Uncovering Research Potentials: Research Areas Evolution Analysis in Scientific Articles

1st Lucas Cerqueira Figueiredo
Programa de Pós-Graduação em Engenharia Elétrica e Computação
Universidade Presbiteriana Mackenzie
São Paulo, Brasil
lucas.cerfig@gmail.com.br

2nd Leandro A. Silva
Programa de Pós-Graduação em Engenharia Elétrica e Computação
Universidade Presbiteriana Mackenzie
São Paulo, Brasil
leandroaugusto.silva@mackenzie.br

***Abstract*—This paper presents a document visualization tool that captures the main topics and similarity between scientific articles. The approach is based on recent techniques of Natural Language Processing and Deep Learning, using document embeddings. The tool combines vector representations of words with visualization techniques to condense a collection of articles useful for researchers performing a literature review and make a planning of research. Document Metadata analysis provides a temporal mapping of the evolution of the collection's areas. The approach employs SCIBERT to extract representations from abstracts, metrics collected from paper metadata, and dimensionality reduction techniques to provide a useful visualization for the researcher exploring the collection.**

***Index Terms*—information representation, natural language processing, deep learning, scientometrics**

I. Introdução

À medida que a literatura científica cresce a um ritmo acelerado, a necessidade de ferramentas eficientes para navegar neste vasto espaço de conhecimento se torna cada vez mais crítica. Embora técnicas de Processamento de Linguagem Natural (PLN) e *Deep Learning* tenham sido exploradas para facilitar a identificação de trabalhos relevantes e tendências emergentes [1, 2], lacunas substanciais na literatura ainda existem.

Diferencia-se o presente trabalho ao adotar modelos de linguagem avançados, particularmente o *SciBERT* [3], para aprimorar a representação textual em uma ferramenta de visualização de artigos científicos. Em comparação com sistemas de recomendação que utilizam representações vetoriais mais básicas como o Saco-de-Palavras [4], a representação textual aqui proposta é projetada para capturar de maneira mais rica e detalhada a semântica e o contexto dos documentos. Além disso, enquanto ferramentas de visualização anteriores como SurVis [5] e TextVis [6] focam na estrutura dos documentos, uma ênfase na semântica profunda é dada neste estudo.

O objetivo definido é a combinação da profundidade semântica proporcionada pelo modelo de linguagem *SciBERT* com métodos avançados de redução de dimensionalidade e visualização, como o UMAP [7]. Esta abordagem não só visa tornar a busca em literatura científica mais eficiente, mas também explorar novas formas de entender a evolução e as inter-relações nos campos científicos.

Este artigo está estruturado em várias seções além desta introdução. Na seção de Trabalhos Relacionados, uma revisão da literatura é fornecida, focalizando esforços anteriores na representação de documentos, análise de artigos científicos e visualização da informação. A seção de Referencial Teórico detalha as técnicas, algoritmos e abordagens previamente estabelecidas na literatura que foram adotadas na experimentação deste trabalho. Em Metodologia Experimental, descreve-se todo o processo de ingestão, manipulação e processamento dos dados de forma clara e detalhada, com um foco particular na reprodutibilidade dos experimentos. Finalmente, a seção de Resultados Experimentais apresenta um protótipo da ferramenta proposta.

Em resumo, a contribuição primária deste trabalho está na inovadora combinação de técnicas de PLN e *Deep Learning* para criar uma ferramenta de visualização que não só torna a literatura científica mais acessível, mas também revela *insights* profundos sobre a estrutura e a evolução dos campos de pesquisa.

II. Trabalhos Relacionados

No campo da análise de literatura científica, diferentes abordagens têm sido propostas para aprimorar a eficácia e a eficiência do processo. Um exemplo notável é o sistema de recomendação de citações introduzido por Bhagavatula et al. [2]. Esta solução se baseia em representações vetoriais de documentos obtidas através do modelo Saco-de-Palavras (Bag-of-Words) [4]. Neste sistema, documentos próximos em um espaço vetorial são selecionados como candidatos a citações e ordenados de acordo com sua relevância para o documento central.

