Universidade de Brasília
Faculdade de Administração, Contabilidade, Economia e Gestão Pública
Departamento de Economia

Innovation and Word Usage Patterns in Machine Learning

Vítor Bandeira Borges Borges

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciências Econômicas

Orientador
Prof. Dr. Daniel Oliveira Cajueiro

Brasília
2024

Universidade de Brasília
Faculdade de Administração, Contabilidade, Economia e Gestão Pública
Departamento de Economia

Innovation and Word Usage Patterns in Machine Learning

Vítor Bandeira Borges Borges

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciências Econômicas

Prof. Dr. Daniel Oliveira Cajueiro (Orientador)
ECO/UnB

Prof. Dr. Victor Rafael Rezende Celestino
ADM/Unb

Prof.a Dr.a Daniela Freddo
Coordenadora do Bacharelado em Ciências Econômicas

Brasília, 19 de fevereiro de 2024

Dedicatória

Dedico este trabalho ao Laboratório de Aprendizado de Máquina em Finanças e Organizações (LAMFO), cuja comunidade de alunos, professores e profissionais dedicados tem sido fundamental na minha trajetória acadêmica desde o terceiro semestre. No LAMFO, encontrei não apenas um espaço de aprendizado e inovação, mas também uma família acadêmica que me acolheu, desafiou e inspirou a superar limites e expandir horizontes.

Agradecimentos

A realização deste trabalho marca a conclusão de uma jornada desafiadora e enriquecedora em minha formação acadêmica e pessoal. No entanto, este caminho não teria sido possível sem o apoio incondicional e amor de pessoas muito especiais.

Gostaria de expressar minha mais profunda gratidão aos meus pais, Marcos Aurélio Perciano Borges e Ana Cristina Bandeira Borges, cujo apoio, compreensão e incentivo foram fundamentais em cada passo desta jornada. Vocês foram minha fonte de inspiração e motivação para superar os desafios e perseguir meus objetivos com determinação. A dedicação e o amor que vocês me demonstraram não só moldaram quem sou como pessoa, mas também contribuíram imensamente para a realização deste trabalho.

A vocês, minha eterna gratidão e amor.

Resumo

Neste estudo, investigamos o cenário dinâmico da evolução da pesquisa em aprendizado de máquina. Inicialmente, por meio do uso da alocação de Dirichlet latente, determinamos temas essenciais e conceitos fundamentais que surgiram no âmbito do aprendizado de máquina. Em seguida, realizamos uma análise abrangente para rastrear as trajetórias evolutivas desses temas identificados. Para quantificar a novidade e a divergência das contribuições de pesquisa, usamos a métrica de Divergência de Kullback-Leibler. Essa medida estatística serve como um indicador de "surpresa", indicando a extensão da diferenciação entre o conteúdo dos artigos acadêmicos e os desenvolvimentos subsequentes na pesquisa. Também analisamos as funções de pesquisadores proeminentes e a importância dos locais acadêmicos (periódicos e conferências) no campo da aprendizagem automática.

Palavras-chave: Descoberta de conhecimento, divergência de Kullback-Leibler, alocação de Dirichlet latente, aprendizado de máquina, processamento de linguagem natural

Abstract

In this study, we investigate the dynamic landscape of machine learning research evolution. Initially, through the use of Latent Dirichlet Allocation, we determine pivotal themes and fundamental concepts that have emerged within the realm of machine learning. Subsequently, we undertake a comprehensive analysis to track the evolutionary trajectories of these identified themes. To quantify the novelty and divergence of research contributions, we use the Kullback-Leibler Divergence metric. This statistical measure serves as a proxy for “surprise", indicating the extent of differentiation between the content of academic papers and the subsequent developments in research. We also analyze the roles of prominent researchers and the significance of academic venues (journals and conferences) in the field of machine learning.

Keywords: Knowledge discovery, Kullback-Leibler Divergence, Latent Dirichlet Allocation, Machine Learning, Natural Language Processing

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introduction

Machine learning has become ubiquitous in many fields today, ranging from economics, finance, medicine, and healthcare to marketing and transportation. It enables businesses and organizations to extract valuable insights from large amounts of data and make predictions based on patterns and trends that would otherwise be difficult or impossible to detect. By extracting valuable information from data and automating several tasks, machine learning can save time and reduce costs while improving accuracy and efficiency.

This paper explores the evolution of research on machine learning, using a large collection of papers in the field. We begin by identifying the main topics and dominant concepts within each area. Next, we trace the emergence of these key concepts and analyze how they have evolved over time. Finally, we examine the relationships between the current state-of-the-art and past developments, providing insight into the ways in which the field has progressed. In addition, we investigate the importance of the main machine learning venues and the more frequent and prominent authors in disseminating the theme.

A fundamental building block of our work is the Latent Dirichlet Allocation (LDA) model [1]. We use it to identify the main topics and concepts of the field of machine learning. Associated with the application of this technique we also use a coherence metric [2] to indicate the suitable number of the topics of the model. With the papers divided by topics, we are able to analyze the trends of the machine learning field. In addition, we use the Kullback-Leibler Divergence (KLD) as the notion of "surprise" that measures the statistical divergence between the contributions of papers in subsequent events. We can use surprise to measure the divergence of one piece of news from previous ones called "Novelty" and the divergence of one piece of news from the later ones called "Transience". With these concepts in hand, we are able to define the concept of "Resonance" as the difference between Novelty and Transience [3]. Thus, we may evaluate the role of the most relevant venues used to disseminate the knowledge of machine learning and the role of the most frequent and prominent researchers.

Our data comes from 25 very relevant machine learning venues that include annals of conferences and prestigious journals. The method we use to choose these venues is based on a two step procedure. In the first step, we look into a series of popular machine learning blogs for the most popular indications. In the second step, we check if these indications belong to the list of the top venues based on the Google scholar h5-index considering the subcategories of "Artificial Intelligence", "Computational Linguistics", "Data Mining and Analysis" and "Engineering and Computer Science".

Our work relates to other studies that tell the history and the evolution of the machine learning field such as [4] and [5]. It also naturally relates to the work of [3] that uses the concepts of novelty, transience and resonance to study how ideas are created, ignored or propagated in the context of the French revolution. In this same context, we may also cite [6] that also applied LDA to to the ACL Anthology to analyze historical trends in the field of Computational Linguistics. In addition, [7] proposes a a LDA based method to estimate a innovation score of a given paper.

Our work is organized as follows. Section 2 describes the procedures we adopt to tune and estimate the models. We detail the data set we use in Section 3 and present the results in Section 4. Section 5 summarizes and concludes the work.

Capítulo 2

Methods

We have divided this section into three segments. In Section 2.1, we provide an overview of the LDA model and the methodology for determining the optimal number of topics. Moving on to Section 2.2, we demonstrate the application of the LDA-derived topics in monitoring the progression of machine learning research. Further in Section 2.3, we revisit the KLD metric, elucidating its utility in delineating the constructs of novelty, transience, and resonance.

2.1 Latent Dirichlet Allocation

The LDA model, introduced by [1], is a generative probabilistic model for topic modeling. It is based on the assumption that documents are mixtures of topics, and each topic is a distribution over words. In order to be mathematically precise, we use the following notation. The vocabulary $\mathcal{V} = \{w_1, \ldots, w_i, \ldots, w_{N_V}\}$ is the set of all distinct words (present in all documents) and $I_V = \{1, \ldots, N_V\}$ is the set of all word indexes, where N_V is the number of distinct terms, i.e., the size of the vocabulary. A document $d = [w_{i_1}, \ldots, w_{i_k}, \ldots, w_{i_{N_d}}]$ consist of a list of N_d non-unique consecutive words ($1 \leq k \leq N_d$ and $i_k \in I_V$), while $\mathcal{V}^d$ is the vocabulary that appears in the document d. Based on the assumption that the number of topics K is fixed and known, LDA assumes the following generative process for each document d in a corpus D:

1. Choose $N_d \sim \text{Poisson}(\xi)$;
2. Choose $\theta \sim \text{Dir}(\alpha)$, where the parameter α is K vector of positive components that we need to estimate;
3. For each of the N_d words of w_n;
   (a) Choose topic $z_n \sim \text{Multinomial}(\theta)$;

(b) Choose word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned to topic z_n, where the word probabilities are parametrized by a $K \times N_V$ matrix $\beta = p(w^j = 1|z^i = 1)$ for all $j \in \{1, ..., N_V\}$ and all $k \in \{1, ..., K\}$.

In the first step, LDA chooses the number of words in the document, denoted by N_d, from a Poisson distribution with parameter ξ. This determines the length of the document. Then, it chooses the document's topic proportions, denoted by θ, from a Dirichlet distribution with parameter α. This step determines the distribution of topics within the document. Thus, for each of the N_d words in the document, it chooses the word's topic, denoted by z_n, from a Multinomial distribution with probabilities determined by the document's topic proportions θ, which determines which topic the word belongs to. And, finally, it chooses the specific word, denoted by w_n, from a Multinomial distribution conditioned on the chosen topic z_n, where, as above-mentioned, the word probabilities are parameterized by a $K \times N_V$ matrix β, where N_V is the vocabulary size. This generative process is repeated for each document in the collection. There are different ways to estimate this model. In our paper, we estimate it using the online variational Bayes algorithm due to [8][1].

Selecting an appropriate number of topics stands as a fundamental prerequisite for executing the LDA model. Notably, it is imperative to acknowledge that evaluating the efficacy of LDA, akin to other unsupervised models, presents challenges stemming from the absence of labels that can serve as benchmarks to validate the accuracy of outcomes. While the most effective approach to appraising unsupervised models involves human assessments, such an evaluation methodology can incur substantial costs and, in cases of extensive datasets, may even become unfeasible. Consequently, within this contextual framework, a prevalent recourse involves the utilization of metrics that capture the frequency of co-occurrences within a given corpus. These metrics find application within the domain of LDA, hinging upon the identification of the words per topic and the analysis of their co-occurrences within the corpus. In our paper, we use the Normalized Pointwise Mutual Information (NPMI) [10] coherence measure . Let the Pointwise Mutual Information (PMI) be given by [11]

$$PMI(w_i, w_j) = log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \tag{2.1}$$

where $P(w_i, w_j)$ is the joint probability of words w_i and w_j measured in a fixed-size window in the text, $P(w_i)$ and $P(w_j)$ are the individual probabilities of the words and ϵ is a small number added to the joint probability to avoid logarithm of zero. Thus, PMI is a measure of how much the actual probability of a particular co-occurrence of words

[1]We use the implementation available in the Gensim Python library [9].

$p(w_i, w_j)$ differs from what we would expect it to be on the basis of the probabilities of the individual words and the assumption of independence $p(w_i)p(w_j)$.

