Exploring AI and ML Trends through Topic Modeling

Mohan Krishna Pasupuleti **Srilakshmi Umamaheswari Mantena** **Naga jyothi Kota**

Abstract:

In today's fast-growing fields of Artificial Intelligence (AI) and Machine Learning (ML), it's important to keep up with the latest trends. This project helps researchers, academic scholars, and companies understand what's new and where to focus their efforts. By using topic modeling, a method of finding key themes in large amounts of text, we can spot what topics are popular in AI and ML right now. This is really useful for researchers who want to find new areas to study, and for companies deciding where to invest.

Our project uses advanced language processing techniques to analyze a lot of data from different scientific areas. We combine older methods with a newer approach called BERTopic to get a deeper understanding of these trends. This way, we can provide valuable insights for people working in AI and ML, helping them make informed decisions and plan future research.

1. Introduction:

In an age where data is the new oil, the realms of Artificial Intelligence (AI) and Machine Learning (ML) are at the forefront of technological innovation and scientific discovery. As these fields continue to expand and intersect with various domains, the sheer volume and complexity of the information generated present a unique challenge: how to effectively discern and interpret the myriad trends, themes, and patterns that emerge from this wealth of data. This is where our project, centered on advanced topic modeling, steps in. Utilizing cutting-edge Natural Language Processing (NLP) techniques, including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and the transformative BERTopic model, our study aims to delve into the depths of AI and ML literature and metadata to unravel the intricate tapestry of topics that define these dynamic fields.

Our pursuit is driven by the recognition that understanding the evolving landscape of AI and ML is not just an academic exercise, but a necessity for guiding future research, informing industry practices, and shaping educational curricula. By meticulously analyzing and interpreting the themes and trends within AI and ML, our project seeks to offer a comprehensive and coherent narrative of these fields' current state and potential future developments. Through this endeavor, we aspire to bridge the gap between the vast amounts of data available and the actionable insights needed by researchers, practitioners, and decision-makers in AI and ML.

The methodology of our project is as diverse as the data it examines. From traditional topic modeling techniques like LSA and LDA to the more recent BERTopic model, we employ a range of tools to ensure a thorough and nuanced analysis. This multi-faceted approach allows us to capture the subtle variations and complexities inherent in the topics of AI and ML, providing a more accurate and rich understanding of these fields. Additionally, by incorporating the latest developments in NLP and topic modeling, our study not only reflects the current state of the art but also contributes to its advancement.

2.Data:

For our project, we gathered a large amount of information from various sources in the fields of Artificial Intelligence (AI) and Machine Learning (ML). The data included different types of content, such as:

- Scientific Publications and Research Papers: These are articles and papers from the AI and ML community. They cover a wide range of topics, from technical research to theoretical discussions.
- Online Databases and Repositories: We used online platforms that store academic and research papers. These databases provided us with a wealth of current and historical information on AI and ML.
- TED Talk Transcripts: We included transcripts of TED talks related to AI and ML. These talks often present new ideas and perspectives in the field, making them a rich source of current trends and themes.
- Industry Reports and White Papers: To get an industry perspective, we added reports and papers published by companies and organizations working in AI and ML. These documents often discuss practical applications and future directions in the industry.

In our project we utilize the different fields from the dataset. Here is a description of these fields.

- **Title** (Title of the Paper): This field contains the title of each paper. Titles are crucial as they often concisely summarize the main topic or focus of the paper. Analyzing titles helps in quickly identifying the core themes and trends in AI and ML research.
- **Abstract** (The Abstract of the Paper): The abstract provides a brief summary of the paper, including its main objectives, methods, findings, and conclusions. It's a valuable source for extracting detailed insights into the content and focus areas of the research without needing to analyze the full text.
- **Categories** (Categories/Tags in the ArXiv System): These are the categories or tags assigned to each paper within the ArXiv system, a repository of electronic preprints. These tags help classify the papers into various AI and ML subfields, making it easier to group and analyze papers by specific topics or research areas.
- **Versions** (A Version History): The version history indicates the various versions of a paper, reflecting updates or revisions. In our project, we used the history date from these versions to determine the year, which is instrumental in analyzing how topics have evolved over time.

- **Authors** (Authors of the Paper): This field lists the authors of each paper. Analyzing the authors can provide insights into collaborations, leading contributors, and influential researchers in the AI and ML fields. It can also help in understanding the geographical and institutional distribution of research in these areas.

3.Methodology:

This project follows a detailed plan to accurately understand and analyze the most important subjects in the fields of Artificial Intelligence (AI) and Machine Learning (ML). The approach involves several main steps:

3.1. Data Collection and Preprocessing:

3.1.1 Source Identification: In our project we focused on AI and ML, data sources typically include academic journals, online repositories, conference proceedings, and industry reports.