Além disso, a visualização textual tem sido aplicada como uma maneira de facilitar a interpretação de literatura científica. Nesse sentido, ferramentas como SurVis [5] e

TextVis [6] foram desenvolvidas. SurVis permite a interação com representações visuais de conjuntos de documentos e agrupamento de artigos usando algoritmos como o k-means [8]. Por outro lado, TextVis usa uma abordagem baseada em grafos para representar e mostrar publicações de tópicos relacionados.

Adnani et al. [9] realizaram um estudo abrangente comparando os cinco índices de similaridade mais usados para três tipos de análise cientométrica: co-palavra, co-citação e co-autoria. Este trabalho representou um marco importante na compreensão das várias técnicas de análise cientométrica.

Embora todas essas abordagens tenham avançado significativamente no campo da análise da literatura científica, elas têm suas limitações. As representações vetoriais do modelo Saco-de-Palavras, embora eficazes, não capturam plenamente a semântica e sintaxe da linguagem. Ferramentas de visualização como SurVis e TextVis focam em aspectos estruturais dos documentos, mas não exploram a profundidade semântica dos textos. Da mesma forma, a análise cientométrica se concentra na comparação de índices de similaridade sem considerar a semântica inerente aos documentos.

Em contrapartida, este trabalho propõe a utilização de embeddings de documentos derivados do modelo SciBERT para capturar e visualizar a semântica dos documentos. Este enfoque na semântica visa superar as limitações das técnicas anteriormente mencionadas, fornecendo uma visão mais profunda da linguagem natural presente na literatura científica. Esta abordagem pode revelar conexões temáticas e ideias emergentes de maneira mais eficaz e intuitiva, avançando nosso entendimento da estrutura e evolução dos campos científicos.

III. Referencial Teórico

1) Representação de Textos: A representação de textos é um campo multidisciplinar que atrai atenção de diversas áreas do conhecimento, como ciência da computação [10], linguística [11] e psicologia [12], para citar algumas. No âmbito da ciência da computação, a análise de textos se estabelece como uma área de pesquisa dinâmica, com o desenvolvimento contínuo de técnicas inovadoras que visam extrair e decifrar informações significativas a partir de grandes conjuntos de texto [13]. Uma técnica particularmente impactante é a extração de *embeddings* de texto [14], a qual mapeia palavras, frases ou documentos inteiros em representações vetoriais de alta dimensão. Essas representações facilitam a aplicação de algoritmos de aprendizado de máquina, fornecendo uma abordagem computacional robusta para o processamento de dados textuais.

O BERT (Bi-directional Encoder Representations from Transformers) [15] é um modelo de linguagem pré-treinado que utiliza grandes conjuntos de dados para aprender representações contextuais de palavras. Os autores apresentam e empregam uma abordagem de máscara de linguagem (Masked Language Model - MLM), onde palavras são ocultadas em uma sentença, e o modelo é treinado para prever essas palavras com base no contexto. Além disso, no trabalho os autores apresentam e utilizam no modelo a técnica de Predição de Próxima Sentença (Next Sentence Prediction - NSP) para aprender sobre as relações entre frases. A arquitetura do BERT é baseada em Transformers, que são modelos de aprendizado profundo com mecanismos de atenção que permitem capturar informações contextuais de forma eficiente ([16].

Já o SCIBERT é uma variante específica do BERT projetada para lidar com textos científicos ([3]). Ele é treinado em um corpus de artigos científicos e preserva a arquitetura do BERT, mas possui adaptações para o contexto científico, como um vocabulário especializado e consideração da estrutura dos artigos. Essas modificações permitem que o SCIBERT seja mais eficaz na análise de textos científicos, incluindo classificação e agrupamento de documentos.

