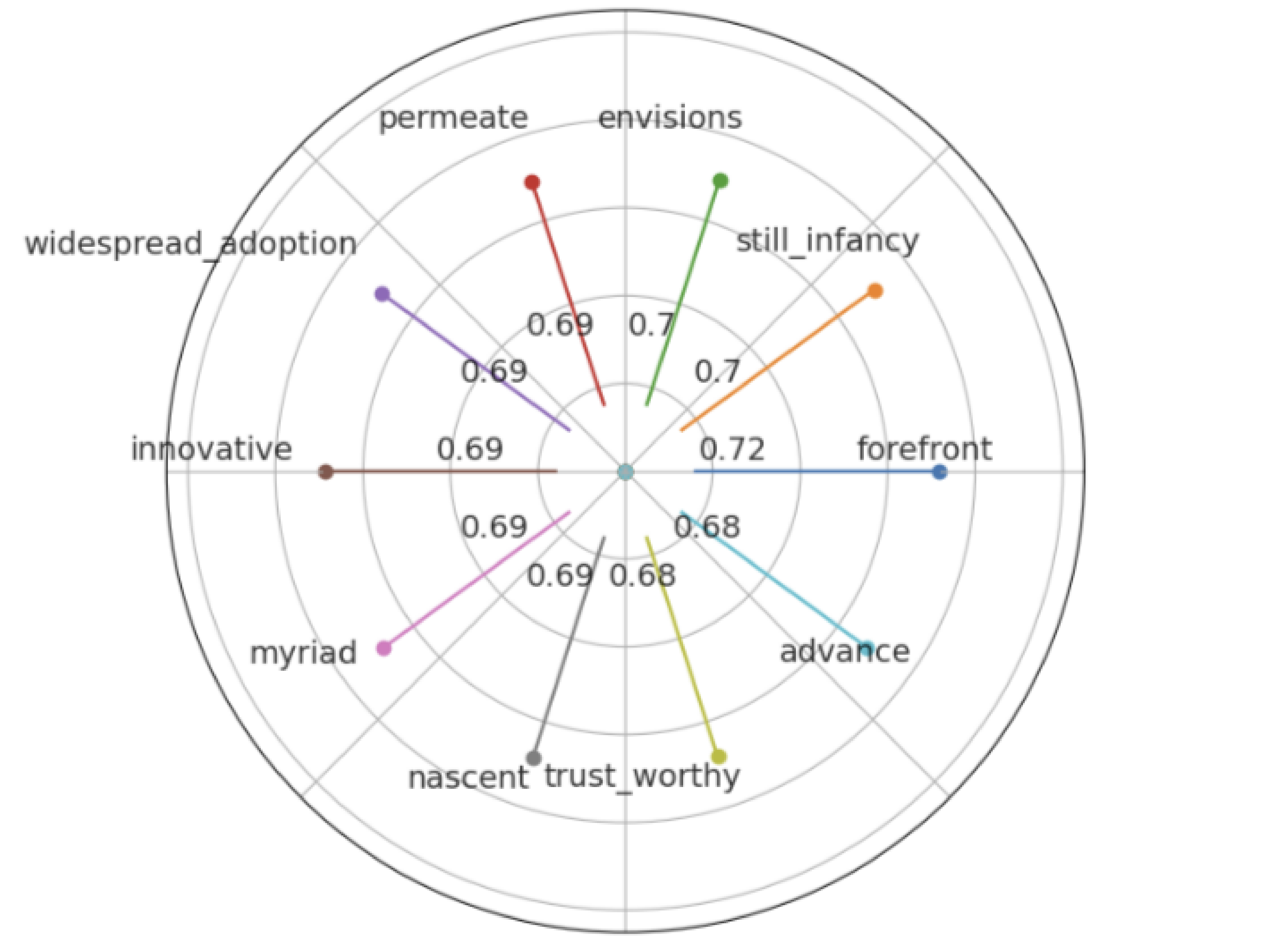
In the second analysis, we compared the similarity between the term artificial intelligence and various other terms always related to different themes connected to artificial intelligence, arbitrarily taken from the results of the previous analyses but using models trained on different time slices.

Results on the slice until 2021

Regarding the time slice from 1993 to 2021, the field of artificial intelligence was still in its early, formative stages. Terms like "**still infancy**," "**nascent**" and "forefront" suggest that the technology was perceived as emerging or still in its early development.

There was also an anticipation of its future influence, as indicated by words such as "**advance**", "**widespread_adoption**" and "**permeate**" .

Words like "**innovative**" and "**myriad**" highlight the novelty and potential wide-ranging applications of artificial intelligence.