Hence we select the arXiv Dataset from kaggle which is rich in content about current research, trends, and developments in AI and ML.

3.1.2 Filtering AI and ML Related Content: To focus our analysis on AI and ML, we employed a keyword-based filtering approach. We identified and extracted documents containing specific AI and ML-related keywords and phrases. This filtering ensured that our dataset was highly relevant to the topics of interest.

3.1.3 Year Extraction from Version Field: The 'version' field in our data often contained date information, which was crucial for analyzing trends over time. We developed a script to parse these version fields, extracting the year information. This allowed us to chronologically organize the data, offering insights into the evolution of AI and ML topics through different years.

3.1.4 Processing Author Information: Author data was critical for understanding collaboration patterns and author contributions in AI and ML. We parsed the author fields, standardizing the format and extracting individual author names for further analysis.

3.1.5 Creating New Category Labels: Using the 'category' column in our dataset, we created new categorical labels: 'AI', 'ML', and 'AI and ML'. This categorization was based on the presence of specific keywords and themes in each document. This new categorization allowed us to segment our analysis more distinctly between AI, ML, and interdisciplinary areas involving both.

3.1.6 Plotting Abstract Length: To understand the depth and detail of our dataset, we analyzed the length of abstracts. By plotting the length distribution, we gained insights into the level of detail and complexity of documents in our dataset. This step was instrumental in assessing the data's comprehensiveness and deciding on further preprocessing needs.

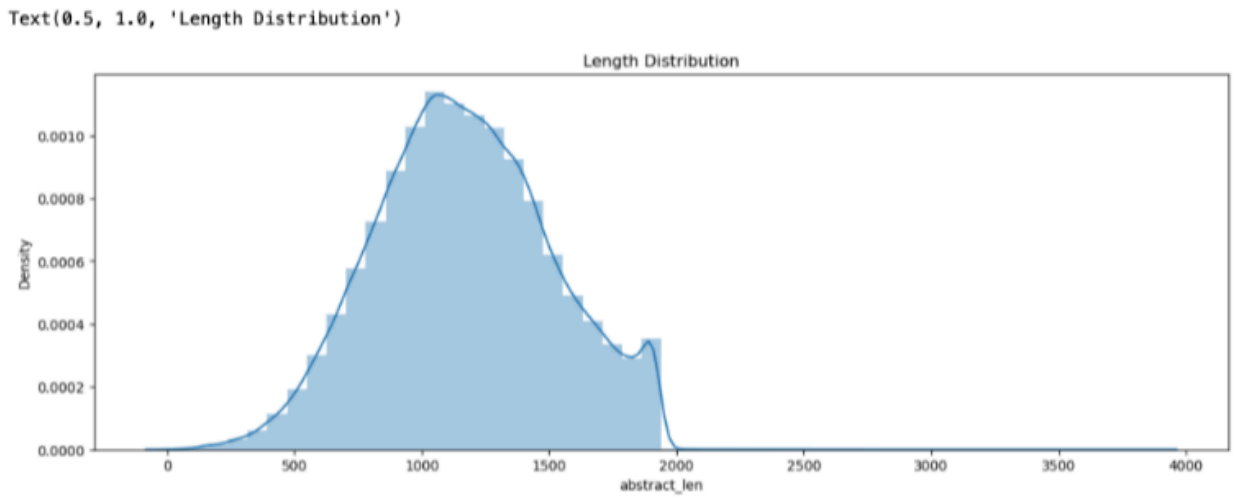

Figure 1: Abstract length graph

3.1.7 Removal of Punctuations: The initial step in our preprocessing was to clean the text data by removing punctuations. This step was essential to eliminate unnecessary characters that could interfere with the textual analysis. By stripping away punctuation marks, we ensured that the text was streamlined for further processing.

3.1.8 Elimination of Stop Words and Lemmatization: Following the punctuation removal, we addressed the presence of stop words – common words that, while integral to sentence structure, often add minimal semantic value to the topic modeling process. We utilized a predefined list of stop words and extended it to include domain-specific terms that were irrelevant to our analysis. Alongside stop words removal, we applied lemmatization to the text. This process involved reducing words to their base or dictionary form (lemma). Unlike stemming, lemmatization considers the context and converts the word to its meaningful base form. For example, 'running' would be lemmatized to 'run'. This step was critical in ensuring that variations of a word were recognized as a single term, providing a more accurate representation in our topic modeling. To implement these steps, we employed NLP libraries, which allowed us to streamline the process and apply these transformations efficiently to our entire dataset.

3.1.9 Visualization of Top N Words: To gain an initial understanding of the most prevalent terms in our dataset, we created visualizations of the top N words. This visualization not only provided a quick overview of the dominant themes in the data but also helped us in fine-tuning our preprocessing steps. By identifying the most frequent words, we could better understand the context and focus areas of our dataset, guiding our subsequent analysis.