The NPMI is a normalized form of the PMI measure. Although there are different ways to normalize the PMI, [10] normalizes it by the $(-log(P(w_i, w_j) + \epsilon))$, since this option normalizes both the upper and the lower bound. Thus, we may write the NPMI by

$$NPMI(w_i, w_j) = \left(\frac{PMI(w_i, w_j)}{-log(P(w_i, w_j) + \epsilon)} \right). \tag{2.2}$$

In order to quantify how semantically related the words within a topic are, we may evaluate the coherence of a topic T_k using

$$C_V(T_k) = \frac{2}{|T_k|(|T_k| - 1)} \sum_{i=1}^{|T_k|-1} \sum_{j=i+1}^{|T_k|} NPMI(w_i, w_j), \tag{2.3}$$

where $|T_k|$ is the number of words in the topic T_k.

Aiming at considering the quality of all the topics together, we average C_V to get

$$\overline{C_V} = \frac{1}{K} \sum_{k=1}^{K} C_V(T_k), \tag{2.4}$$

where K is the number of topics. An important characteristic of this coherence measure is its high correlation with human judgment in assessing the quality of topics [2].

2.2 Modelling trends

As in [6], in order to capture the temporal dynamics among topics, we evaluate the observed probability of each topic within specific time intervals. This probability assessment involves calculating the average likelihood of each topic across the papers published during that period. This process is repeated for all topics across all distinct time periods under consideration.

2.3 Kullback Leibler Divergence

The Kullback-Leibler Divergence (KLD) [12], also known as relative entropy, is a measure of information loss when an *observed* probability distribution p is estimated using a *theoretical* distribution q. If the observed and theoretical distributions are the same ones, the divergence is zero. On the other hand, if we consider two vastly different distributions, the divergence is very high, meaning a great loss of information due to misspecification.

In the context of topic modeling, we can use KLD to quantify the dissimilarity between a document's topic distribution and a reference topic distribution. From an information retrieval perspective, we may interpret relative entropy as a measure of "surprise" when one document is expected and another is observed [3]. Given an LDA-generated set of probability distributions $p^{(j)} = (p_1^{(j)}, p_2^{(j)}, \ldots, p_K^{(j)})$, where j indexes chronological order and K is the number of topics, we may evaluate the *surprise* between times j and i as

$$\text{KLD } (p^{(j)}|p^{(i)}) = \sum_{k=1}^{K} p_k^{(j)} \log_2 \frac{p_k^{(j)}}{p_k^{(i)}}, \tag{2.5}$$

where K, as before, is the number of topics[2].

We may define the *novelty* $\mathcal{N}_w(j)$ of the j-th document by the average surprise between itself and the past documents that took place in a time scale w:

$$\mathcal{N}_w(j) = \frac{1}{w} \sum_{d=1}^{w} \text{KLD } (p^{(j)}|p^{(j-d)}). \tag{2.6}$$

On the other hand, we may define the *transience* $\mathcal{T}_w(j)$ of the j-th document by the average surprise between itself and the future documents that will take place in a time scale w:

$$\mathcal{T}_w(j) = \frac{1}{w} \sum_{d=1}^{w} \text{KLD } (p^{(j)}|p^{(j+d)}). \tag{2.7}$$

We measure *resonance* $\mathcal{R}_w(j)$ as the difference between novelty and transience:

$$\mathcal{R}_w(j) = \mathcal{N}_w(j) - \mathcal{T}_w(j). \tag{2.8}$$

We may interpret the resonance of a document in a corpus of news stories as an indicator of a novel subject that is capable of influencing the general direction of outlets, being written about again in the future.

In addition, we may measure the expected resonance of any document given some level of novelty with a linear model

$$E[\mathcal{R}|\mathcal{N}] = \beta_{\text{int}} + \beta_{\mathcal{N}}\mathcal{N} \tag{2.9}$$

[2]It is worth noting that, unlike the paper by [3], our approach does not adhere strictly to a chronological ordering of the papers. Instead, we arrange the papers in chronological order by month, which serves as the temporal unit allowing us to reconstruct the paper sequence based on the publication dates provided by the academic venues. In addition, to assess the novelty and transience (that we will define below), we compute these measures for each paper relative to all papers published in the preceding period and calculate the average value.

and, using this linear equation, we may define *novelty effectiveness* Γ as the rate at which resonance increases with novelty:

$$\Gamma = \frac{\partial E[\mathcal{R}|\mathcal{N}]}{\partial \mathcal{N}} = \beta_{\mathcal{N}}. \tag{2.10}$$

Novelty effectiveness provides a nuanced understanding of the dynamics of speech influence. It highlights the delicate balance speakers must strike between novelty and resonance, and the inherent risk and reward associated with introducing novel ideas.

The time period parameter for calculating the average innovation between papers, denoted as w, was set to be equal to 12 months.

Capítulo 3

Dataset

Our dataset consists of 25 venues related to machine learning, including both conference proceedings and periodicals. We choose these venues using a two step procedure. In the first step, we look into a series of popular machine learning sources for the most popular indications presented in Appendix A. In the second step, we check if these indications belong to the list of the top 20 publications based on the Google Scholar h5-index[1] considering the subcategories of "Artificial Intelligence", "Computational Linguistics", "Data Mining and Analysis" and "Engineering and Computer Science". Of these, 24 venues were indexed on the Web of Science (WoS) database. The International Conference on Learning Representations, however, was not accessible in WoS, necessitating manual extraction from the [13] API.

The dataset, comprising 168,757 publications, serves as the foundation for this research, which aims to scrutinize abstracts and their interrelationships throughout the field's history. The dataset includes 95,626 (56.66%) papers in academic periodicals, 72,188 (42.78%) conference papers, 940 publication series, and 3 books. Among these papers, we are not able to use 4,001 of them because they do not have abstracts. We also exclude from our study 3,750 publications that we are not able to recover the date. It is worth mentioning that in most cases, the date of the publication is directly available in the data extracted from the WoS. However, in some specific cases we have to directly deal with it. In particular, in some conferences, the date of publication was not available and we replaced this missing value by the date that the conference took place. In a small number of cases, the description of the date of the conference was given by the station of the year, namely *Summer*, *Autumn*, *Winter* and *Spring*. In these cases, we carefully looked into the correct date of the publication and replaced this information by it. Due to the lack of precision of this piece of data, we adopt the monthly granularity for our time series.

[1]This Google scholar tool is available at https://scholar.google.com/citations?view_op=top_venues.

The exponential growth of publications in our database is illustrated in Figure 3.1. This remarkable trend mirrors the recent upsurge of interest and resources devoted to machine learning [14], which has facilitated the swift generation and obsolescence of innovative concepts. The expansion of publications in both number and velocity is indicative of the dynamic nature of this field, where novel findings and ideas emerge at a rapid pace.

Figura 3.1: Number of Publications per Year.

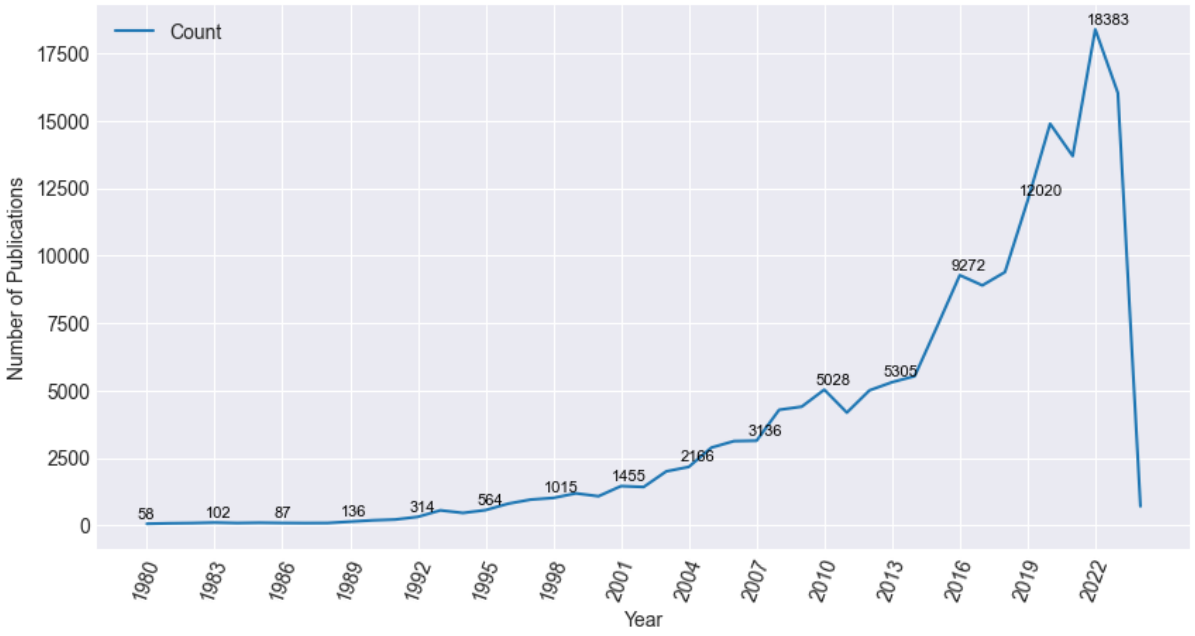


Table 3.1 offers a summary of the dataset, detailing the number of papers, accessible date ranges, and predominant publication types by venue. The complete dataset and associated code can be accessed through this paper's Zenodo.