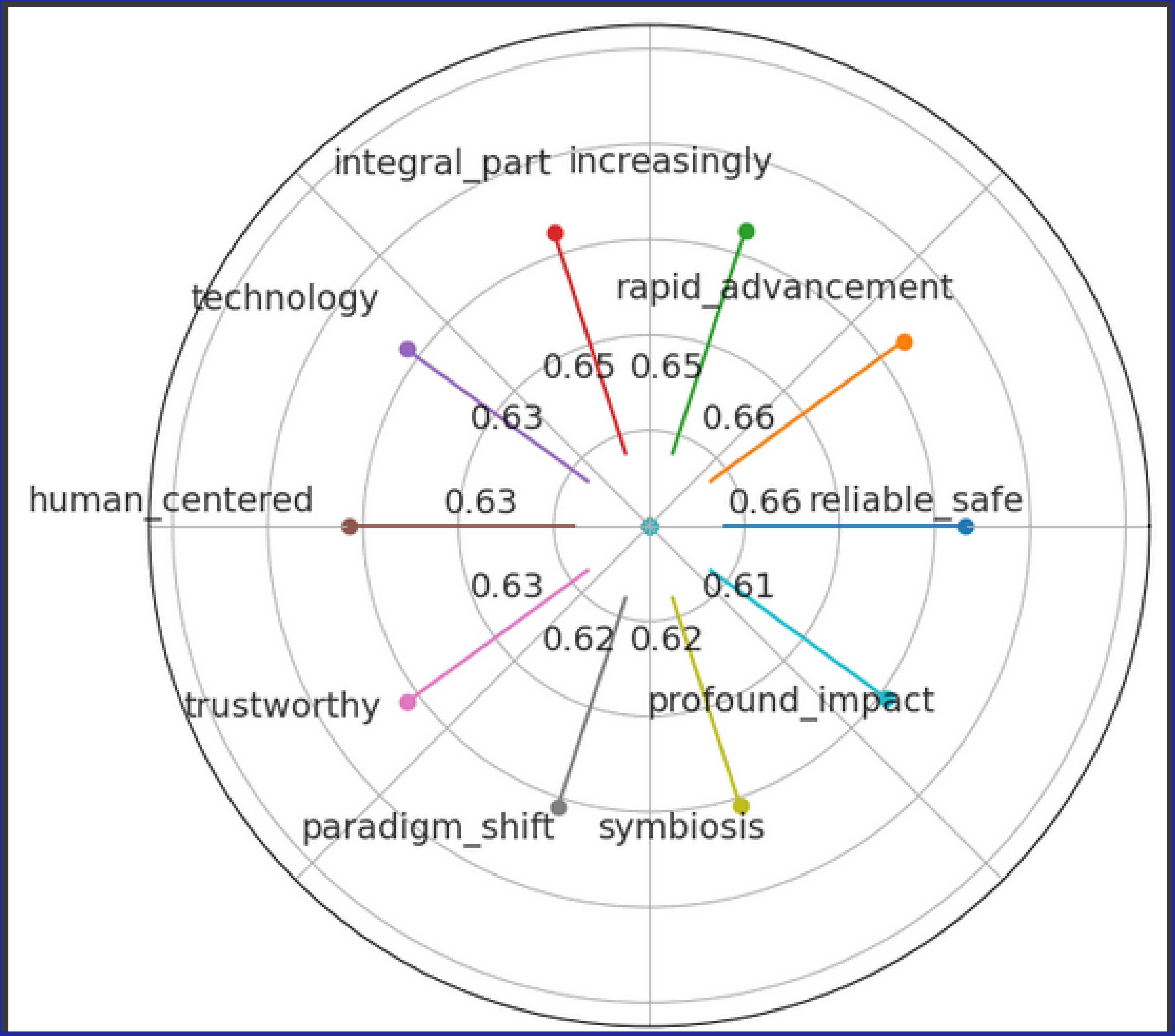


Results on the slice from 2022

Regarding the time slice from 2022 to mid 2023, there's a strong emphasis on the growth and integration of AI in various aspects of life and technology, suggested by terms such as "increasingly", "rapid_advancement", "profound_impact", "integral_part", and "technology".

The presence of "**paradigm_shift**" and "**symbiosis**" hints at a significant transformation in the way AI is understood and its relationship with humans and society. They suggest an evolving recognition of AI not merely as an external tool, but something more closely interwoven with human activities.

"**human_centered**" and "**trustworthy**" point to an increased focus on designing AI that is centered on human needs, ethical considerations, and building trust. The term "**safe**" further underscores this focus on reliability and trustworthiness



Results: differences between term 'artificial_intelligence' and other related words within 2 temporal slices

1. "**Future**": The term maintains a similar association with "artificial intelligence" across both time slices, indicating a consistent perception of AI as linked to the future.
2. "**Rich**": The similarity decreases significantly in the second time slice, suggesting that the association between wealth and AI has diminished over time.
3. "**Chatbot**": The similarity increases after 2021, which could reflect the growing prevalence of AI-driven chatbot applications in recent years.
4. "**Problem**": The similarity increases after 2021, which may imply an evolving view of AI as connected to problem-solving or possibly the emergence of new challenges associated with AI.
5. "**Regulation**": A considerable increase in similarity after 2021 suggests growing discussions around the regulation of AI.
6. "**Law**": Like regulation, the increase in similarity could indicate a heightened focus on the legal implications and considerations of AI after 2021.

	reference_word	comparison_word	similarity_until_2021	similarity_after_2021
0	artificial_intelligence	future	0.444482	0.441436
1	artificial_intelligence	rich	0.214934	0.141122
2	artificial_intelligence	chatbot	0.285324	0.320134
3	artificial_intelligence	problem	0.086613	0.146434
4	artificial_intelligence	regulation	0.316589	0.436491
5	artificial_intelligence	law	0.210496	0.326952

Conclusions

The comparison of artificial intelligence discourse before and after 2021 reveals a distinct shift in context. Initially, the focus was on the potential and growth of AI, reflecting its nascent state.

Post-2021, the discourse reflects a broader incorporation and impact of AI in society, with emphasis on safety and ethics.

This evolution in language suggests a maturing understanding and practical application of AI, moving from the realm of potential to tangible real-world implications.

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art **machine learning model** that understands the context and structure of sentences, enabling more accurate **natural language processing** tasks such as translations, sentiment analysis, and intent recognition.

In our use case, by understanding the context and structure of sentences, **BERT** enables more accurate analysis of text, facilitating tasks such as finding relevant papers and determining authorship patterns.

Main goals with **BERT**:

1. **Paper search**
2. **Collaborators finder**

What has been done with BERT

1. Paper search

Our goal is try to compare BERT with other popular ways of search, e.g. Bag Of Words, in order to understand if BERT is able to better find scientific papers from an input phrase, thanks to its peculiarity to find the context of the words bidirectionally in the texts.

Steps to implement BERT in paper search:

- load a pretrained model
- generate word embeddings for each Summary
- take a phrase in input to search
- embed the words of the input phrase
- calculate the cosine similarity within the summary's embeddings
- return the most similar

BERT comparison with BOW

For the comparison, we implemented a simple paper search with Bag Of Words

Steps to implement BOW in paper search:

- create the sparse matrix with the rows represented by the documents and the columns are the unique words in the summaries, the cells are the count of the single words present in each summary
- do the same for the input phrase
- calculate the cosine similarity within the summary's matrix and the one of the input
- return the most similar