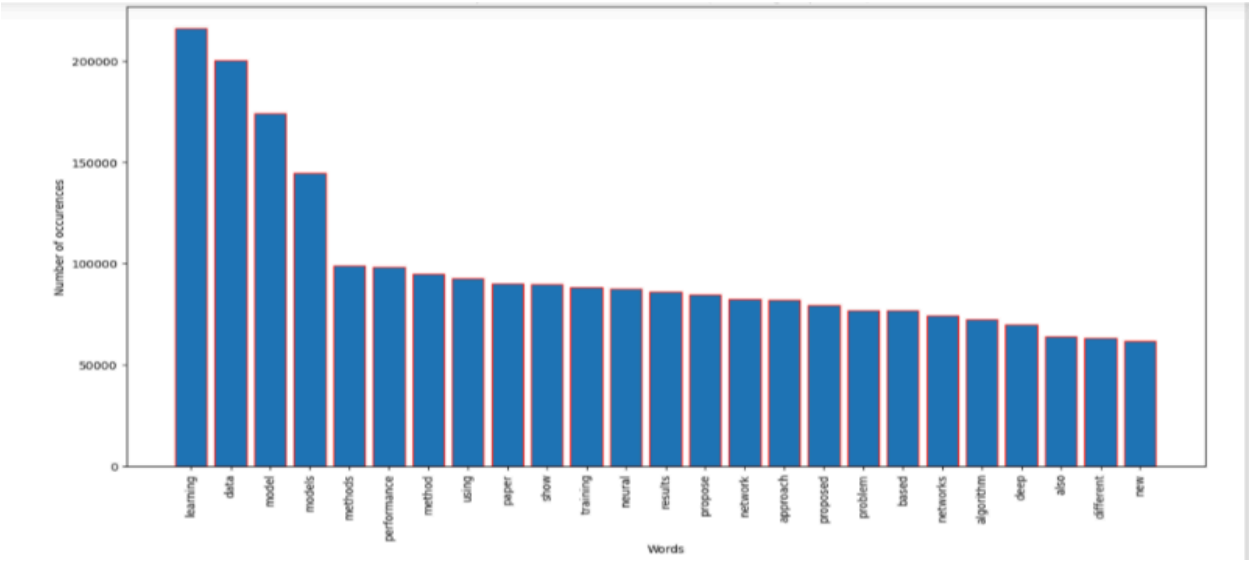

Figure 2: Top 25 words and frequency

3.1.10 Growth of AI and ML Over the Years: Our analysis began with extracting the year information from each document in our dataset. This step was crucial to track the temporal progression in AI and ML research. We then categorized the documents based on whether they pertained to AI, ML, or both, using keyword-based filtering. By aggregating the number of papers per year for each category, we were able to visualize the growth trends in AI and ML. This analysis provided us with insights into the evolution of interest and research intensity in these domains over time. The growth trend analysis revealed significant insights into how the focus areas and research intensity in AI and ML have shifted and expanded over the years.