Tabela 3.1: Summary of venues, counts, date ranges, and publication types in the dataset

Venue	Count	Date Range	Primary Publication Type
Neurocomputing	17830	Feb-1992 to Jan-2023	Journal
Expert Systems with Applications	17273	Dec-1992 to Apr-2023	Journal
International Conference on Machine Learning	15785	Oct-1997 to Dec-2022	Conference
International Conference on Computer Vision (ICCV)	12051	Jun-1995 to Feb-2022	Conference
AAAI Conference on Artificial Intelligence	9246	Nov-2008 to Feb-2021	Conference
Applied Soft Computing	8369	Feb-2004 to Nov-2022	Journal
Neural Computing and Applications	8296	Sep-1997 to Dec-2022	Journal
IEEE Transactions on Pattern Analysis and Machine Intelligence	7226	Jan-1986 to Dec-2022	Journal
Knowledge-Based Systems	6767	Mar-1991 to Jan-2023	Journal
Neural Information Processing Systems	6020	Dec-1992 to Dec-2020	Conference
Computer Vision and Pattern Recognition Conference	5786	Jun-2000 to Jun-2019	Conference
Meeting of the Association for Computational Linguistics (ACL)	5771	Jun-1993 to May-2022	Conference
International Joint Conferences on Artificial Intelligence Organization	5291	Jul-2005 to Sep-2020	Conference
Neural Networks	5078	Feb-1990 to Jan-2023	Journal
IEEE Transactions on Neural Networks and Learning Systems	5059	Jan-2012 to Dec-2022	Journal
Applied Intelligence	4706	Feb-1993 to Dec-2022	Journal
Engineering Applications of Artificial Intelligence	4515	Jan-1992 to Jan-2023	Journal
International Conference on Learning Representations	4353	May-2013 to Apr-2022	Conference
European Conference on Computer Vision	4124	Sep-1997 to Mar-2020	Conference
IEEE Transactions on Fuzzy Systems	3766	Feb-1994 to Dec-2022	Journal
Journal of Machine Learning Research	2657	Oct-2000 to Dec-2015	Journal
International Conference on Artificial Intelligence and Statistics	2526	Apr-2014 to Mar-2022	Conference
Conference on Empirical Methods in Natural Language Processing	2479	Feb-1999 to Nov-2021	Conference
IEEE Transactions on Systems, Man, and Cybernetics, Part B	2104	Feb-1996 to Dec-2012	Journal
Artificial Intelligence Review	1679	Feb-1993 to Dec-2022	Journal
Total	168757	Jan-1986 to Apr-2023	Journal

Capítulo 4

Results

AIn this section, we present our results. In Subsection 4.1, we present the discovered topic trends uncovered during our study. In Subsection 4.2, we delve into the assessment of novelty, transience, and resonance as key characteristics of machine learning research. Here, we examine the roles of authors and venues in shaping this field, evaluating their impact and influence. We present the details of our LDA implementation in Appendix B.

4.1 Topics Trends

The dynamics of scientific progress and the elements influencing the ascent and descent of academic interest in diverse subjects have been extensively debated among historians, sociologists, philosophers of science, and scientists themselves [15]. By reducing a corpus of scientific documents to a set of topics, we can enhance our understanding of the development of scientific pursuits and the driving forces behind these shifts.

In the following subsections, we utilize LDA and observed probability of each topic to extract the trends of specific relevant topics in the field of machine learning, including deep learning (Section 4.1.1), computer vision (Section 4.1.2), natural language processing (Section 4.1.3), reinforcement learning (Section 4.1.4), and expert systems (Section 4.1.5). We conclude this section with Section 4.1.6, which examines the potential impact of certain real-world events on machine learning research.

4.1.1 Deep Learning

As a prominent subfield of machine learning, deep learning focuses on the design and application of artificial neural networks, particularly those with multiple hidden layers, to address complex computational problems. Influential researchers in deep learning, such as [16], [17], and [18], have been instrumental in the development of the field.

This powerful approach has significantly propelled advancements in various areas of machine learning, such as computer vision [19], natural language processing [20], and speech recognition [21], by enabling the extraction of hierarchical features and promoting the development of end-to-end learning systems.

Figura 4.1: Deep Learning Related Topics

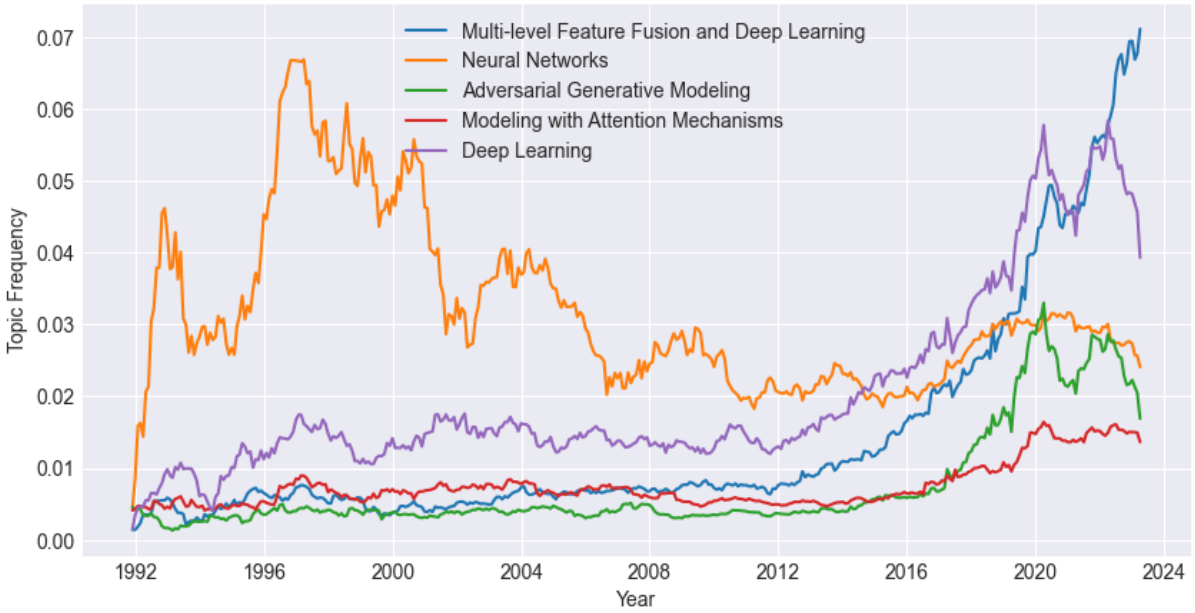


Figure 4.1 illustrates the frequency of Deep Learning related topics over the past 30 years, clearly demonstrating the substantial evolution of the subfield within the last five to eight years. This figure exemplifies [22] model of scientific evolution, which posits that a community's adoption of a new paradigm triggers a shift in focus, provoking debates and promoting advancements in novel areas. In the subfield's literature, the prevailing paradigm transitioned from *Neural Networks* — a topic that primarily emphasized data representations and the design of network architectures for capturing features within data — to subjects that focus on improving model performance through novel training techniques, optimization algorithms, and architectural innovations.

Comparatively recent approaches, such as *Adversarial Generative Modeling* (AGM) [23] and *Modeling with Attention Mechanisms* [24], experienced an upsurge in a more condensed timeframe than *Deep Learning*. The foundational work of Hinton, LeCun, and Bengio influenced the surge of publications in *Multi-level Feature Fusion and Deep Learning*, which subsequently led to the emergence of AGM and Attention. These subjects encompass various aspects of model training, such as discovering better optimization

methods, understanding the benefits of depth in neural networks, regularization techniques, and novel architectures that facilitate learning in complex domains.

4.1.2 Computer Vision

Computer vision, a multidisciplinary subfield of machine learning, focuses on enabling machines to interpret and comprehend visual information from their surroundings. Drawing on techniques from image processing [25], pattern recognition [26], and statistical learning [27], it plays a crucial role in artificial intelligence by empowering systems to interact with and make sense of the visual world, facilitating applications in robotics, surveillance, healthcare, and autonomous vehicles.

During the 1990s, computer vision techniques were primarily based on rigorous mathematical analysis and quantitative aspects. Examples of models from this period include the concept of scale-space [28], contour models known as snakes [29], and projective 3-D reconstructions [30]. Researchers also utilized optimization frameworks such as regularization [31] and Markov random fields [32]. In addition, statistical learning techniques, like Eigenface [33], were employed for facial recognition in images. However, these traditional methods relied on handcrafted feature extraction and shallow models, often struggling to generalize and capture complex patterns in visual data.

The advent of deep learning revolutionized computer vision in recent years, significantly advancing performance and capabilities. Convolutional neural networks (CNNs) [34] have enabled automatic learning of hierarchical representations from raw images, bypassing manual feature engineering. This success has been further bolstered by the progress in GPU computing power, which allows for efficient training of increasingly complex and deep models. As a result, deep learning-based computer vision systems have achieved unprecedented success in tasks like object recognition [19], semantic segmentation [35], and image generation [36], surpassing human-level performance in certain benchmarks and enabling practical applications across various sectors.

In this study, we manually subdivided topics into "Groups" or subfields of machine learning, as shown in Tables C.1 and C.2. The evolution of the *Deep Learning Group* is compared to the *Computer Vision Group* in Figure 4.2. The two series exhibit a statistically significant negative Pearson correlation coefficient of -0.61749, suggesting a shift in the scientific community's preference.

Figura 4.2: Computer Vision vs. Deep Learning

During the 1990s, computer vision constituted between a third and a quarter of all publications, while in recent times, it represents merely 10%. This negative relationship can be attributed to deep learning's capability to automatically learn hierarchical feature representations from raw data, outperforming traditional techniques reliant on manual feature engineering. Consequently, the research focus has shifted toward data-driven methods, leading to a decline in the proportion of theoretical computer vision publications in the field.

4.1.3 Natural Language Processing

The 1950s marked the beginning of NLP as a subfield of artificial intelligence. Alan Turing's test, which involved the automated interpretation and generation of natural language, laid the groundwork for the field [37]. At this stage, a fundamental development is the classical and sparse n-grams model, which serves as a precursor to the contemporary large language models we are familiar with today [38]. Some decades later, we may cite the research on information retrieval that developed techniques such as TF-IDF (Token Frequency-Inverse Document Frequency) [39, 40]. In the 1980s, NLP shifted towards statistical and machine learning algorithms, driven by increased computational power, the contributions in the field of informational retrieval and the decline of Chomsky's Transformational Grammar linguistic theories [41].

The 2000s saw a surge in available raw, unannotated language data, prompting a focus on unsupervised and semi-supervised learning algorithms. At this time, different data-driven approaches were applied to deal with important machine learning tasks. Among them, we may cite the matrix factorization based methods [42], the graph based methods [43, 44], and the topic modeling based methods [1].

Since 2015, NLP has shifted from statistical methods to neural networks, streamlining feature engineering. Techniques such as word embeddings, end-to-end learning of higher-level tasks, and the use of Long Short-Term Memory (LSTM) networks [45] have gained popularity, leading to significant changes in NLP system design. Deep neural network-based approaches now represent a new paradigm, distinct from statistical natural language processing.

One significant development in NLP is the introduction of attention mechanisms [24], which have improved the performance of models by allowing them to focus on specific parts of input sequences while processing information. The groundbreaking work by [46], "Attention is All You Need", introduced the Transformer architecture, which has revolutionized NLP. Transformers leverage self-attention mechanisms to process input sequences in parallel, rather than sequentially, resulting in improved efficiency and performance, as illustrated in Figure 4.1. Since then, Transformer-based models, such as Google's BERT [47], OpenAI's GPT-series [48, 49], and numerous other variations, have consistently achieved state-of-the-art results in various NLP tasks, transforming the field and its applications.