Results

	Input phrase:	"computer vision for breast cancer recognition"
	BERT results	**BOW results**
1	Deep-LIBRA: Artificial intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment Authors: Omid Haji Maghsoudi, Aimilia Gastounioti, Christopher Scott, Lauren Pantalone, Fang-Fang Wu, Eric A. Cohen, Stacey Winham, Emily F. Conant, Celine Vachon, Despina Kontos	Deep-LIBRA: Artificial intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment Authors: Omid Haji Maghsoudi, Aimilia Gastounioti, Christopher Scott, Lauren Pantalone, Fang-Fang Wu, Eric A. Cohen, Stacey Winham, Emily F. Conant, Celine Vachon, Despina Kontos
2	Biomarker Gene Identification for Breast Cancer Classification Authors: Sheetal Rajpal, Ankit Rajpal, Manoj Agarwal, Naveen Kumar	Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5,060 patients and a deep learning model
3	Predicting Axillary Lymph Node Metastasis in Early Breast Cancer Using Deep Learning on Primary Tumor Biopsy Slides Authors: Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, Mulan Jin	A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks
4	Interpretable HER2 scoring by evaluating clinical Guidelines through a weakly supervised, constrained Deep Learning Approach Authors: Manh Dan Pham, Cyprien Tilmant, Stéphanie Petit, Isabelle Salmon, Saima Ben Hadj, Rutger H. J. Fick	Biomarker Gene Identification for Breast Cancer Classification Authors: Sheetal Rajpal, Ankit Rajpal, Manoj Agarwal, Naveen Kumar
5	Automated Scoring System of HER2 in Pathological Images under the Microscope Authors: Zichen Zhang, Lang Wang, Shuhao Wang	AI assisted method for efficiently generating breast ultrasound screening reports

Results

	Input phrase:	"NLP for legal documents"
	BERT results	**BOW results**
1	An Argumentation-Based Legal Reasoning Approach for DL-Ontology Authors: Zhe Yu, Yiwei Lu	SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval
2	Legal Sentiment Analysis and Opinion Mining (LSAOM): Assimilating Advances in Autonomous AI Legal Reasoning Authors: Lance Eliot	Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings
3	Burden of Persuasion in Argumentation . Authors: Roberta Calegari, Giovanni Sartor	Indian Legal NLP Benchmarks : A Survey
4	CaseEncoder: A Knowledge-enhanced Pre-trained Model for Legal Case Encoding Authors: Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, Yiqun Liu	Sim2Real Docs: Domain Randomization for Documents in Natural Scenes using Ray-traced Rendering .
5	Lawmaps: Enabling Legal AI development through Visualisation of the Implicit Structure of Legislation and Lawyerly Process Authors: Scott McLachlan, Evangelia Kyrimi, Kudakwashe Dube, Norman Fenton, Lisa Webley	How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact

Results

	Input phrase:	"Deep learning in finance"
	BERT results	**BOW results**
1	Data science and AI in FinTech: An overview Authors: Longbing Cao, Qiang Yang, Philip S. Yu	Deep Reinforcement Learning for Conversational AI .
2	Asset Pricing and Deep Learning Authors: Chen Zhang	Opening the Black Box of Financial AI with CLEAR-Trade: A CLass-Enhanced Attentive Response Approach for Explaining and Visualizing Deep Learning-Driven Stock Market Prediction
3	Forecasting Implied Volatility Smile Surface via Deep Learning and Attention Mechanism Authors: Shengli Chen, Zili Zhang	The many faces of deep learning
4	Theoretically Motivated Data Augmentation and Regularization for Portfolio Construction Authors: Liu Ziyin, Kentaro Minami, Kentaro Imajo	Deep Reinforcement Learning for Cybersecurity Threat Detection and Protection: A Review
5	Explaining AI in Finance: Past, Present, Prospects Authors: Barry Quinn	A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU

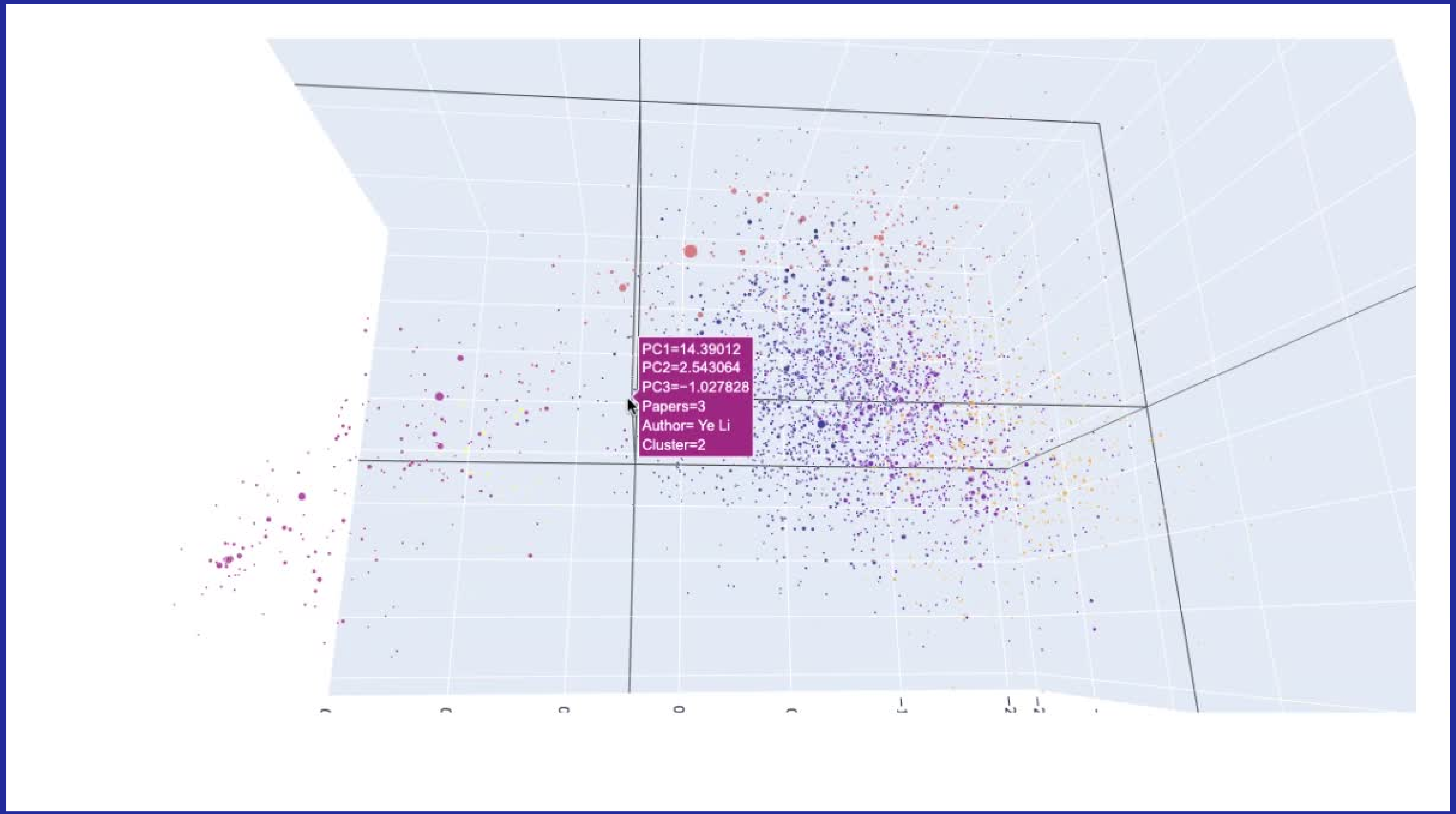
Conclusions

It should be noted that the embeddings were exclusively performed on the summaries of the papers. While conducting embeddings on the entire paper would have yielded more precise results, such an approach would have required a substantial amount of time.

Nonetheless, from an academic standpoint, we have successfully demonstrated how the utilization of BERT, in certain circumstances, surpasses other research methods, such as the Bag of Words. This superiority stems from BERT's ability to capture the semantic meaning of words, rather than simply quantifying their occurrences within the text.

Potential co-authors finder BERT

Utilizing the pre-trained **BERT** model, we generated **embeddings** for authors, which are visualized in a 3D interactive chart. The purpose of this visualization is to enable users to identify potential collaborators for a given author by entering their name. The proximity of authors within the chart indicates the similarity of their written papers, suggesting potential collaboration opportunities.



What has been done with BERT

2. Collaborators finder

The goal is to find, given an author name, the most '**compatible**' authors that belongs to the same cluster that never collaborate with him (that is not a co-author) and that wrote at least 4 papers. This might be useful to find **potential similar authors** in relation to the style of writing or for similar topics..

Steps to implement BERT in collaborators finder:

- load a pretrained **BERT** model
- create a new data frame in which for each author we save the number of papers that they wrote and the summaries of each paper, excluding author that wrote less than 3 papers
- preprocessing of the summaries (lowercase, punctation, tokenization, ...)
- With the **BERT** model generate the vector embeddings for each summary
- Calculate the mean between the different vector embeddings for each author and save the resulted vector
- Standardize the data
- Perform **PCA** to decrease dimensionality to 3 in order to visualize the plot
- Perform Agglomerative Clustering to divide authors in clusters
- Process the whole dataset to retrieve all the categories for each author
- Given an author name, we check the fuzz similarity with the authors in our data frame, and if it is more similar than 80% the retrieval begins
- we get the authors belonging to the same cluster, we calculate the cosine similarity
- return the first 5 author that has the higher similarity score

Results

- Rushin Shah, Papers: 4 Categories: cs.CV, cs.GR, cs.LG
- Xiaoyan Zhu, Papers: 3 Categories: cs.CL, cs.AI, cs.LG
- Marie-Francine Moens, Papers: 3 Categories: cs.LG, cs.AI, cs.CL
- Hannes Westermann, Papers: 6 Categories: cs.CV, cs.LG, cs.AI, cs.GR
- Ai Ti Aw, Papers: 5 Categories: cs.AI, cs.CL, cs.CV, cs.HC, cs.LG, cs.IR, cs.NE
- Karim Benyekhlef, Papers: 3 Categories: cs.LG, cs.AI, cs.CL
- Jaromir Savelka, Papers: 3 Categories: cs.LG, cs.AI, cs.CL

Author's name: Matteo Palmonari

Categories: cs.CL, cs.LG, cs.AI

Number of papers on Arxiv: 5

Rushin Shah

Google
Verified email at google.com
Natural Language Understa… Dialog Systems Machine Learning

Hannes Westermann

Cyberjustice Laboratory, Faculté de droit, Université de Montréal
Verified email at umontreal.ca
artificial intelligence & law natural language processing machine learning

Marie-Francine Moens

Professor of Computer Science KU Leuven
Verified email at cs.kuleuven.be
Natural language processin… machine learning information retrieval multimedia and text mining archaeology

Aw Ai Ti
Department Head (Aural & Language Intelligence)
Email: aaiti@i2r.a-star.edu.sg

Research areas:
Human Language Technologies, Audio Analytics & Speech Recognition, Speech Generation, Machine Translation, Question Answering & Dialogue Technology, Summarization

Publications

Ms Aw Ai Ti is the Head of the Aural & Language Intelligence (ALI) department at the Institute for Infocomm Research (I^2R), A*STAR, Singapore. She leads the department in the development and implementation of A*STAR's Audio, Speech and Language R&D strategies.

Results

- Jiajun Wu, Papers: 7 Categories: cs.CV, cs.RO, cs.AI, cs.LG, cs.HC
- Chengshu Li, Papers: 6 Categories: cs.AI, cs.CV, cs.RO, cs.LG
- Roberto Martín-Martín, Papers: 10 Categories: cs.CV, cs.LG, cs.AI, cs.RO, cs.HC
- Silvio Savarese, Papers: 22 Categories: cs.CV, cs.RO, cs.LG, cs.MA, cs.AI, cs.GR, cs.CL,
- Aniruddha Kembhavi, Papers: 16 Categories: cs.CV, cs.RO, cs.AI, cs.LG, cs.MA
- Luca Weihs, Papers: 20 Categories: cs.CV, cs.RO, cs.AI, cs.LG, cs.MA,
- Roozbeh Mottaghi, Papers: 13 Categories: cs.CV, cs.RO, cs.AI, cs.LG, cs.GR, cs.CL, cs.MA
- Ali Farhadi, Papers: 6 Categories: cs.RO, cs.AI, cs.CV, cs.LG, cs.GR,

Author's name: Li Fei-Fei

Categories: cs.AI, cs.CV, cs.RO, cs.LG, cs.GR, cs.HC

Number of ai papers on Arxiv: 10

Interests fields: Artificial Intelligence, Machine Learning, Computer Vision, Neuroscience

Jiajun Wu

Stanford University
Verified email at cs.stanford.edu - Homepage
Computer Vision Machine Learning Artificial Intelligence Cognitive Science

Silvio Savarese

Associate Professor of Computer Science at Stanford University
Verified email at stanford.edu - Homepage
Computer vision