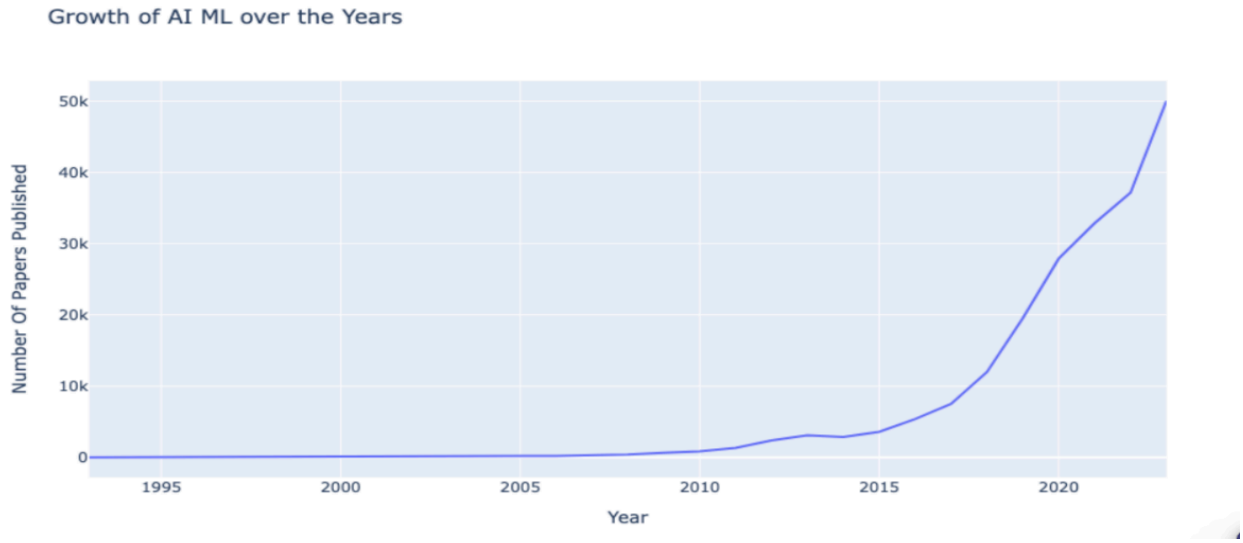

Figure 3: Growth of AI and ML over Years

3.1.11 Number of Papers Published Each Year: To quantify the research output in AI and ML, we calculated the total number of papers published each year. This involved aggregating the data based on the year of publication, providing us with a clear picture of the research trajectory in AI and ML. The annual publication count allowed us to assess the growing or waning interest in these fields and identify any notable trends or patterns. The visualization of this data helped in easily communicating these trends to a broader audience, showcasing the dynamic nature of AI and ML research.

	Year	new_category_label	Number of Papers
84	2023	Machine Learning	27259
83	2023	Artificial Intelligence and Machine Learning	11394
82	2023	Artificial Intelligence	11390
81	2022	Machine Learning	22514
80	2022	Artificial Intelligence and Machine Learning	7914
..	...	...	...
6	1997	Machine Learning	1
3	1996	Artificial Intelligence	28
2	1995	Artificial Intelligence	27
1	1994	Artificial Intelligence	14
0	1993	Artificial Intelligence	6

[85 rows x 3 columns]

Figure 4: Count of papers published each year

3.2. Topic Modeling Techniques:

3.2.1 Latent Semantic Analysis (LSA)[1]:

We used LSA as one of our primary topic modeling techniques. LSA is a method that helps in identifying patterns in the relationships between the documents and the terms they contain. By applying singular value decomposition (SVD) to our document-term matrix, we could extract latent themes from our dataset. LSA was particularly useful in reducing the dimensionality of our data, making it more manageable while preserving the essential structures for topic identification.

```
Topic 0:

learning
data
model
models
based
training
performance
methods
method
using

Topic 1:

data
model
models
training
privacy
synthetic
real
sets
augmentation
clustering

Topic 2:

learning
data
machine
algorithms
algorithm
reinforcement
policy
deep
supervised
rl
```

Figure 5: LSA output

3.2.2 Latent Dirichlet Allocation (LDA)[2]: In our project, we employed Latent Dirichlet Allocation (LDA) for topic modeling to uncover the underlying themes in our AI and ML dataset. Here's an overview of the key steps and methodologies we used in our LDA analysis:

3.2.2.1 Calculating Coherence Score: To assess the quality of the topics generated by our LDA model, we calculated the coherence score. This metric helped us determine how meaningful and interpretable the topics were. We experimented with different numbers of topics and hyperparameters, using the coherence score as a guide to find the optimal model configuration. The coherence score played a pivotal role in refining our LDA model, ensuring that the topics were both statistically sound and interpretable.

3.2.2.2 Implementing LDA: We implemented the LDA model using a well-established NLP library, ensuring robust and efficient topic modeling. The model was trained on our preprocessed dataset, where it learned to group words into topics based on their distribution across documents.

3.2.2.3 Identifying Dominant Topics: Once the LDA model was trained, we analyzed the output to identify the dominant topics in each document. This involved mapping each document to the topic it most strongly associated with, based on the distribution of topic probabilities. This step was crucial for understanding which topics were most prevalent in our dataset and how they were represented across different documents.

3.2.2.4 Displaying Topics: We displayed the topics generated by the LDA model, listing the most representative words for each topic. This provided a clear and concise view of the thematic content of each topic. The displayed topics were then used for further qualitative analysis, helping us interpret and understand the broader themes in our dataset.