Figura 4.3: Natural Language Processing Group

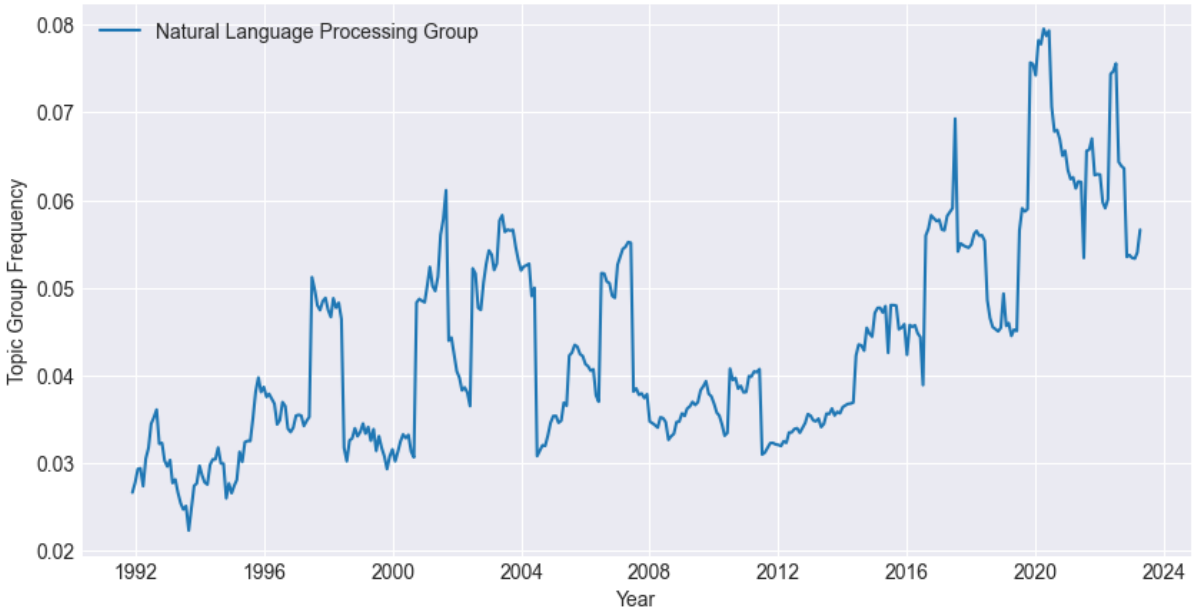


Figure 4.3 illustrates the evolution of NLP, highlighting that, in contrast to the *Computer Vision Group*, the *Natural Language Processing Group* has experienced a surge in frequency since the 1990s. This is further reinforced by its statistically significant positive Pearson correlation coefficient of 0.7426 with the *Deep Learning Group*, indicating a strong association between the growth of NLP and the advancements in deep learning techniques.

4.1.4 Reinforcement Learning

Reinforcement Learning (RL) is a subfield of machine learning that focuses on training intelligent agents to make optimal decisions by interacting with their environment. In contrast to supervised learning, which relies on labeled data to learn from, RL is inspired by the trial-and-error learning process observed in humans and animals. The primary components of a reinforcement learning system are an agent, an environment, states, actions, and rewards.

In RL, an agent observes the current state of the environment and takes an action based on its internal policy. The policy, represented by a function, maps states to actions, determining the agent's behavior. After performing an action, the agent receives feedback in the form of a reward signal from the environment. The goal of the agent is to maximize its cumulative reward over time, which requires finding an optimal balance between exploration (trying new actions) and exploitation (relying on actions that have been successful in the past).

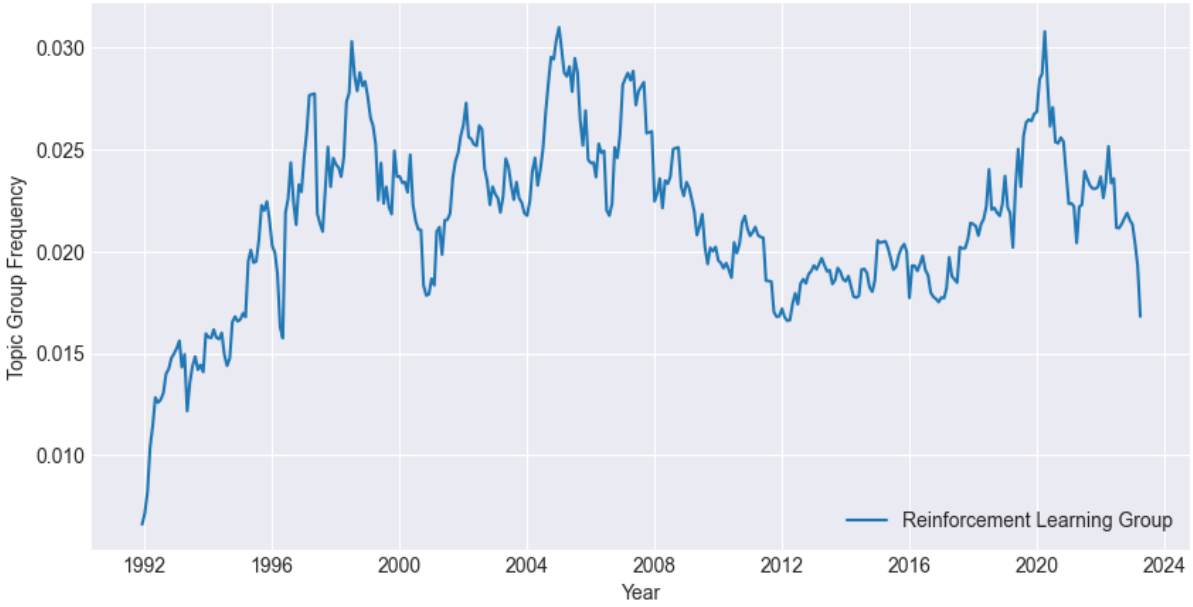
The 1990s marked a significant period in the growth of Reinforcement Learning (RL), with groundbreaking advancements shaping the field's trajectory. Q-learning, introduced by Chris Watkins in 1989 [50], emerged as a key model-free RL algorithm that learns optimal policies without explicitly modeling the environment's dynamics. Furthermore, Richard Sutton's development of Temporal Difference (TD) Learning [51] combined dynamic programming and Monte Carlo methods, allowing agents to learn directly from experience.

The exploration of function approximation methods, including neural networks, enabled RL algorithms to tackle problems with large state and action spaces. These pivotal developments in the 1990s propelled RL into prominence and solidified its importance in subsequent decades. The 2000s, 2010s, and 2020s witnessed consistent progress in RL, with major breakthroughs such as Deep Q-Networks (DQN) [52] and AlphaGo [53] demonstrating the power and versatility of RL algorithms in solving complex real-world problems.

Figure 4.4 illustrates the substantial impact of the 1990s' pivotal advancements in Reinforcement Learning (RL) on the field's enduring significance. Notably, the figure

reveals a marked surge in RL publication frequency after the transformative innovations of 2015 and 2016, further emphasizing the lasting influence of early RL breakthroughs on the discipline.

Figura 4.4: Reinforcement Learning Group



4.1.5 Expert Systems

As an early branch of artificial intelligence, Expert Systems emerged in the 1970s and 1980s, focusing on developing rule-based systems emulating human expert decision-making capabilities [54, 55]. These systems, comprising a knowledge base, an inference engine, and a user interface, captured domain-specific knowledge in rules and facts, applying logical reasoning to draw inferences and provide recommendations. While expert systems played a relevant role in AI's historical evolution, their influence in the current machine learning landscape has diminished due to the advent of more sophisticated techniques like representation learning [56].

Representation learning is a method in which models automatically discover and learn relevant features or representations from raw data, without relying on manually engineered features. This contrasts with the feature engineering approach used in Expert Systems, where domain experts would design and handcraft features to capture the most relevant aspects of the problem. The shift towards representation learning has allowed for more flexible, scalable, and adaptive models capable of handling complex, high-dimensional data.

In the 1990s, expert systems represented one of the most frequent topics in machine learning literature, peaking at 17.42% of all publications. During this period, expert systems gained widespread recognition and were employed in various applications, such as medical diagnosis [57], business decision-making [58], and fault detection [59]. However, today, they account for only around 1.53%, as shown in Figure 4.5, highlighting the decline in this topic's importance. The shift in focus towards representation learning and the rise of deep learning have contributed to the reduced emphasis on expert systems in contemporary AI research.

Figura 4.5: Expert Systems & Design Topic

4.1.6 Historical Contexts

Examining the evolution of scientific ideas, methodologies, and paradigms provides researchers with insights into factors influencing past discoveries, limitations of prevailing techniques, and driving forces behind major shifts in scientific thinking. Understanding historical context allows scientists to appreciate current theories and practices, identify foundations for new knowledge, and recognize the social, economic, and political forces shaping scientific inquiry. This contextual awareness deepens the understanding of the scientific process, informs future research direction, and fosters a holistic and nuanced perspective on scientific progress.

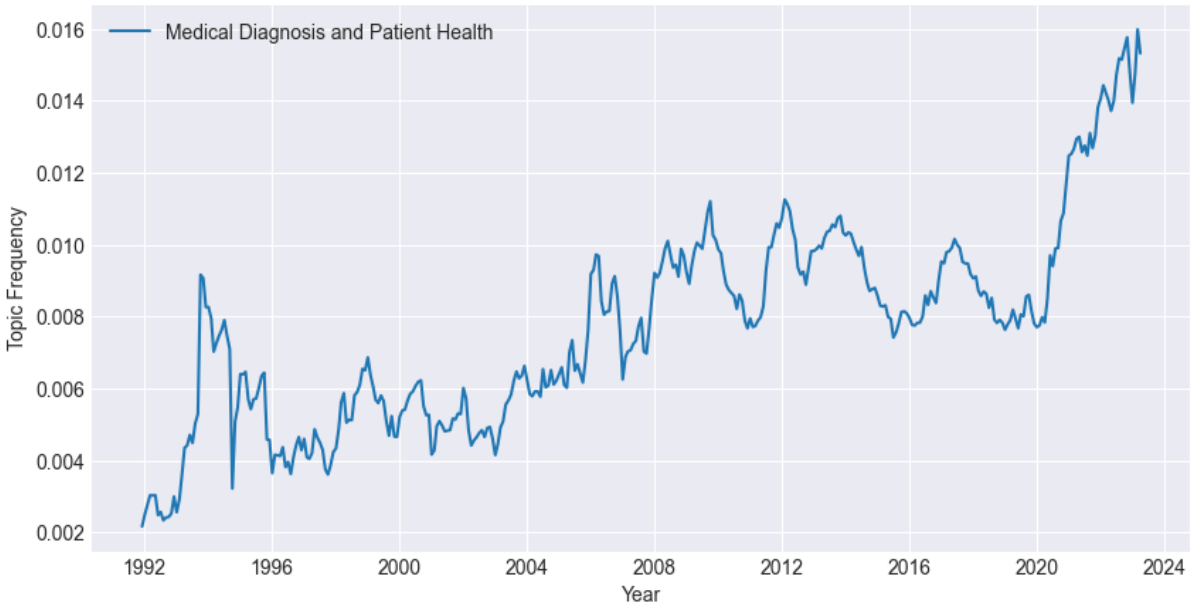
As Figure 4.6 demonstrates, the 2008 financial crisis precipitated a marked increase in machine learning publications addressing financial markets and risk. This event exposed

the inadequacies of conventional risk management and forecasting methods, prompting researchers to seek innovative solutions. Machine learning proved to be a powerful resource, offering precise, data-driven insights for market trends and decision-making processes, which exemplifies how historical events can significantly influence the trajectory of scientific research and technology within a particular domain.