Roberto Martín-Martín

Assistant Professor of Computer Science, The University of Texas at Austin
Verified email at cs.utexas.edu - Homepage
Robotics Artificial Perception Machine Learning Interactive Perception

Results

- Pushmeet Kohli, Papers: 3 Categories: cs.LG, cs.AI, stat.ML, cs.NE
- Yujia Li, Papers: 3 Categories: cs.CL, stat.ML, cs.AI, cs.CV, cs.LG, cs.NE
- Samuel J. Gershman, Papers: 3 Categories: cs.AI, cs.NE
- Jan Feyereisl, Papers: 3 Categories: cs.CV, eess.IV, stat.AP, cs.AI, cs.LG, stat.ML
- Jan N. van Rijn, Papers: 3 Categories: cs.LG, cs.AI, cs.MA, cs.OH, cs.LG, math.OC
- Yong Yu, Papers: 3 Categories: quant-ph, cs.AI, cs.CV
- Hans J. Briegel, Papers: 4 Categories: cs.SE, cs.LG, cs.PL
- Aftab Hussain, Papers: 3 Categories: cs.AI, cs.LG

Author's name: Yoshua Bengio

Categories: cs.LG, cs.AI, stat.ML, cs.CV, cs.DM, q-bio.NC, stat.CO, eess.AS, cs.CY, ...

Number of ai papers on Arxiv: 21

Interests fields: Artificial Intelligence, Machine Learning, Computer Vision, Neuroscience

Pushmeet Kohli

DeepMind

Verified email at google.com - Homepage

Machine Learning AI for Science Reliable and Trustworthy AI Computer Vision

Yujia Li

Other names

Research Scientist, Google DeepMind

Verified email at google.com - Homepage

Machine Learning Computer Vision Natural Language Processing Optimization

Samuel Gershman

Professor, Harvard University

Verified email at fas.harvard.edu - Homepage

Computational neuroscience cognitive science machine learning

Conclusions

It should be noted again that the embeddings were exclusively performed on the summaries of the papers. While conducting embeddings on the entire paper would have yielded more precise results, such an approach would have required a substantial amount of time.

Nonetheless, from an academic standpoint, we have successfully implemented BERT to find authors that have a similar way to write papers, in similar topics that had never collaborate. This was possible thanks to BERT's ability to capture the semantic meaning of words.

CADE - Compass Aligned Distributional Embeddings

CADE (Compass Aligned Distributional Embeddings) is a method for aligning distributional word embeddings across multiple corpus. It is designed to create cross-lingual word embeddings by leveraging parallel corpora. The goal of CADE is to produce embeddings that capture the semantic similarities and relationships between words across different environment.

This method uses a general corpus called compass which acts as a reference point for the alignment for the slices.

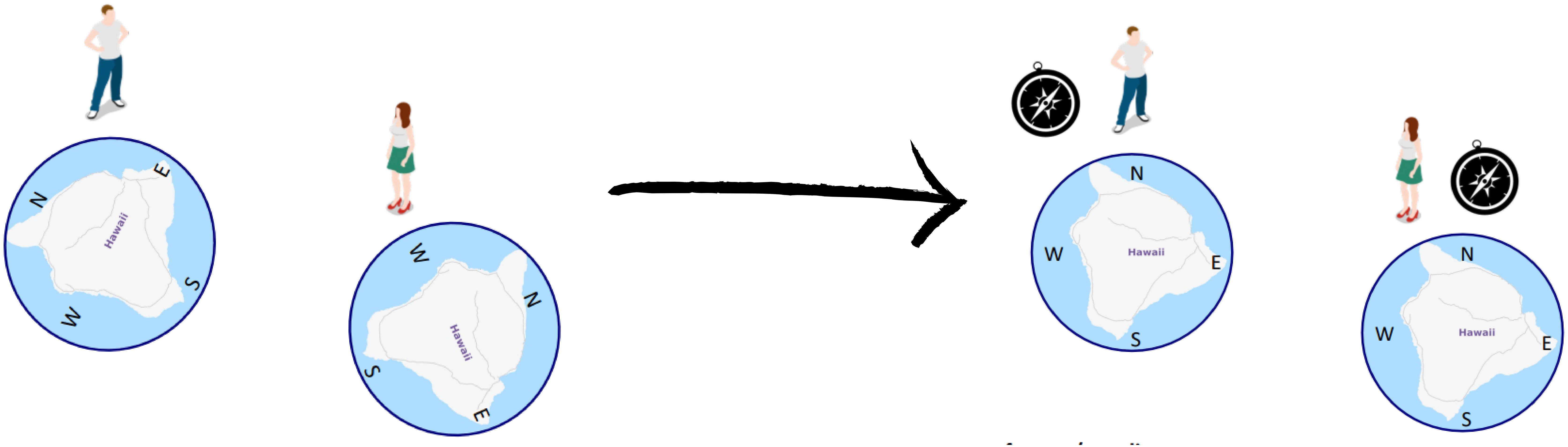
What we have to do with CADE

We made three compass text composed by different slices of the main corpus:

- first compass: slice computer science + slice biology slices
- second compass: slice computer science + slice astrophysics
- third compass: slice 1993_2018 + slice 2023

We found the 3 words most similar to the name of the field of each corpus with the ***model.wv.most_similar*** function
And we analyzed the vectorial difference that these had plus the word "artificial intelligence" in the two respective models using the function: ***1 - cosine(model.wv[word], model_bio.wv[word].***

The compass metaphor for the alignment problem



"Computer science"
most similar

Word	Similarity
HCI	0.877533
academic	0.858417
data_science	0.850618

"Biology"
most similar

Word	Similarity
psychology	0.90998
chemistry	0.891764
sciences	0.880741

Results

Model_1	Model_2	Word	Similarity between the two models
Computer science	Biology	artificial_intelligence	0.928258
Computer science	Biology	HCI	0.877673
Computer science	Biology	academic	0.905563
Computer science	Biology	data_science	0.884163
Computer science	Biology	psychology	0.955707
Computer science	Biology	chemistry	0.949927
Computer science	Biology	sciences	0.964024