2) Redução de Dimensionalidade: A técnica UMAP (Uniform Manifold Approximation and Projection) [7] é um algoritmo de redução de dimensionalidade e visualização de dados de grandes dimensões. Foi desenvolvida para preservar a estrutura e a relação dos dados em espaços de alta dimensionalidade ao projetá-los em espaços de menor dimensão. O UMAP utiliza conceitos da teoria dos grafos e otimização para mapear os dados de forma eficiente e preservar a proximidade e a conectividade entre as instâncias.

3) Bases de Dados Acadêmicas: O arXiv [17] é uma base de dados composta por artigos científicos nas áreas de matemática, física, astronomia, engenharia elétrica, ciência da computação, biologia, estatística, finanças e economia. Nessa plataforma, os pesquisadores têm a possibilidade de compartilhar seus trabalhos de forma ágil e aberta, sem a necessidade de passarem pelo processo tradicional de revisão por pares. Isso resulta em um acesso rápido a uma vasta quantidade de conhecimento científico atualizado. É importante destacar que muitos artigos disponibilizados no arXiv são posteriormente publicados em periódicos científicos renomados, o que evidencia a relevância dessa base de dados como uma fonte precursora de pesquisas e descobertas científicas.

IV. Metodologia Experimental

Os experimentos implementados foram desenvolvidos usando a linguagem Python 3 [18], com o apoio de bibliotecas especializadas em manipulação de dados (Pandas [19]), modelagem de linguagem (HuggingFace Tranformers [20]) e visualização interativa de dados (Plotly [21]).

Para obter o conjunto de dados do arXiv [17], importamos o dataset *arXiv Dataset* hospedado na plataforma Kaggle [22]. Este conjunto de dados inclui 2.2 milhões de artigos científicos. Cada artigo é representado pelos seguintes metadados:

- **id**: Identificador único do artigo no repositório do arXiv.
- **submitter**: Nome do submissor.
- **authors**: Autores do artigo.
- **title**: Título do artigo.
- **comments**: Comentários a respeito da submissão.
- **journal-ref**: Referências de jornais onde o artigo foi publicado (se publicado).