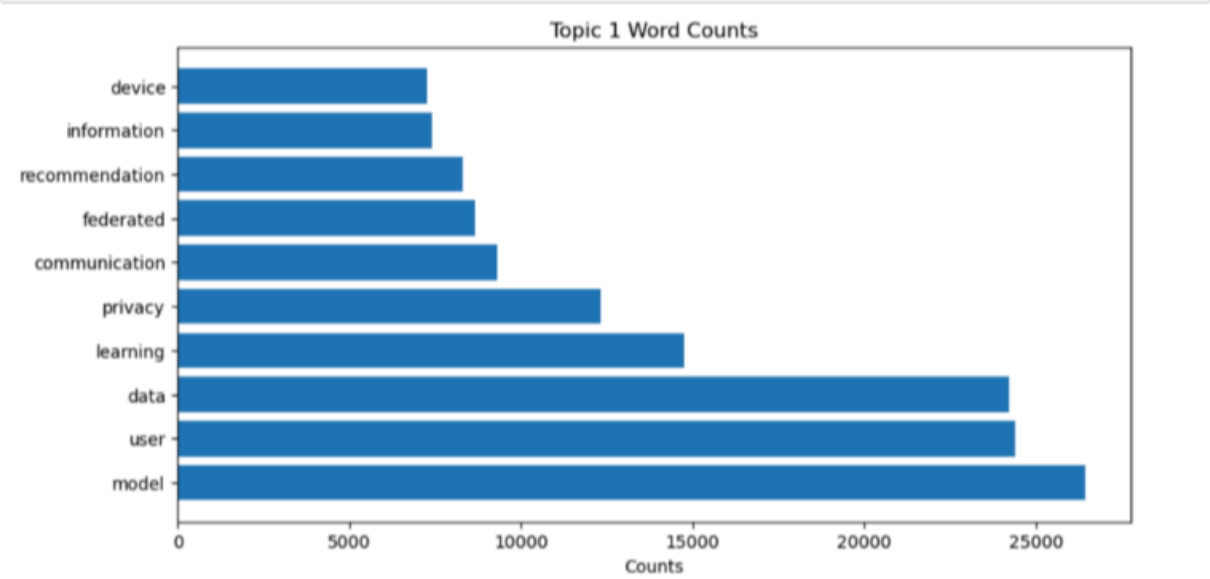

Figure 6: Output of LDA for topic 1

3.2.2.5 Using Word Clouds for Visualization: To visually represent the results of our LDA analysis, we created word clouds for each topic. These word clouds highlighted the most frequent and significant words within each topic, offering an intuitive and engaging way to visualize the data.The word clouds served as an effective tool for communicating our findings, making the complex results of the LDA analysis accessible to a broader audience.

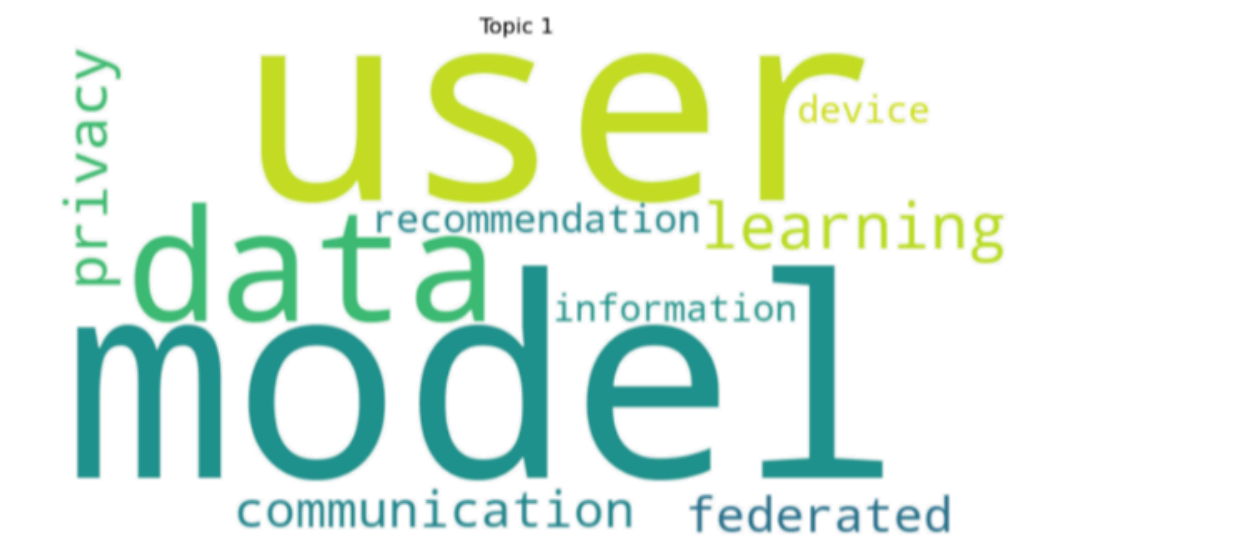

Figure 7: Output of Topic 1

3.2.3 BERT Topic Modeling Implementation

In our project, we leveraged BERTopic [3], a topic modeling technique that utilizes the BERT (Bidirectional Encoder Representations from Transformers) model, known for its exceptional language processing capabilities. BERTopic combines the power of transformer models and C-TF-IDF to create dense clusters, leading to highly interpretable topics while retaining important words in descriptions. This method was particularly effective for our deep, data-driven analysis of the arXiv dataset, focusing on AI and ML-related metadata. By fine-tuning BERT with this domain-specific collection, we successfully discovered various intricate topics, demonstrating the advantages of BERTopic in contextual understanding and adaptability. In our exploration of AI and ML trends, we integrated the cutting-edge capabilities of BERT, a transformer-based model known for its superior understanding of context and language nuances. Here's how we incorporated BERT into our topic modeling project:

3.2.3.1 Model Selection and Setup: We chose a pre-trained BERT model due to its state-of-the-art performance in processing natural language. This decision was grounded in BERT's ability to capture the context and intricacies of language, which is paramount for accurate topic modeling. The setup involved configuring the BERT model with our dataset, ensuring it was primed to analyze and understand our specific corpus related to AI and ML.