"Computer science" *most similar*

Word	Similarity
academic	0.870069
industry	0.854898
economics	0.833945

"Astrophysics" *most similar*

Word	Similarity
robotic	0.867881
upcoming	0.85429
healthcare	0.847909

Results

Model_1	Model_2	Word	Similarity between the two models
computer_science	astrophysics	artificial_intelligence	0.90858
computer_science	astrophysics	academic	0.879875
computer_science	astrophysics	industry	0.862692
computer_science	astrophysics	economics	0.90615
computer_science	astrophysics	robotic	0.859473
computer_science	astrophysics	upcoming	0.843924
computer_science	astrophysics	healthcare	0.869497

"AI" on 1993_2018

most similar

Word	Similarity
Explainable_AI	0.803734
Machine_Learning	0.797711
Future	0.769789

"AI" on 2023

most similar

Word	Similarity
Chatbots	0.798417
Blockchain	0.791107
Designing	0.789808

Results

Model_1	Model_2	Word	Similarity between the two models
1993_2018	2023	artificial_intelligence	0.952275
1993_2018	2023	Chatbots	0.945083
1993_2018	2023	Blockchain	0.920149
1993_2018	2023	Designing	0.944458
1993_2018	2023	Explainable_AI	0.913639
1993_2018	2023	Machine_learning	0.94599
1993_2018	2023	Future	0.956867

SWEAT: Scoring Polarization of Topics across Different Corpora

To understanding differences of viewpoints across corpora we have used the SWEAT, a novel statistical measure of relative semantic polarization with respect to a given topic for pairs of textual corpora.
It aims to quantify the degree of divergence or disagreement between the discussions of specific topics in different contexts or sources.
The polarization score indicates the level of divergence or disagreement in the sentiment expressed towards a particular topic in different contexts and corpus.

What we have to do with SWEAT

We made three three SWEAT analysis composed by different slices of the main corpus:

- first SWEAT: slice computer science + slice math
- second SWEAT: slice financial + slice biology
- third SWEAT: slice 1993_2018 + slice 2023

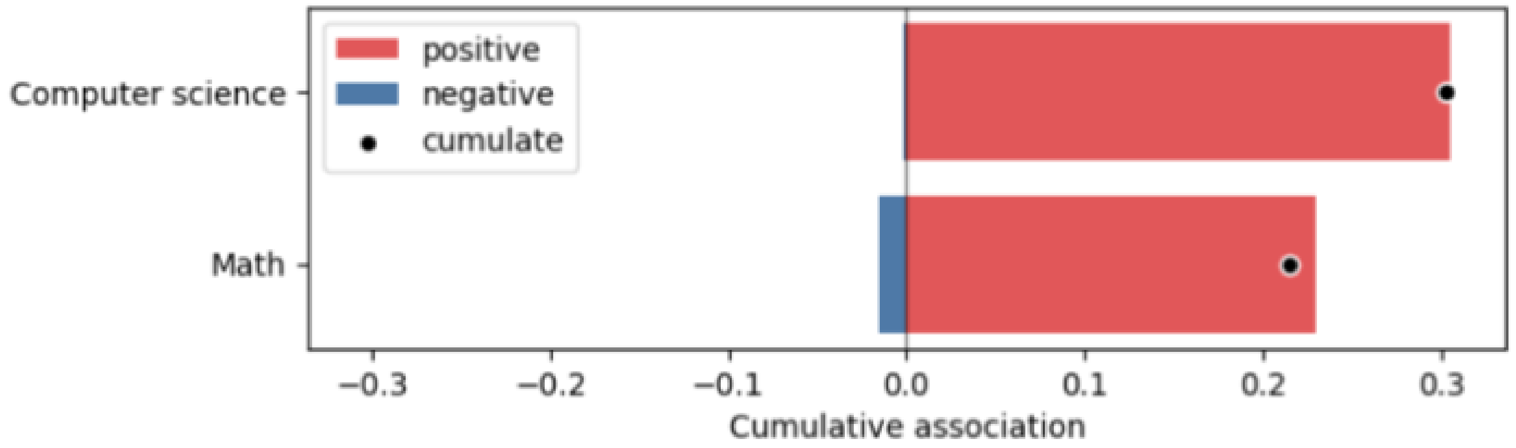
We used a sets of negative and positive words related at the words "articial intelligent" and for the topics of the analysis we chose two words most simlar at the field plus the word "artificial intelligence"

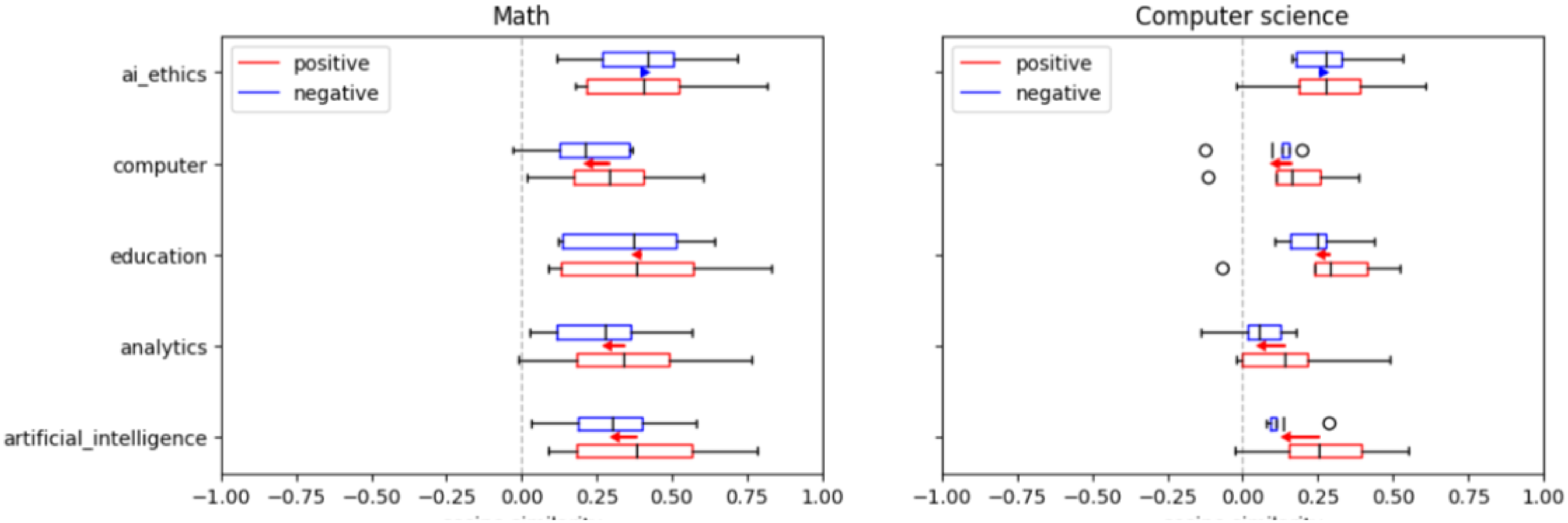
1. SWEAT computer_scince-math

topic = ['artificial_intelligence',"analytics", "education", "computer", "ai_ethics"]

artificial_intelligence	1615	4464
analytics	466	648
education	706	3910
computer	3614	3384
ai_ethics	143	957

```
Math ~ negative
Computer science ~ positive
{'score': -0.0884, 'eff_size': -0.4424, 'p-val': 0.555}
```