Figura 4.6: Financial Markets & Risk Topic

Figure 4.7 highlights the substantial increase in machine learning publications focused on medical diagnosis and patient health following the 2020 COVID-19 pandemic. This event emphasized the necessity for advanced diagnostic tools and personalized healthcare solutions, leading researchers to explore machine learning applications in disease detection, treatment, and patient care [60, 61, 62, 63]. This example further illustrates the profound influence of historical context on scientific progress and technology development within specific fields, such as healthcare and medical diagnosis.

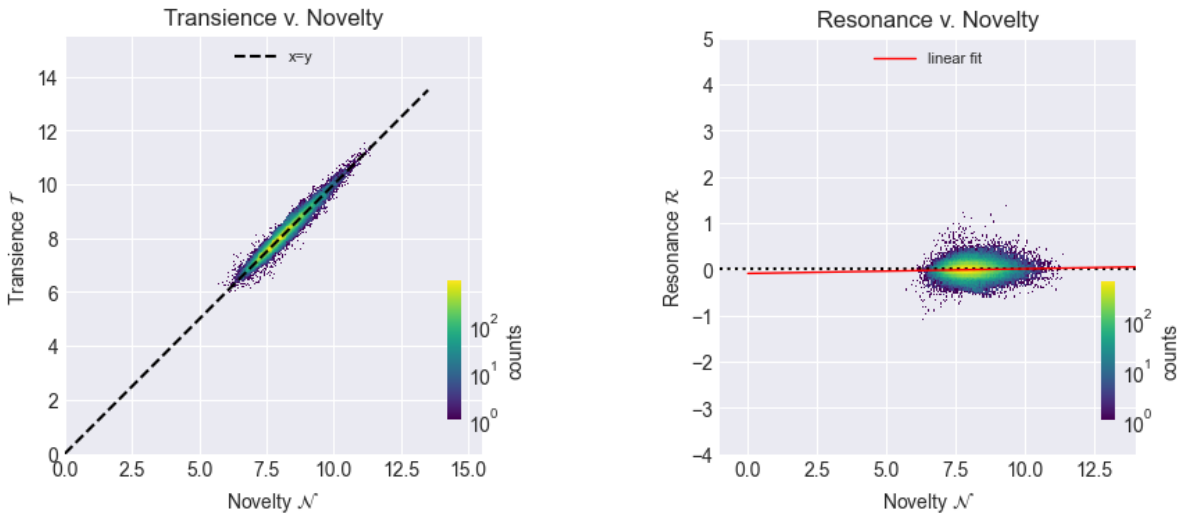
Figura 4.7: Medical Diagnosis and Patient Health



4.2 Novelty, Transience, Resonance

Figure 4.8a shows that the relation between Transience and Novelty is close to the identity line ($x = y$). This suggests that an increase in novelty is generally matched by an equal increase in transience. In simpler terms, the more novel a research work is, the less likely it is for that content to propagate into subsequent works. However, this symmetry is broken by resonant works, which differ more from their past and align more with their future. These works are found below the identity line, where novelty outweighs transience.

Figura 4.8: Innovation Bias for $w = 12$



(a) (b)

In Figure 4.8b, we see that the red line, representing the novelty effectiveness defined in Equation 2.10, is close to zero. This indicates that there is no systematic relation between novelty and resonance in the entire dataset. Despite the general trend of increased novelty leading to increased transience, the lack of a systematic relationship between novelty and resonance suggests that the influence of a paper is not solely determined by its novelty.

4.2.1 Authors

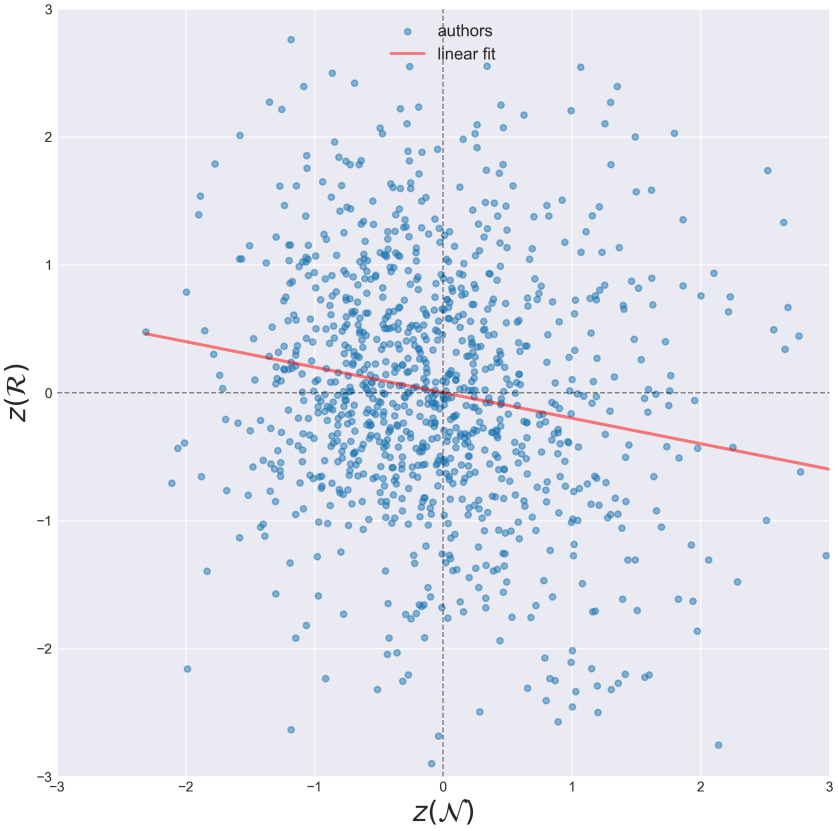
In order to evaluate the author capabilities, we need to attribute publications to their respective authors. However, the extensive diversity of venues in the dataset presented challenges in accurately matching authors, particularly when dealing with identical names or those publishing under multiple name variations (e.g., YOSHUA, BENGIO and BENGIO, YOSHUA). With 225,825 unique names in the dataset, manually verifying each case was not feasible. The difficulty of dealing with unmatching names is a well-known issue in scientific research involving large datasets [64, 65]. To address this challenge, several measures were taken to improve the accuracy of author name matching.

Initially, the focus was narrowed to the top 1000 authors with the largest number of publications. Due to some authors having the same number of papers, this reduced dataset comprised 1039 unique author names. Subsequently, the [13] API and [66]'s Names Matching Fuzzy Algorithm[1] were employed to automatically identify duplicate

[1]This algorithm is structured around the subsequent steps: (1) Parsing, normalizing, and segmenting the names within each identity, resulting in a set of strings for each one. (2) Establishing the similarity between identities. (3) Creating the distance matrix between identities within two designated lists. (4) Addressing the Linear Assignment Problem (LAP) associated with this matrix.

names, further reducing the unique names to 1029.

Figura 4.9: Innovation Bias for the 1029 most frequent authors



It is important to emphasize that the comparison being conducted pertains to the most prolific researchers in the field, whose contributions undoubtedly hold considerable significance. Nonetheless, within this group of scholars, Figure 4.9 reveals a notable observation: among the 1029 most frequent writers, there appears to be a significant novelty avoidance. This result suggests that established authors may have a preference for working within their areas of expertise and familiarity, leading to a more conservative approach in their research and a lower degree of novelty compared to less-established researchers who are more likely to explore uncharted territory or take risks with novel ideas.

Additionally, the top 1000 authors, who may have a higher degree of influence in their respective fields, could be more focused on refining and consolidating existing knowledge rather than pursuing radical innovations, potentially stemming from the pressure to maintain their status and reputation within the scientific community. However, it is important to note that some authors defied this tendency and managed to achieve high resonance in their work, even as they pursued high novelty.

Tabela 4.1: Highest and lowest scoring authors for Novelty and Resonance

	High resonance				Low resonance			
	Name	$z(\mathcal{N})$	$z(\mathcal{R})$	$\Delta z(\mathcal{N})$	Name	$z(\mathcal{N})$	$z(\mathcal{R})$	$\Delta z(\mathcal{N})$
High novelty	Qiu, Xipeng	1.352	2.394	2.664	Tong, Shaocheng	2.138	-2.755	-2.328
	Huang, Xuanjing	1.299	2.268	2.527	Chen, Huayou	6.271	-2.572	-1.321
	Sun, Xu	1.256	2.102	2.353	Hua, Changchun	1.201	-2.500	-2.260
	Zhao, Dongyan	1.796	2.026	2.385	Liao, Huchang	5.121	-2.496	-1.474
	Wang, William Yang	1.490	1.999	2.297	Mesiar, Radko	4.408	-2.439	-1.560
	Name	$z(\mathcal{N})$	$z(\mathcal{R})$	$\Delta z(\mathcal{N})$	Name	$z(\mathcal{N})$	$z(\mathcal{R})$	$\Delta z(\mathcal{N})$
Low novelty	Pang, Yanwei	-1.181	2.760	2.524	Hsu, Chun-fei	-1.182	-2.634	-2.870
	Lu, Huchuan	-1.085	2.394	2.178	Raja, Muhammad A. Z.	-1.986	-2.158	-2.555
	Ouyang, Wanli	-1.350	2.272	2.002	Veeraraghavan, Ashok	-1.146	-1.917	-2.145
	Shen, Jianbing	-1.255	2.214	1.964	Li, Kenli	-1.301	-1.571	-1.831
	Huang, Feiyue	-1.580	2.013	1.697	Hu, Bin	-1.834	-1.394	-1.760

Table 4.1 displays the authors with the highest and lowest scores in terms of novelty and resonance. We identified these authors by applying a z-score transformation to the data, selecting the top 100 most novel authors and the top 100 least novel authors. We then chose the 5 authors with the highest resonance and the 5 authors with the lowest resonance from each group.