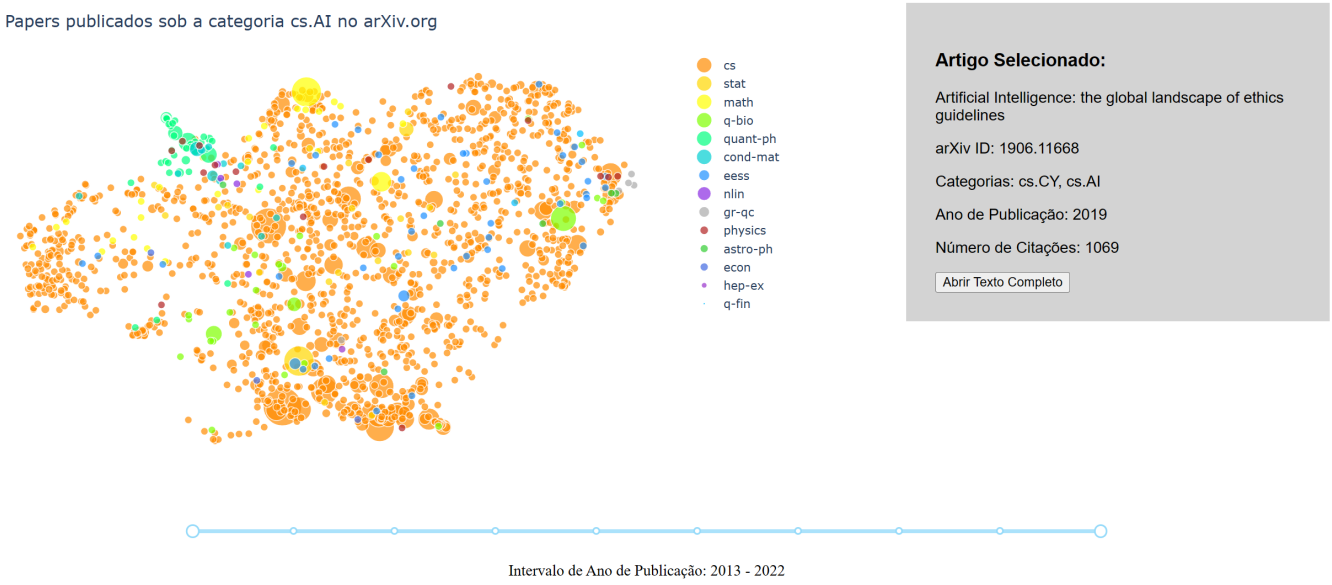

Fig. 1. Estrutura da ferramenta

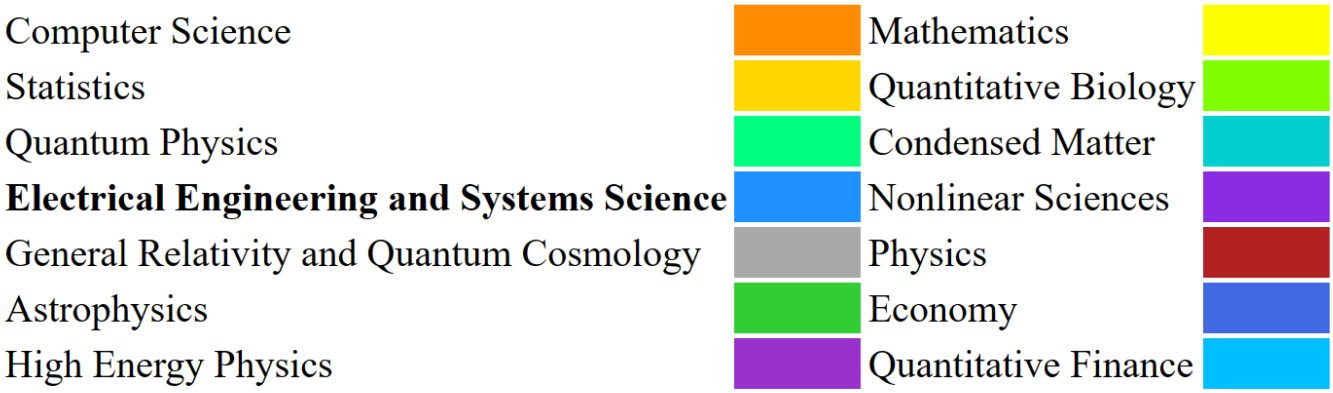

Fig. 2. Legenda das categorias presentes no conjunto visualizado na Figura 1

- **doi**: Digital Object Identifier – identificador para versões digitais do artigo.
- **report-no**: Identificador atribuído pela instituição responsável pela submissão.
- **categories**: Categorias atribuídas ao artigo, referentes à taxonomia do arXiv.
- **license**: Licença da submissão.
- **abstract**: Resumo do artigo.
- **versions**: Registro das versões da publicação.
- **update-date**: Data da última atualização do artigo.
- **authors_parsed**: Lista de autores.

Para tratar e manipular tais metadados, a biblioteca pandas [19] foi utilizada. Foram mantidas apenas as colunas *id*, *title*, *journal-ref*, *doi*, *categories*, *abstract* e *authors_parsed*.

A biblioteca *datasets* do *Hugging Face* [20] foi utilizada para otimizar o carregamento dos dados, dado que o volume de artigos é de 3.7GB, em formato JSON.

Os metadados adicionais de cada artigo, como o ano de publicação e o número de citações, foram obtidos para aprofundar a análise. O ano de publicação foi extraído do campo *journal-ref* por meio de expressões regulares. Este método foi aplicado aos artigos com um campo *journal-ref* válido. Dos 703065 artigos processados, não foi possível identificar o ano de publicação em 64659 deles e por isso foram descartados da análise.