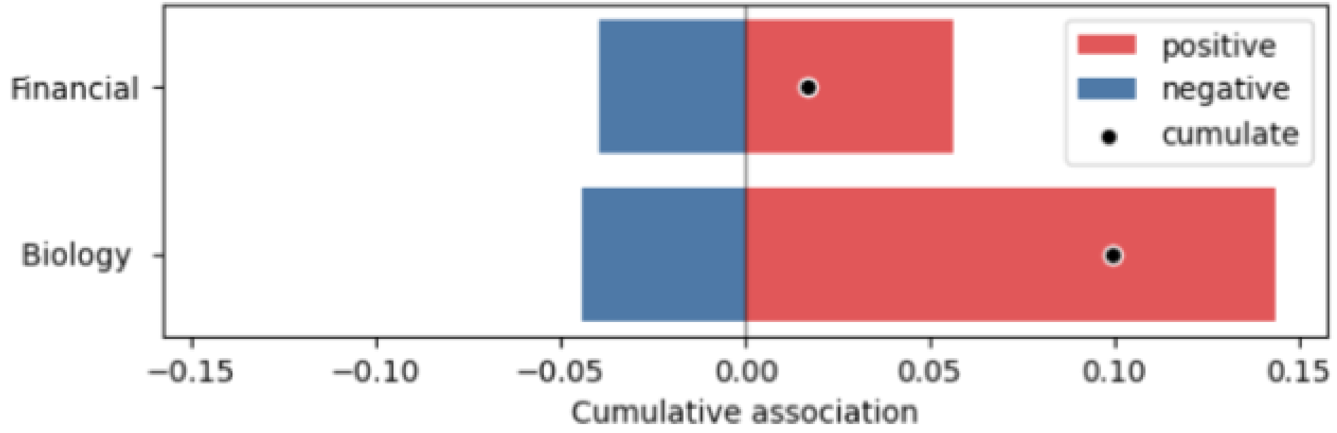
2.SWEAT financial-biology

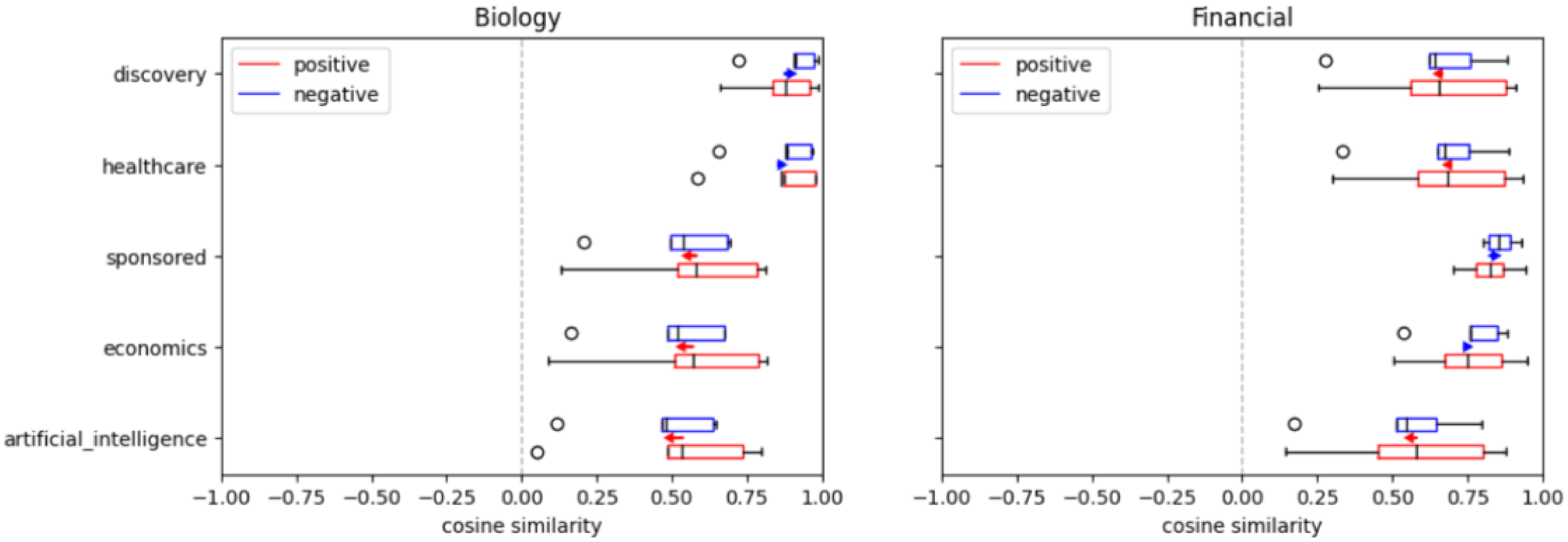
topic = ['artificial_intelligence',"economics", "sponsored", "healthcare", "discovery"]

```
                            ['future', 'strong', 'solver', 'positive', 'innovative'] ['unreliable', 'difficult', 'issue', 'unethical', 'negative']

artificial_intelligence    587              1066
economics                  32               585
sponsored                  4                7
healthcare                 309              101
discovery                  663              67

Biology  ~ positive
Financial ~ negative
{'score': 0.0824, 'eff_size': 0.5662, 'p-val': 0.4288}
```