The metric $\Delta z(\mathcal{N})$, defined as $z(\mathcal{R}) - E[z(\mathcal{R})|z(\mathcal{N})]$, quantifies the deviation of an author's resonance score from the expected resonance score given their novelty score. Essentially, it shows the extent to which an author's resonance diverges from the overall trend observed between novelty and resonance in the data.

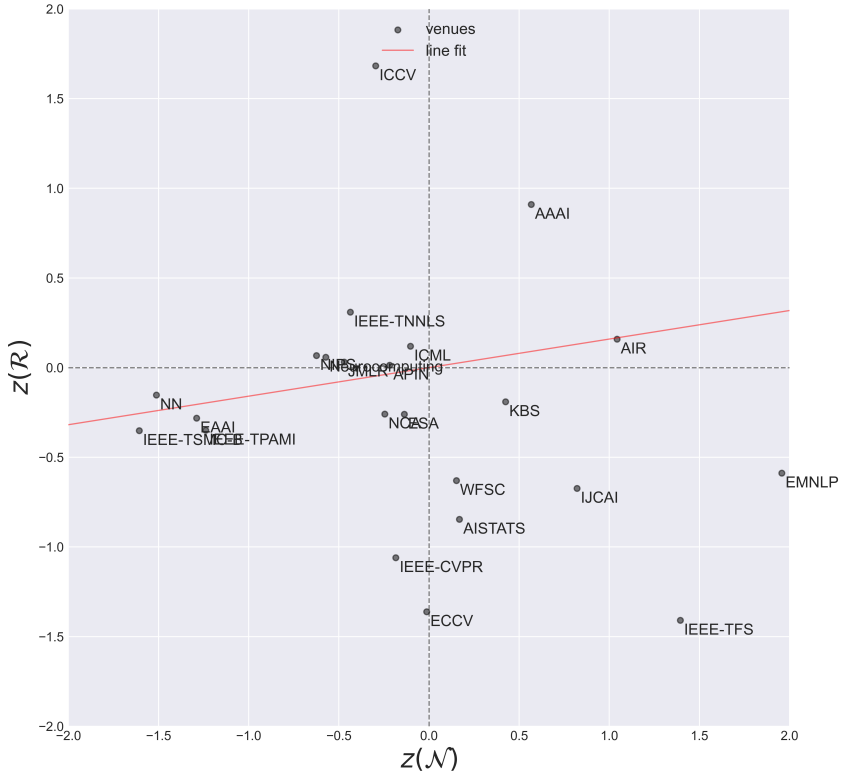
Observing a high $\Delta z(\mathcal{N})$ for authors with high novelty implies that these authors are able to achieve a greater than expected impact on their field, even as they pursue novel ideas. This deviation from the general trend of innovation avoidance might be attributed to the individual abilities and skills of these authors, which enable them to explore new concepts while still making a significant impact in their respective domains.

4.2.2 Venues

To be able to compare the academic venues in terms of resonance and novelty, we evaluate these metrics for individual papers using the information of the topic of the paper. Subsequently, we categorize the papers based on their respective venues. Finally, we calculate the average resonance and novelty values for each specific venue. These values are presented in Figure 4.10, along with the linear regression of these values. This regression can provide valuable insights into whether a given venue falls above or below the common average. It is important to acknowledge the presence of inherent sampling bias in our analysis. The selected venues were included based on their perceived significance, influence, and anticipated novelty in the field of Machine Learning. Therefore, in this figure, we are inherently comparing highly relevant venues. Deviations from the straight line

should not be interpreted as diminishing a venue's worthiness, given the distinguished nature of the venues being compared.

Figura 4.10: Innovation Bias for venues



Furthermore, it is worth mentioning the unique nature of the Computer Science field, which, unlike many other fields, accords conferences a distinctive significance. Notably, numerous significant findings are exclusively disseminated through conferences in this field due to their rapid information dissemination.

Table 4.2 presents the venues considered in this work ranked by the deviations to the expected resonance score given their novelty score. It is amazing the relevant role of conferences in specific fields, such as ICLR (International Conference on Learning Representations), ACL (Meeting of the Association for Computational Linguistics) ICCV (International Conference on Computer Vision).

Tabela 4.2: Venues normalized novelties and resonances ranked by $\Delta z(\mathcal{R})$

Venue	Abbreviation	$z(\mathcal{N})$	$z(\mathcal{R})$	$\Delta z(\mathcal{R})$
International Conference on Learning Representations	ICLR	-0.12159	2.587853	2.607214
Meeting of the Association for Computational Linguistics (ACL)	ACL	2.539549	2.471487	2.067111
International Conference on Computer Vision (ICCV)	ICCV	-0.29588	1.682636	1.729749
AAAI Conference on Artificial Intelligence	AAAI	0.567189	0.910322	0.820008
IEEE Transactions on Neural Networks and Learning Systems	IEEE-TNNLS	-0.43608	0.309849	0.379287
Neural Information Processing Systems	NIPS	-0.62429	0.06708	0.166486
Neurocomputing	Neurocomputing	-0.57209	0.058609	0.149704
International Conference on Machine Learning	ICML	-0.10303	0.119577	0.135983
Journal of Machine Learning Research	JMLR	-0.47041	0.031591	0.106495
Neural Networks	NN	-1.51144	-0.15239	0.088282
Applied Intelligence	APIN	-0.21868	0.013124	0.047944
Artificial Intelligence Review	AIR	1.042796	0.158838	-0.00721
Engineering Applications of Artificial Intelligence	EAAI	-1.28932	-0.28147	-0.07617
IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)	IEEE-TSMC-B	-1.6068	-0.35138	-0.09553
IEEE Transactions on Pattern Analysis and Machine Intelligence	IEEE-TPAMI	-1.2382	-0.34967	-0.15251
Neural Computing and Applications	NCA	-0.24493	-0.25899	-0.21999
Expert Systems with Applications	ESA	-0.13667	-0.25988	-0.23812
Knowledge-Based Systems	KBS	0.42484	-0.19057	-0.25822
Applied Soft Computing	WFSC	0.15193	-0.62989	-0.65409
International Joint Conferences on Artificial Intelligence Organization	IJCAI	0.821203	-0.67314	-0.8039
International Conference on Artificial Intelligence and Statistics	AISTATS	0.168854	-0.8456	-0.87248
Conference on Empirical Methods in Natural Language Processing	EMNLP	1.95683	-0.58854	-0.90013
Computer Vision and Pattern Recognition Conference	IEEE-CVPR	-0.18398	-1.05983	-1.03053
European Conference on Computer Vision	ECCV	-0.01344	-1.36115	-1.35901
IEEE Transactions on Fuzzy Systems	IEEE-TFS	1.393643	-1.40848	-1.63039

Capítulo 5

Conclusion

In this study, we have employed LDA to delve into the evolution of machine learning (ML) research. Through LDA, we discern key themes and foundational concepts within the field. By segmenting these themes, we trace their temporal trends. Ultimately, leveraging the Kullback-Leibler Divergence metric, we ascertain the roles of prominent authors and machine learning venues in shaping the landscape.

Our findings unveil the swift evolution of the machine learning field towards emerging technologies, occurring concurrently with the diminishing relevance of other technologies that are gradually receding from the forefront. Remarkably, deep learning emerges as the focal point of utmost interest within the field, while expert systems, once of paramount importance, have irreversibly slipped into obscurity. Moreover, domains such as computer vision and natural language processing have substantially integrated into the realm of deep learning research.

We have also investigated the roles of the authors in generating novel insights and the academic venues for disseminating this knowledge. Notably, our exploration has revealed that certain prominent authors exhibit an inclination towards innovation, while others adopt a more conventional stance. Regarding academic venues, we have identified the distinct significance of conferences and broadly scoped periodicals in spreading this wealth of knowledge.

References

[1] Blei, David M., Andrew Y. Ng e Michael I. Jordan: *Latent dirichlet allocation.* J. Mach. Learn. Res., 3(null):993–1022, mar 2003, ISSN 1532-4435. 1, 3, 15

[2] Röder, Michael, Andreas Both e Alexander Hinneburg: *Exploring the space of topic coherence measures.* Em *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, página 399–408, New York, NY, USA, 2015. Association for Computing Machinery, ISBN 9781450333177. `https://doi.org/10.1145/2684822.2685324`. 1, 5

[3] Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang e Simon DeDeo: *Individuals, institutions, and innovation in the debates of the french revolution.* Proceedings of the National Academy of Sciences, 115(18):4607–4612, 2018. `https://www.pnas.org/content/115/18/4607`. 1, 2, 6

[4] Langley, Pat *et al.*: *The changing science of machine learning.* Machine learning, 82(3):275–279, 2011. 2

[5] Fradkov, Alexander L: *Early history of machine learning.* IFAC-PapersOnLine, 53(2):1385–1390, 2020. 2

[6] Hall, David, Daniel Jurafsky e Christopher D. Manning: *Studying the history of ideas using topic models.* Em *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, página 363–371, USA, 2008. Association for Computational Linguistics. 2, 5

[7] Savov, Pavel, Adam Jatowt e Radoslaw Nielek: *Identifying breakthrough scientific papers.* Information Processing & Management, 57(2):102168, 2020. 2

[8] Hoffman, Matthew, Francis Bach e David Blei: *Online learning for latent dirichlet allocation.* advances in neural information processing systems, 23, 2010. 4

[9] Řehůřek, Radim e Petr Sojka: *Software Framework for Topic Modelling with Large Corpora.* Em *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta, maio 2010. ELRA. `http://is.muni.cz/publication/884893/en`. 4

[10] Bouma, Gerlof: *Normalized (pointwise) mutual information in collocation extraction.* Proceedings of GSCL, 30:31–40, 2009. 4, 5

[11] Church, Kenneth e Patrick Hanks: *Word association norms, mutual information, and lexicography.* Computational linguistics, 16(1):22–29, 1990. 4

[12] Kullback, S. e R. A. Leibler: *On Information and Sufficiency*. The Annals of Mathematical Statistics, 22(1):79 – 86, 1951. https://doi.org/10.1214/aoms/1177729694. 5

[13] DBLP: *dblp computer science bibliography*. https://dblp.org/, acesso em 2023-04-06, Provides bibliographic information on major computer science conferences and journals. 8, 21

[14] Jordan, M. I. e T. M. Mitchell: *Machine learning: Trends, perspectives, and prospects*. Science, 349(6245):255–260, 2015. https://www.science.org/doi/abs/10.1126/science.aaa8415. 9