Após a extração do ano de publicação, uma filtragem adicional foi implementada para incluir apenas artigos publicados entre 1915 e 2022. Um recorte baseado na categoria de Inteligência Artificial (cs.AI) de acordo com a taxonomia do arXiv foi feito, removendo do conjunto de dados qualquer artigo que não apresentasse o item cs.AI em sua lista de categorias. Isso resultou em um total de 1966 artigos.

Para determinar o número de citações de cada um dos 1966 artigos, utilizou-se a API da plataforma *Crossref* [23]. Com o

DOI de cada artigo, foram feitas chamadas à API do *Crossref* para obter as respectivas contagens de citações. Dos 1966 artigos, um total de 1910 teve seus metadados extraídos com sucesso a partir da API e compôs o conjunto de dados final para uso nos experimentos. Os atributos e a estrutura final do conjunto estão detalhados na Tabela I, utilizando como exemplo o trabalho de Jobin et al. [24].

TABELA I
AMOSTRA DE UM ITEM DO DATASET AO FINAL DO PRÉ-PROCESSAMENTO

Atributos Categóricos	**Valor**
id	'1906.116680'
title	'Artificial Intelligence: the global landscape of ethics guidelines'
journal-ref	'Nat. Mach. Intell. (2019)'
doi	'10.1038/s42256-019-0088-2'
categories	['cs.CY', 'cs.AI']
abstract	'In the last five years, private companies [...]'
authors_parsed	[['Jobin', 'Anna'], ['Ienca', 'Marcello'], ['Vayena', 'Effy']]
Atributos Numéricos	**Valor**
year	2019
abstract_word_count	138
citations	1069

Na preparação dos experimentos, as bibliotecas *PyTorch* [25] e *Transformers* do Hugging Face foram utilizadas. O módulo *AutoModel* foi usado para importar o modelo pré-treinado *scibert-scivocab-uncased* do *SciBERT*. Optou-se por um tamanho máximo de entrada de 512 *tokens*, adequado ao resumo mais longo em nossa base de dados que possuía 315 palavras. A partir do modelo *SciBERT*, foram extraídas representações vetoriais de 768 dimensões para cada resumo no conjunto de 1910 artigos. As fases de filtragem e seleção podem ser observadas na Tabela II.

O modelo de Transformer utilizado é conhecido por sua eficiência computacional e capacidade de captar contextos complexos por meio de mecanismos de atenção multi-cabeça [16]. Diferentemente de representações vetoriais tradicionais, esses modelos fornecem *embeddings* dinâmicos, que variam com o contexto, capacidade crucial para compreender o significado ambíguo ou polissêmico de termos em textos científicos. O *SciBERT*, treinado em um corpus específico de literatura científica, foi empregado para produzir representações de 768 dimensões para cada resumo no nosso conjunto de 1.910 artigos, servindo como base para as análises subsequentes do presente estudo.

TABELA II
FASES DE SELEÇÃO E PRÉ-PROCESSAMENTO

Fase	**Operação**	**Artigos Resultantes**
1	Origem	2.238.880
2	Validação de DOI	1.111.732
3	Artigos publicados em periódicos	703.085
3	Filtragem temporal (1915-2022)	638.424
4	Seleção da categoria cs.AI	1966
5	Extração de metadados pelo Crossref	1910

A seguir, procedeu-se com a redução de dimensionalidade utilizando a técnica UMAP, com a implementação da biblioteca umap-learn [7]. A dimensão foi reduzida de 768 para apenas 2, a fim de permitir a visualização bidimensional dos dados. A decisão de reduzir a dimensão dos dados para duas permite uma visualização intuitiva e fácil de interpretar, o que é particularmente útil para observadores humanos. Isso facilita a identificação de padrões e agrupamentos em um espaço bidimensional. No entanto, essa redução substancial na dimensão também traz riscos, como a perda de nuances e detalhes que poderiam ser capturados em dimensões mais altas. Com isso, algumas relações complexas ou menos óbvias entre os artigos podem não ser tão claramente delineadas ou até mesmo perdidas. Se optássemos por uma representação de dimensão 10, por exemplo, a ferramenta poderia capturar uma gama mais rica de relações entre os artigos, mas ao custo de tornar a visualização e interpretação desses dados mais complexa. É provável que, em uma dimensão mais alta, diferentes agrupamentos ou até mesmo sub-agrupamentos poderiam emergir, oferecendo *insights* adicionais sobre a estrutura intrínseca das áreas de pesquisa. Portanto, a escolha da dimensão é um compromisso entre a facilidade de visualização e a fidelidade na representação dos dados.