3.2.3.2 Data Processing for BERT: Our dataset underwent a series of preprocessing steps tailored for BERT, including tokenization and encoding. This process converted our textual data into a format that BERT could efficiently process and analyze..

3.2.3.3 Topic Extraction Using BERT: With the model set up and data preprocessed, we utilized BERT to extract topics from our dataset. The model's bidirectional nature allowed it to

understand the context of words in a sentence more effectively than traditional unidirectional models, leading to more nuanced topic identification. The topics identified by BERT were both diverse and representative of the underlying content, showcasing the model's proficiency in handling complex linguistic structures.

3.2.3.4 Refinement and Evaluation: We iteratively refined the BERT model's parameters, including adjusting the number of topics and tuning other hyperparameters, to optimize the quality and coherence of the topics generated. The evaluation of BERT's performance in topic modeling was conducted through qualitative assessments and coherence measurements, ensuring that the topics were not only statistically significant but also meaningful and interpretable.

Figure 8: Output of BERT for Topics

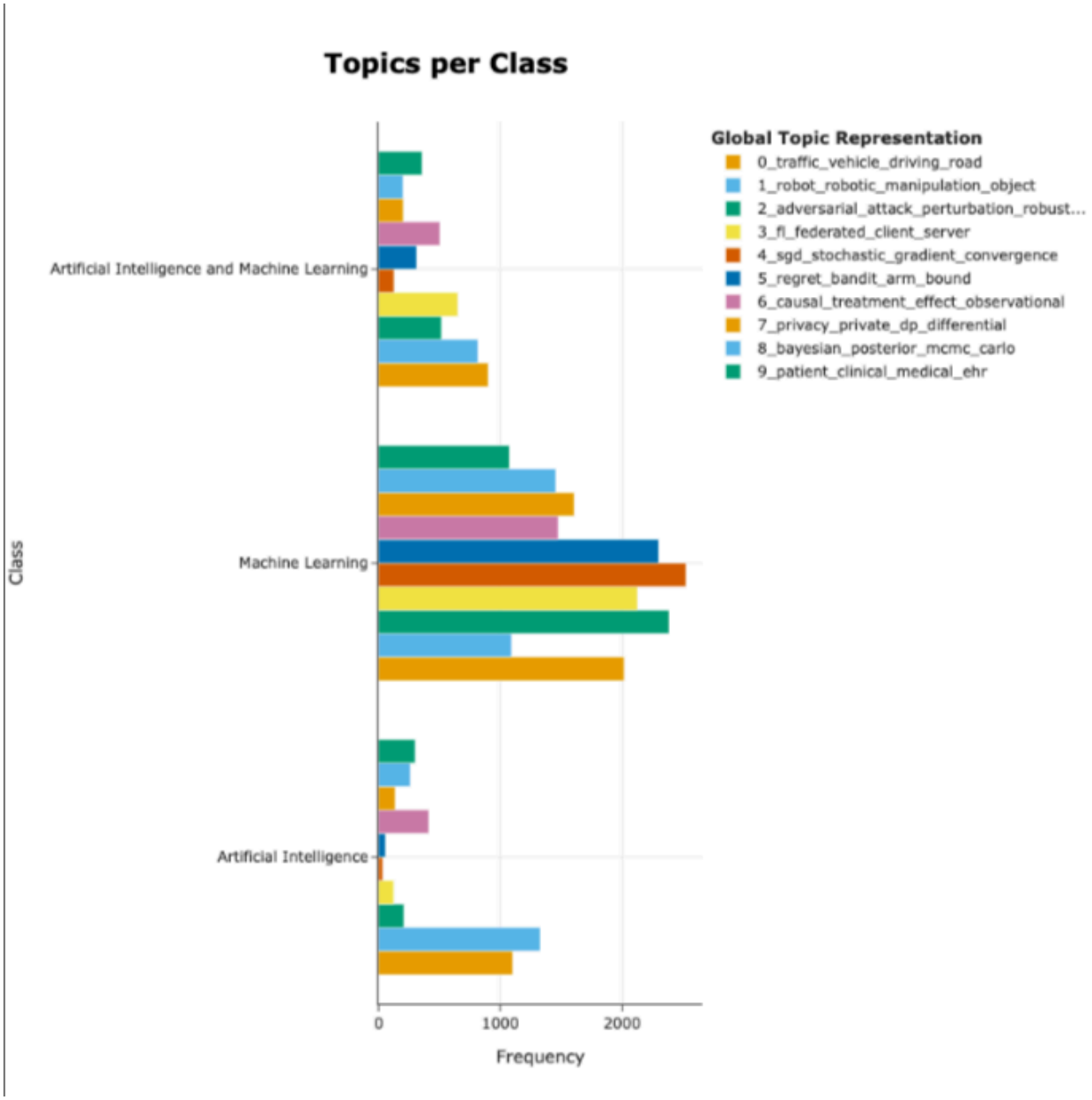

Figure 9: Topics per class output

Related Work:

The field of topic modeling in AI and Machine Learning (ML) has been extensively explored, with numerous significant contributions that have shaped current understanding and methodologies. This section reviews key works that have informed and influenced this project.