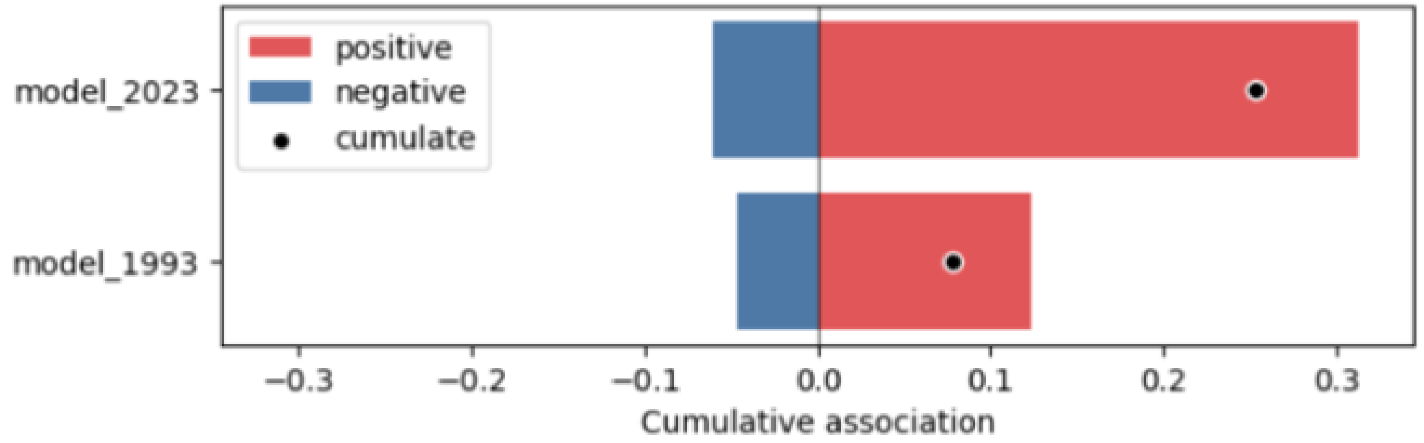
3.SWEAT 1993_2018 - 2023

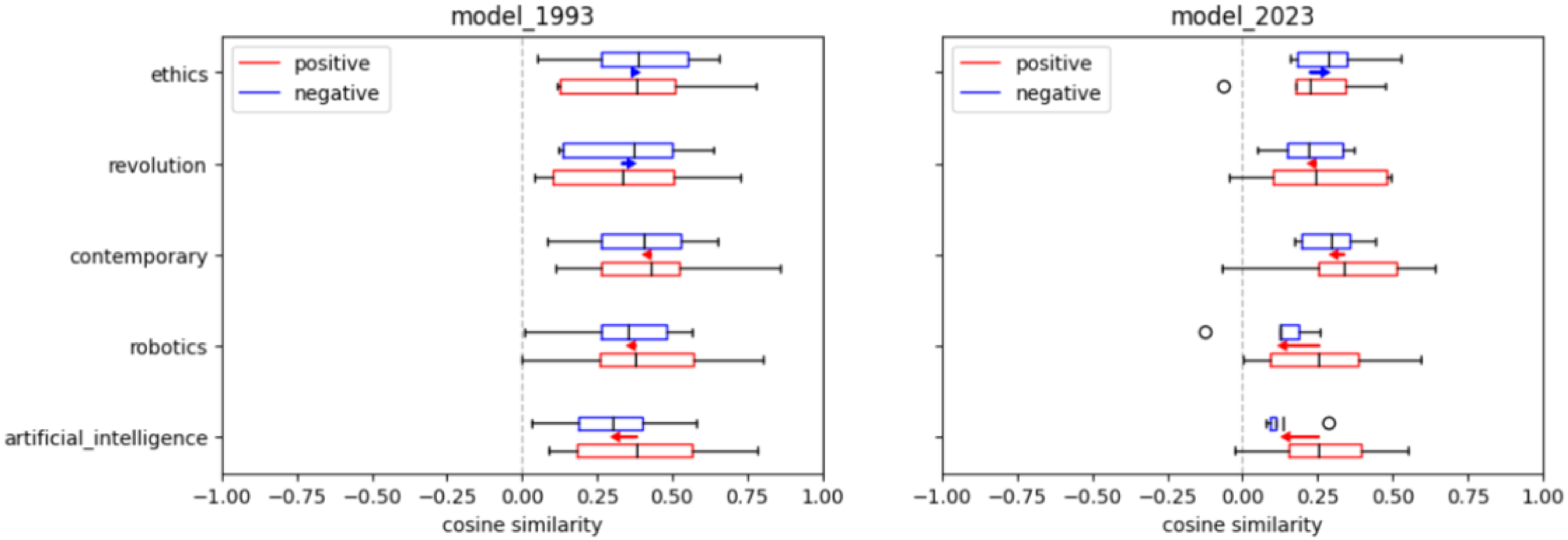
topic = ['artificial_intelligence',"robotics", "contemporary", "revolution", "ethics"]

```
                         ['future', 'strong', 'solver', 'positive', 'innovative'] ['unreliable', 'issue', 'problem', 'unethical', 'negative']

artificial_intelligence  1615            4464
robotics                 1252            1174
contemporary             257             302
revolution               209             249
ethics                   652             1970

model_1993 ~ negative
model_2023 ~ positive
{'score': -0.1756, 'eff_size': -0.6222, 'p-val': 0.3901}
```





Future developments

- **Refinement of Paper Retrieval:** Incorporate advanced algorithms to improve the quality of paper retrieval.
- **Increasing the Dataset:** Enrich the analysis by incorporating more papers into the database.
- **Enhanced Data Structuring:** Improve data organization with more comprehensive category slices.
- **Temporal Analysis Enhancement:** Increase granularity of time-based analysis to understand AI evolution in finer detail.
- **Expanded Use of BERT:** Extend BERT application to include full-text embedding for more accurate semantic understanding.
- **Authors same name recognition:** Detect and provide an id for authors with the same name

References

Compass-aligned Distributional Embeddings for Studying Semantic Differences across Corpora
Matteo Palmonari, Federico Bianchi,Valerio Di Carlo,Paolo Nicoli

Efficient Estimation of Word Representations in Vector Space
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

Attention Is All You Need
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

Modern hierarchical, agglomerative clustering algorithms
Daniel Müllner

SWEAT: Scoring Polarization of Topics across Different Corpora
Anonymous EMNLP submission