[15] Griffiths, Thomas L. e Mark Steyvers: *Finding scientific topics*. Proceedings of the National Academy of Sciences, 101(suppl_1):5228–5235, 2004. https://www.pnas.org/doi/abs/10.1073/pnas.0307752101. 11

[16] Hinton, Geoffrey E., Simon Osindero e Yee Whye Teh: *A fast learning algorithm for deep belief nets*. Neural Comput., 18(7):1527–1554, jul 2006, ISSN 0899-7667. https://doi.org/10.1162/neco.2006.18.7.1527. 11

[17] LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard e L. D. Jackel: *Backpropagation applied to handwritten zip code recognition*. Neural Computation, 1(4):541–551, 1989. 11

[18] Bengio, Yoshua: *Learning deep architectures for AI*. Foundations and Trends in Machine Learning, 2(1):1–127, 2009, ISSN 1935-8237. 11

[19] Krizhevsky, Alex, Ilya Sutskever e Geoffrey E Hinton: *Imagenet classification with deep convolutional neural networks*. 25, 2012. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. 12, 13

[20] Sutskever, Ilya, Oriol Vinyals e Quoc V. Le: *Sequence to sequence learning with neural networks*. 2014. 12

[21] Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath e Brian Kingsbury: *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*. IEEE Signal Processing Magazine, 29(6):82–97, 2012. 12

[22] Kuhn, Thomas Samuel: *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1427 East 60th Street Chicago, IL 60637 U.S.A, 1962. 12

[23] Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville e Yoshua Bengio: *Generative adversarial networks*. 2014. 12

[24] Bahdanau, Dzmitry, Kyunghyun Cho e Yoshua Bengio: *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473, 2014. 12, 15

[25] Canny, John: *A computational approach to edge detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(6):679–698, 1986. 13

[26] Duda, Richard O., Peter E. Hart e David G. Stork: *Pattern Classification*. Wiley, 111 River St, United States, 2000. 13

[27] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning*. Springer, New York, NY, 2009. 13

[28] Lindeberg, Tony: *Scale-space theory: A basic tool for analysing structures at different scales*. Journal of Applied Statistics, 21:224–270, setembro 1994. 13

[29] Kass, Michael, Andrew P. Witkin e Demetri Terzopoulos: *Snakes: Active contour models*. Int. J. Comput. Vis., 1(4):321–331, 1988. http://dblp.uni-trier.de/db/journals/ijcv/ijcv1.html#KassWT88. 13

[30] Hartley, Richard e Andrew Zisserman: *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2ª edição, 2003, ISBN 0521540518. 13

[31] Tikhonov, Andrey N. e Vasiliy Y. Arsenin: *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics. 13

[32] Geman, Stuart e Donald Geman: *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6(6):721–741, 1984. http://dx.doi.org/10.1080/02664769300000058. 13

[33] Turk, Matthew e Alex Pentland: *Eigenfaces for recognition*. J. Cognitive Neuroscience, 3(1):71–86, 1991. 13

[34] Lecun, Y., L. Bottou, Y. Bengio e P. Haffner: *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, 86(11):2278–2324, novembro 1998, ISSN 00189219. http://dx.doi.org/10.1109/5.726791. 13

[35] Long, Jonathan, Evan Shelhamer e Trevor Darrell: *Fully convolutional networks for semantic segmentation*, 2014. http://arxiv.org/abs/1411.4038, cite arxiv:1411.4038Comment: to appear in CVPR (2015). 13

[36] Radford, Alec, Luke Metz e Soumith Chintala: *Unsupervised representation learning with deep convolutional generative adversarial networks*, 2015. http://arxiv.org/abs/1511.06434, cite arxiv:1511.06434Comment: Under review as a conference paper at ICLR 2016. 13

[37] TURING, A. M.: *I.—COMPUTING MACHINERY AND INTELLIGENCE*. Mind, LIX(236):433–460, outubro 1950, ISSN 0026-4423. https://doi.org/10.1093/mind/LIX.236.433. 14

[38] Shannon, Claude E: *A mathematical theory of communication.* The Bell system technical journal, 27(3):379–423, 1948. 14

[39] Baeza-Yates, Ricardo e Berthier Ribeiro-Neto: *Modern Information Retrieval.* Addison-Wesley Publishing Company, USA, 2nd edição, 2008, ISBN 9780321416919. 14

[40] Manning, Christopher D., Prabhakar Raghavan e Hinrich Schütze: *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008, ISBN 0521865719, 9780521865715. 14

[41] Chomsky, Noam: *Aspects of the Theory of Syntax.* The MIT Press, One Broadway, E70. Floor 12. Cambridge, MA 02142, 50ª edição, 1965, ISBN 9780262527408. `http://www.jstor.org/stable/j.ctt17kk81z`, acesso em 2023-04-26. 14

[42] Gong, Yihong e Xin Liu: *Generic text summarization using relevance measure and latent semantic analysis.* Em *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, página 19–25, New York, NY, USA, 2001. Association for Computing Machinery, ISBN 1581133316. 15

[43] Mihalcea, Rada e Paul Tarau: *Textrank: Bringing order into text.* Em *Proceedings of the 2004 conference on empirical methods in natural language processing*, páginas 404–411, 2004. 15

[44] Erkan, Günes e Dragomir R. Radev: *Lexrank: Graph-based lexical centrality as salience in text summarization.* Journal Artificial Intelligence Research, 22(1):457–479, 2004, ISSN 1076-9757. 15

[45] Hochreiter, Sepp e Jürgen Schmidhuber: *Long short-term memory.* Neural computation, 9(8):1735–1780, 1997. 15

[46] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin: *Attention is all you need.* Advances in neural information processing systems, 30, 2017. 15

[47] Devlin, Jacob, Ming Wei Chang, Kenton Lee e Kristina Toutanova: *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018. 15

[48] Radford, Alec, Karthik Narasimhan, Tim Salimans e Ilya Sutskever: *Improving language understanding by generative pre-training.* OpenAI Blog, 2018. 15

[49] Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei e Ilya Sutskever: *Language models are unsupervised multitask learners.* OpenAI Blog, 2019. 15

[50] Watkins, Christopher John Cornish Hellaby: *Learning from delayed rewards.* 1989. 16

[51] Sutton, Richard S.: *Learning to predict by the methods of temporal differences.* Mach. Learn., 3(1):9–44, aug 1988, ISSN 0885-6125. https://doi.org/10.1023/A:1022633531479. 16

[52] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg e Demis Hassabis: *Human-level control through deep reinforcement learning.* Nature, 518:529–533, 2015. 16

[53] Silver, David, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel e Demis Hassabis: *Mastering the game of go with deep neural networks and tree search.* Nature, 529:484–489, janeiro 2016. 16

[54] Hayes-Roth, Frederick, Donald A. Waterman e Douglas B. Lenat: *Building Expert Systems.* Addison-Wesley Longman Publishing Co., Inc., USA, 1983, ISBN 0201106868. 17

[55] Lucas, Peter J. e Linda C. Gaag: *Principles of Expert Systems.* janeiro 1991. 17

[56] LeCun, Yann, Yoshua Bengio e Geoffrey Hinton: *Deep learning.* nature, 521(7553):436–444, 2015. 17

[57] Miller, Randolph, Melissa McNeil, Sue Challinor, Fred Masarie e Jack Myers: *The internist-1/quick medical reference project-status report.* The Western journal of medicine, 145:816–22, janeiro 1987. 18

[58] Wong, Bo K e John A Monaco: *Expert system applications in business: A review and analysis of the literature (1977-1993).* Information & Management, 29(3):141–152, 1995. 18

[59] Venkatasubramanian, Venkat: *A review of process fault detection and diagnosis: Part i: Quantitative model-based methods.* Computers & Chemical Engineering, 27:293–311, 2003. 18

[60] Vaishya, Raju, Mohd. Javaid, Ibrahim Haleem Khan e Abid Haleem: *Artificial intelligence (ai) applications for covid-19 pandemic.* Diabetes & Metabolic Syndrome, 14:337 – 339, 2020. 19

[61] Alimadadi, Ali, Sudip Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe e Xi Cheng: *Artificial intelligence and machine learning to fight covid-19.* Physiological Genomics, 52(4):200–202, 2020. 19

[62] Khan, Muzammil, Muhammad Taqi Mehran, Zeeshan Ul Haq, Zahid Ullah, Salman Raza Naqvi, Mehreen Ihsan e Haider Abbass: *Applications of artificial intelligence in covid-19 pandemic: A comprehensive review.* Expert systems with applications, 185:115695, 2021. 19

[63] Utku, Anil: *Deep learning based hybrid prediction model for predicting the spread of covid-19 in the world's most populous countries*. Expert Systems with Applications, página 120769, 2023. 19

[64] Tang, Li e John P. Walsh: *Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps*. Scientometrics, 84(3):763–784, September 2010. `https://ideas.repec.org/a/spr/scient/v84y2010i3d10.1007_s11192-010-0196-6.html`. 21

[65] Strotmann, Andreas e Dangzhi Zhao: *Author name disambiguation: What difference does it make in author-based citation analysis?* Journal of the American Society for Information Science and Technology, 63:1820–1833, setembro 2012. 21

[66] Athenianco: *names-matcher*. `https://github.com/athenianco/names-matcher`, 2021. Accessed: April 29, 2023. 21

Apêndice A

List of popular machine leaning sources

Table A.1 presents a list of the popular machine learning sources used to find the list of most popular machine learning venues.

Tabela A.1: URL's used in the first step of our search

https://aclanthology.org/
https://proceedings.mlr.press/
https://www.quora.com/What-are-the-best-conferences-and-journals-about-machine-learning
https://research.com/conference-rankings/computer-science/machine-learning
https://deepai.space/top-ai-conferences-and-journals/
https://www.junglelightspeed.com/the-top-10-nlp-conferences/

Apêndice B

Details of the LDA implementation

In this section, we describe the process we used to determine the values of the LDA parameters.

We start by applying Document Frequency (TF) and Token Frequency-Inverse Document Frequency (TF-IDF) in order to eliminate words exhibiting low importance within the corpus. Then, we search for the optimal value of the number of topics K.

As we mention before, we have implemented the LDA using the Python library Gensim. It provides three alternatives for setting priors: (1) 'symmetric' (default), utilizing a fixed symmetric prior of $1/num\_topics$, (2) 'asymmetric', implementing a fixed normalized asymmetric prior of $1/(topic\_index + \sqrt{num\_topics})$, and (3) 'auto', which learns an asymmetric prior from the corpus. We incorporate these these configurations alongside document frequency and tf-idf filters to optimize the LDA model's performance during the hyperparameter search process.