An Introduction to Latent Semantic Analysis [11], this paper on LSA is critical for understanding the method's application in our project's dimensionality reduction and topic discovery.

Probabilistic Latent Semantic Analysis [12], Presents pLSA, a probabilistic version of LSA, enhancing our project's topic modeling accuracy. A Correlated Topic Model of Science [13], discusses advanced topic models, contributing to our project's approach in handling correlated topics in AI and ML.Finding Scientific Topics [14], provides insights into using probabilistic methods for scientific topic discovery, relevant to our project's focus on AI and ML research. Comparing Twitter and Traditional Media Using Topic Models [16], Offers insights into different media content analysis, relevant for our project's comparative study of AI and ML topics. Short and Sparse Text Topic Modeling via Self-Aggregation [17] discusses handling short and sparse texts in topic modeling, applicable to our project's diverse data types. Improving Topic Models with Latent Feature Word Representations [18], Explores topic model improvements, enhancing the sophistication of our project's topic discovery process. Topic Modeling Over Short Texts by Incorporating Word Embeddings [19], examines topic modeling in short texts, relevant for our project's analysis of abstracts and short documents.

Results:

The findings of the topic modeling project reveal insightful trends and themes in the AI and ML domains, providing a comprehensive view of the current state and future directions of these fields. The key results are as follows:

Identification of Core Topics: The project successfully identified a diverse range of topics within AI and ML, including but not limited to neural networks, deep learning, AI ethics, and AI applications in various industries like healthcare and finance. Our project highlights the broad scope and interdisciplinary nature of AI and ML research.

The comparison between LSA, LDA, and BERTopic models showed varying levels of effectiveness in topic identification. BERTopic, with its advanced contextual understanding, was particularly effective in identifying nuanced and emerging topics, providing a deeper insight into the field.

LSA and LDA offered a foundational perspective on common and recurrent themes but were less effective in capturing more complex topics.

Temporal Trends and Evolution of Topics:

A significant aspect of the project was analyzing how topics have evolved over time. There was a noticeable shift from foundational ML techniques to more advanced topics like deep learning, AI ethics, and real-world applications, reflecting the rapid advancement in the field.

The project also highlighted emerging trends, such as the increasing focus on ethical implications of AI and the integration of AI with other cutting-edge technologies.

Quantitative Analysis: Coherence and Perplexity Scores:

- The topics generated by each model were quantitatively evaluated using coherence and perplexity scores. BERTopic generally showed higher coherence scores, indicating more meaningful and cohesive topic representation.
- LDA demonstrated reasonable perplexity scores, suggesting a good statistical fit to the data, albeit sometimes at the cost of missing nuanced topics.

Visual Representations of Topics:

- Visual tools such as word clouds and PyLDAVis were employed to represent the topics visually, aiding in the interpretation and understanding of the results. These visualizations provided a clear and intuitive way to grasp the relationships and significance of different topics.

Insights from Qualitative Assessments:

- Qualitative assessments reinforced the quantitative findings, with BERTopic's results aligning closely with current trends and discussions in the AI and ML community compared to the LDA's performance but the coherence score of LDA was higher than that of Bertopic. This proves that quantitative evaluation is not enough and qualitative evaluation is needed as proved in the paper "Is Automated Topic Model Evaluation Broken? " [20]

Comparison with Expert Insights:

- The topics identified were also cross-verified with insights from domain experts. This validation process confirmed the relevance and accuracy of the topics, ensuring that the project's findings were aligned with current expert knowledge in AI and ML.

Discussion:

The results of the topic modeling project offer significant insights into the current landscape and future trends in AI and ML. This section discusses the implications of these findings, the strengths and limitations of the methodologies used, and potential areas for future research.

Interpretation of Identified Topics:

- The wide array of topics identified, from technical aspects like neural networks to broader issues like AI ethics, highlights the diverse and interdisciplinary nature of AI and ML research. This diversity underscores the need for a holistic approach in AI and ML education and policy-making.
- The emergence of specific themes, such as the ethical implications of AI and the integration of AI with other technologies, points to a maturation in the field, where ethical and practical considerations are gaining prominence alongside technological advancements.