Para a visualização de dados, a biblioteca Plotly [21] foi empregada. Esta permite a criação de visualizações interativas dos dados, possibilitando a exploração individual com cada ponto e fornecendo acesso aos metadados de cada artigo.

Por fim, a biblioteca Dash [26] foi utilizada para criar uma interface de usuário interativa para os gráficos gerados, permitindo a filtragem dos dados com base no ano e no número de citações.

V. RESULTADOS EXPERIMENTAIS

Esta seção abrange as descobertas derivadas da implementação da ferramenta proposta, para visualizar o contexto e a evolução de conjuntos de artigos científicos dentro da categoria cs.AI do arXiv. A Figura 1 - acompanhada pela legenda na Figura 2 - ilustra um exemplo prático dessa aplicação, exibindo uma coleção de publicações categorizadas como Ciência da Computação (CS) no arXiv, cujos artigos foram publicados entre 2013 e 2022. Dispostos em um gráfico de dispersão, cada ponto representa um artigo, com seu tamanho correspondendo ao número de citações recebidas, indicando sua relevância dentro da comunidade científica. Este sistema interativo facilita a filtragem e visualização de informações ao longo de algumas facetas, tais como categoria do artigo e ano de publicação. Isso possibilita um acompanhamento da evolução das publicações ao longo do tempo, oferecendo *insights* sobre o desenvolvimento e o progresso de determinadas áreas de pesquisa, como demonstrado na figura 4

O sistema foi alimentado com dados do ArXiv, uma base composta por pré-prints, isto é, artigos que ainda não passaram pelo processo de revisão por pares. A nossa análise, no entanto, se concentra em uma subseção desse banco de dados. Do conjunto original de 2.2 milhões de artigos, apenas 703.085 eram artigos que foram aprovados em processos de revisão por pares, todos validados e publicados em periódicos ou outras

modalidades científicas. Destes 703.085, 1910 artigos podem ser visualizados através da ferramenta.

Os pontos podem ser selecionados para revelar detalhes adicionais de cada publicação, incluindo a possibilidade de acessar o texto completo do artigo diretamente do arXiv. A cor de cada ponto é determinada pela categoria hierárquica superior do artigo, que indicamos como "categoria principal". Por exemplo, se um artigo pertence à subcategoria de "Machine Learning" (cs.LG), sua cor será baseada na categoria principal "Computer Science" (CS). Essas categorias principais permitem uma visão geral das grandes áreas de pesquisa representadas no conjunto de artigos.

Para exemplificar a utilidade da ferramenta, focamos na subárea de Inteligência Artificial (cs.IA). Observamos que mesmo dentro desta área específica, as representações vetoriais foram capazes de capturar agrupamentos e constelações de artigos com temas semelhantes, refletindo a organização intrínseca do campo de estudo. Na figura 3 podemos notar uma constelação de artigos em torno do tema de ética na inteligência artificial. Este agrupamento possui como representante mais relevante o trabalho de Jobin et al. [24], com 1.069 citações (usado como exemplo na Tabela I), ilustrando a capacidade da ferramenta de retratar adequadamente a similaridade entre os tópicos tratados pelos artigos.

É importante destacar que a formação dessas constelações de artigos baseou-se exclusivamente na representação vetorial dos resumos dos artigos. Isso demonstra que, apenas a partir do conteúdo semântico dos textos, é possível capturar a essência do tópico de pesquisa de cada artigo e proporcionar uma representação precisa das áreas e subáreas de estudo.