Table B.1 displays the results, with document frequency values set at 0.5, meaning words appearing in more than 50% of the documents were removed from the corpus. The second value, 1.0, indicates no words were removed for models trained with this parameter. The tf-idf parameter compared a low value (0.0075), which removed fewer words, against a high value (0.015), which removed more words. The search space for α was ('symmetric', 'asymmetric') and for η it was ('symmetric', 'auto'). The best-performing model had $K = 60$, df $= 0.5$, tf-idf $= 0.0075$, $\alpha =$ asymmetric, and $\eta =$ auto. The reader can refer to Tables C.1 and C.2 to examine the resulting topics derived from this optimal combination of parameters, please access this paper's Zenodo for a more detailed overview on the topics and their word distributions.

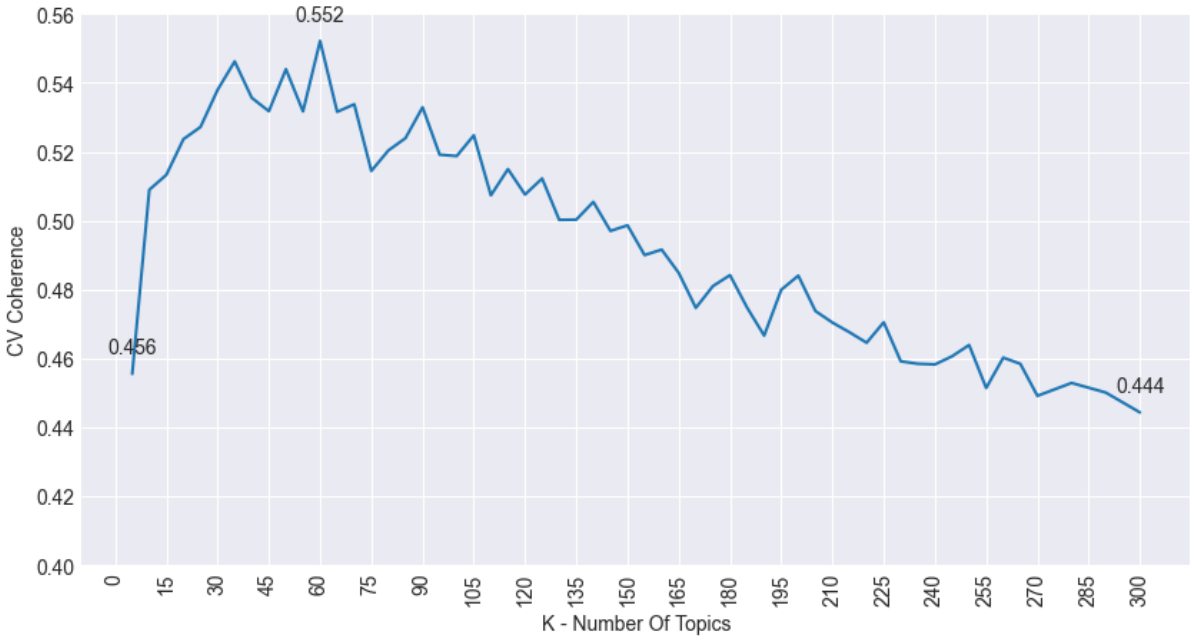
Figure B.1 displays graphically the results of the search for K values ranging from 5 to 300, with coherence peaking at 0.552 for $K = 60$. Lower K values fail to adequately capture the complex structure of the machine learning field's literature, likely combining disparate topics. As the K value increases, the optimal representation of the

Tabela B.1: Parametric space search results

DF	TF-IDF	K	α	η	C_V coherence
0.5	0.0075	60	symmetric	symmetric	0.531903
0.5	0.0075	60	symmetric	auto	0.531045
0.5	0.0075	60	asymmetric	symmetric	0.545108
0.5	**0.0075**	**60**	**asymmetric**	**auto**	**0.552183**
0.5	0.015	60	symmetric	symmetric	0.524149
0.5	0.015	60	symmetric	auto	0.543684
0.5	0.015	60	asymmetric	symmetric	0.528428
0.5	0.015	60	asymmetric	auto	0.523044
1	0.0075	60	symmetric	symmetric	0.526615
1	0.0075	60	symmetric	auto	0.536799
1	0.0075	60	asymmetric	symmetric	0.527374
1	0.0075	60	asymmetric	auto	0.520164
1	0.015	60	symmetric	symmetric	0.532143
1	0.015	60	symmetric	auto	0.534907
1	0.015	60	asymmetric	symmetric	0.518847
1	0.015	60	asymmetric	auto	0.504365

underlying semantic structure diminishes, resulting in topics primarily driven by statistical co-occurrence.

Figura B.1: Model Coherence vs. K - Number of Topics



Apêndice C

Summary of topics

Tables C.1 and C.2 display the topic labels generated by a collaboration between researchers and ChatGPT. The researchers provided the most frequent words for each topic, and ChatGPT, leveraging its language understanding capabilities, responded with an appropriate label for the topic. After that, the authors reviewed the labels generated by ChatGPT. These tables offer a succinct and interconnected representation of the dataset, showcasing the associated groups and the top words within each topic, thus providing a more comprehensive overview of the underlying themes in the data. Groups with an * were not decisively classified.

Tabela C.1: Summary of topics, groups and the main terms in each topic.

Topics	Labels	Groups	Terms
0	Graph Theory and Network Structures	Networks	graph, structure, node, tree, edge
1	Control Systems and Dynamic Systems	Systems & Control	control, system, time, controller, design
2	Systems and Energy Management	Systems & Control	system, power, time, energy, sensor
3	Multi-level Feature Fusion and Deep Learning	Deep Learning	multi, level, feature, module, dataset
4	Probability Distribution and Estimation Algorithms	Statistics	distribution, algorithm, probability, show, estimate
5	Evolutionary Algorithm and Optimization	Genetic Algorithms	algorithm, optimization, search, performance, local
6	Urban Navigation and Localization	Navigation	location, trajectory, vehicle, mobile, road
7	Object Recognition and Visual Saliency	Computer Vision	object, visual, image, scene, category
8	Object Detection	Computer Vision	detection, detect, detector, community, proposal
9	Cybersecurity & Network Communication	Cybersecurity	event, attack, service, communication, distribute
10	Information Retrieval & Data Mining	Data*	pattern, query, search, retrieval, question
11	Image Processing and Resolution	Computer Vision	surface, resolution, image, color, light
12	Relational Data and Entity Relationships	Data*	group, relation, entity, link, knowledge
13	Failure Diagnosis and Causal Effects	Fault Diagnosis	fault, effect, variable, failure, causal
14	Medical Diagnosis and Patient Health	Medical	diagnosis, patient, medical, disease, clinical
15	Neural Network and Synaptic Functioning	Deep Learning	neuron, spike, mechanism, unit, channel
16	Time Series Forecasting and Prediction Modeling	Statistics	prediction, time, series, model, predict
17	Camera Tracking and Motion Estimation	Computer Vision	camera, tracking, point, depth, image
18	Bias and Evidence Analysis	Statistics	bias, evidence, hypothesis, explanation, argument
19	Game Theory and Strategic Behavior	Decision Making	strategy, game, play, player, equilibrium
20	Facial Recognition & Identification	Computer Vision	face, recognition, image, identification, person
21	Computational Efficiency & Performance	Complexity	time, large, computational, scale, reduce
22	Shape and Gesture Recognition	Computer Vision	shape, recognition, hand, body, line
23	Semantic Word Embeddings in NLP	Natural Language Processing	word, semantic, sentence, embedding, representation
24	Reinforcement Learning and Robotics	Reinforcement Learning	agent, policy, environment, learning, robot
25	Human Attributes & Interactions	Reinforcement Learning*	human, attribute, cell, interaction, behavior
26	Neural Networks	Deep Learning	network, neural, layer, architecture, deep
27	Expert Systems & Design	Expert Systems	system, knowledge, design, process, expert
28	Regression Analysis & Modeling	Statistics	error, regression, model, test, parameter
29	Information Imputation and Completion	Machine Learning	information, miss, mutual, incomplete, side

Tabela C.2: Summary of topics, fields and the main terms in each topic.

Topics	Labels	Groups	Terms
30	Constraint Planning & Logic	Logic	constraint, path, show, set, property
31	Feature Extraction & Selection	Machine Learning	feature, extract, selection, extraction, select
32	Adversarial Generative Modeling	Deep Learning	generate, model, generation, adversarial, training
33	Sentiment and Emotion Analysis	Natural Language Processing	model, dataset, sentiment, emotion, task
34	Probabilistic Modeling & Inference	Statistics	model, parameter, process, modeling, inference
35	Social Recommendation Systems	Recommendation Systems	user, recommendation, item, social, system
36	Classification & Performance	Machine Learning	classification, classifier, machine, accuracy, performance
37	Multi-Dimensional Data Representation	Statistics	space, representation, view, dimensional, sparse
38	Robust Denoising	Statistics	noise, loss, robust, noisy, robustness
39	Data Clustering & Segmentation	Machine Learning	cluster, clustering, algorithm, datum, partition
40	Image Partitioning & Texture Analysis	Computer Vision	image, segmentation, region, pixel, texture
41	Decision Making & Evaluation	Decision Making	decision, criterion, make, preference, selection
42	Speech & Language Processing	Natural Language Processing	language, code, translation, model, speech
43	Domain Adaptation & Transfer Learning	Machine Learning*	domain, target, source, transfer, adaptation
44	Semi-Supervised Data Annotation	Machine Learning	label, datum, training, semi_supervised, annotation
45	Modeling with Attention Mechanisms	Deep Learning	sequence, attention, model, long, memory
46	Deep Learning	Deep Learning	learning, learn, task, training, deep
47	Signal Processing	Synaptic Functioning*	signal, phase, brain, frequency, activity
48	Video Action Recognition	Computer Vision	video, temporal, action, frame, motion
49	Document Summarization	Natural Language Processing	document, topic, content, web, summary
50	Similarity Measures and Rankings	Statistics*	measure, similarity, metric, correlation, rank
51	Research and Development Trends	Research	research, study, analysis, application, technique
52	Fuzzy Logic Systems	Fuzzy Algorithms	fuzzy, rule, system, function, logic
53	Fuzzy Value & Set Theory	Fuzzy Algorithms	value, set, weight, interval, uncertainty
54	Linear and Nonlinear Optimization	Optimization	function, matrix, linear, problem, non
55	Financial Markets & Risks	Finance	risk, market, student, financial, price
56	Adaptive optimization algorithms	Optimization	algorithm, parameter, rate, filter, convergence
57	Class Imbalance & Sampling	Machine Learning	class, sample, instance, problem, learning
58	Data Management and Utilization	Data	datum, data, set, real, technique
59	Optimization & Problem-Solving	Optimization	problem, algorithm, solution, solve, optimization