VI. Conclusão e Trabalhos Futuros

Há um amplo espaço para aprimoramento da ferramenta desenvolvida, com destaque para a necessidade de validação quantitativa dos resultados. Um avanço significativo seria a implementação de uma métrica ou índice capaz de indicar o potencial de inovação de um artigo específico. Com base na habilidade da ferramenta de acompanhar a evolução temporal de um conjunto de artigos, é possível extrair um indicador de inovação com base apenas no resumo do artigo, o que poderia constituir uma contribuição valiosa e se tornar objeto de futuras pesquisas.

Além disso, outra perspectiva para trabalhos futuros envolve a ingestão de dados diretamente do arXiv ou até mesmo outras bases de dados científicas por meio de técnicas de mineração de dados. Isso permitiria, por exemplo, acompanhar a evolução diária de várias áreas de pesquisa, aumentando a relevância e a utilidade da ferramenta para a comunidade científica.

Em resumo, a ferramenta desenvolvida provou ser eficaz para visualizar a estrutura e evolução dos campos de pesquisa científica. Ela oferece aos pesquisadores uma visão abrangente de uma área de estudo e facilita a exploração aprofundada de tópicos específicos. Sua aplicação prática é evidente, permitindo o acesso direto aos textos completos, auxiliando na análise e compreensão das tendências atuais. A relevância e o potencial da ferramenta para contribuir com a pesquisa científica são destacados, e suas possibilidades de expansão e aprimoramento indicam um futuro promissor na área de visualização e análise de dados científicos.

VII. Agradecimentos

Agradecemos pelo apoio do Fundo Mackenzie de Pesquisa e Inovação (MackPesquisa) da Universidade Presbiteriana Mackenzie (UPM)

Referências Bibliográficas

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021.

[2] C. Bhagavatula, R. Power, J. L. Wiandt, V. Khandelwal, and K. Toutanova, "Content-based citation recommendation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2381–2391.

[3] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.

[4] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: A statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, 12 2010.

[5] F. Beck, S. Koch, and D. Weiskopf, "Visual analysis and dissemination of scientific literature collections with survis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 180–189, 2016.

[6] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *Proceedings of the 8th IEEE Pacific Visualization Symposium (PacificVis '15)*, 2015, pp. 117–121.

[7] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," 1967.

[9] H. Adnani, M. Cherraj, and H. Bouabid, "Similarity indexes for scientometric research: A comparative analysis," *Malaysian Journal of Library and Information Science*, vol. 25, no. 3, p. 31–48, Dec. 2020.

[10] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999. [Online]. Available: http://nlp.stanford.edu/fsnlp/

[11] N. Chomsky, *Aspects of the Theory of Syntax*, 50th ed. The MIT Press, 1965. [Online]. Available: http://www.jstor.org/stable/j.ctt17kk81z

[12] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our

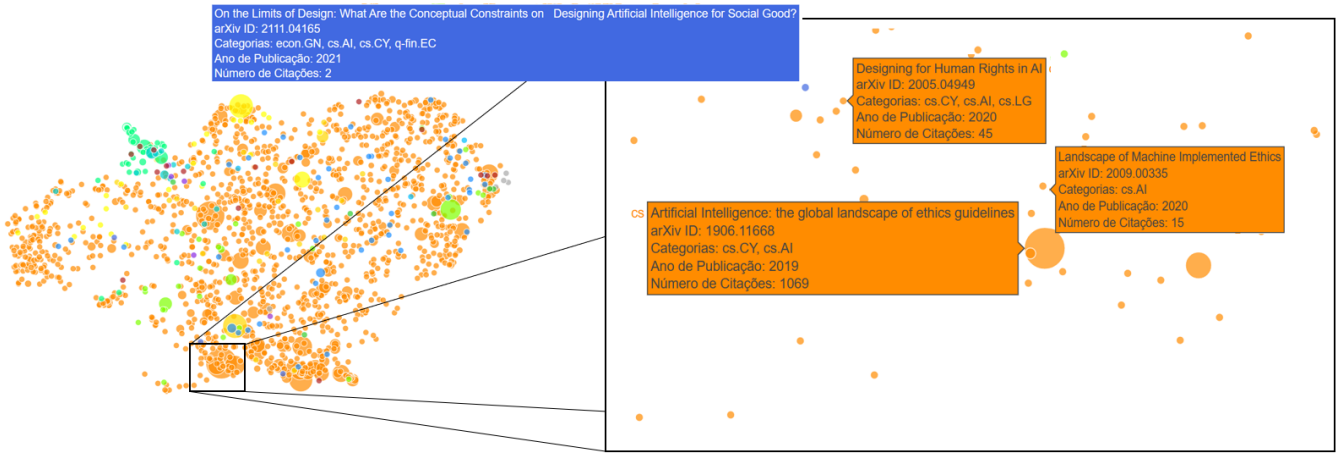

Fig. 3. Estrutura da constelação de trabalhos envolvendo Ética em Inteligência Artificial

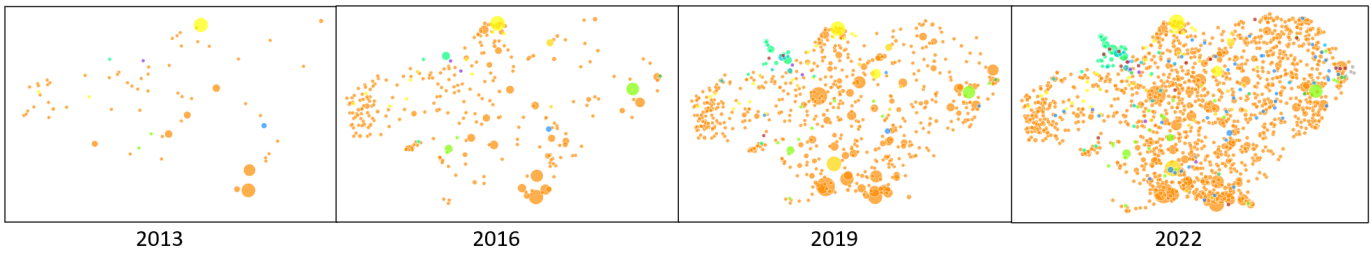

Fig. 4. Evolução da categoria cs.AI a partir de 2013

words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[13] N. Indurkhya and F. J. Damerau, Eds., *Handbook of Natural Language Processing*, 2nd ed. Chapman and Hall/CRC, 2010. [Online]. Available: https://doi.org/10.1201/9781420085938

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017, cite arxiv:1706.03762Comment: 15 pages, 5 figures. [Online]. Available: http://arxiv.org/abs/1706.03762

[17] C. University, "Arxiv.org," https://arxiv.org, [Online; accessed 01-June-2023].

[18] P. S. Foundation, "Python, version 3.10," http://www.python.org, [Online; accessed 01-June-2023].

[19] T. pandas development team, "Pandas," https://pandas.pydata.org/, [Online; accessed 01-June-2023].

[20] H. Face, "Transformers," https://huggingface.co/docs/transformers/index, [Online; accessed 01-June-2023].

[21] P. development team, "Plotly," https://plotly.com/, [Online; accessed 01-June-2023].

[22] Kaggle.com, "Kaggle.com," https://www.kaggle.com, [Online; accessed 01-June-2023].

[23] I. D. Foundation, "Crossref.org," https://www.crossref.org/, [Online; accessed 02-June-2023].

[24] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0088-2

[25] PyTorch, "Pytorch," https://pytorch.org/, [Online; accessed 01-June-2023].

[26] P. development team, "Dash," https://plotly.com/dash/, [Online; accessed 01-June-2023].