Comparative Efficacy of Topic Modeling Techniques:

- The project showcased the varying capabilities of LSA, LDA, and BERTopic in topic identification. While LSA and LDA provided a solid foundation for

identifying prevalent themes, BERTopic's context-aware approach proved more adept at uncovering nuanced and emerging topics.
- This comparative analysis highlights the importance of selecting appropriate topic modeling techniques based on the specific objectives and characteristics of the dataset.

Evolution and Trends in AI and ML:

- The temporal analysis of topics revealed a shift in focus within the AI and ML fields. This evolution from foundational concepts to advanced applications and ethical considerations reflects the rapid pace of change and the increasing complexity of challenges in these domains.
- These findings are vital for researchers and practitioners in AI and ML, as they provide direction for future research and development efforts.

Quantitative Metrics: Coherence and Perplexity:

- The coherence and perplexity scores served as important quantitative measures for evaluating the quality of the topics generated. However, these metrics have their limitations and should be complemented with qualitative assessments to fully grasp the effectiveness of the topic modeling techniques.

Implications for Research and Practice:

- The project's results have significant implications for academic research, industry practices, and policy-making in AI and ML. Understanding the evolving trends can help steer research in directions that are both innovative and aligned with societal needs.
- The insights gained can also inform curriculum development in educational institutions, ensuring that students are exposed to both the foundational aspects and the latest developments in AI and ML.

Limitations and Future Research Directions:

- One limitation of the project is the potential bias in the dataset, which might influence the topics identified. Future research could expand the dataset to include more diverse and comprehensive sources.
- There is also scope for incorporating additional topic modeling techniques and exploring hybrid models that combine the strengths of different approaches.
- The integration of ChatGPT API and other advanced NLP tools in future iterations could further enhance the depth and accuracy of the topic modeling process.

Ethical Considerations and Responsible AI:

- The project underscores the importance of ethical considerations in AI research and development. As AI technologies become more pervasive, it is crucial to address issues like bias, privacy, and accountability proactively.
- The discussion of ethical topics in the project reflects a growing awareness in the AI community and points to the need for ongoing dialogue and responsible AI practices.

Conclusion:

This topic modeling project has successfully demonstrated the power and versatility of advanced NLP techniques in dissecting the complex landscape of AI and ML research. Through the application of LSA, LDA, and particularly BERTopic, the project has unearthed a rich tapestry of themes that define current research and discourse in these fields. The findings reveal not only the technical and scientific advancements but also the increasing significance of ethical, societal, and practical considerations in AI and ML. These insights provide a valuable roadmap for researchers, practitioners, and policymakers, offering a nuanced understanding of where the field currently stands and where it may be headed.

The project underscores the dynamic nature of AI and ML, reflected in the evolving focus areas and emerging trends. This evolution speaks to the continuous advancement and adaptation within these disciplines, driven by technological breakthroughs and changing societal needs. The comparative analysis of different topic modeling techniques also highlights the importance of choosing the right method to match the specific characteristics and objectives of the research.

Future Work:

Looking forward, there are several avenues for extending and enhancing this research:

- Expanding the Dataset: Future iterations of this project could benefit from a more expansive and diverse dataset, incorporating a wider range of sources, including more recent publications and datasets from different geographical regions and languages. This expansion would help in capturing a more global perspective of AI and ML trends.
- Advanced NLP Techniques: Incorporating newer and more sophisticated NLP techniques, such as transformer-based models like GPT-3 or BERT variations, could provide even deeper insights. These models, known for their superior understanding of context and semantics, could further refine topic detection and analysis.
- Hybrid Modeling Approaches: Exploring hybrid models that combine the strengths of different topic modeling techniques could lead to more accurate and comprehensive topic identification and trend analysis. Using BERTopic and ChatGPT API for text summarization.
- Integration of ChatGPT API: Utilizing APIs like ChatGPT for enhanced topic modeling and trend analysis could add another layer of depth to the study, particularly in understanding and generating nuanced text interpretations.
- Temporal Dynamics and Predictive Analysis: Further studies could focus on the temporal dynamics of topics, employing predictive models to forecast future trends and shifts in AI and ML research.
- Interdisciplinary Applications: Applying the findings of this project in interdisciplinary contexts, such as AI's role in healthcare, finance, or environmental studies, could yield interesting insights into how AI and ML are influencing various sectors.

Ethical and Societal Impact Studies: Given the prominence of ethical considerations in the findings, dedicated studies on the societal impact and ethical implications of AI and ML would be both timely and beneficial.

Collaborative Research Initiatives: Engaging in collaborative research with academic and industry partners could provide diverse perspectives and expertise, enriching the quality and impact of the research.

References:

1. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). "Indexing by Latent Semantic Analysis." Journal of the American Society for Information Science, 41(6), 391–407. http://wordvec.colorado.edu/papers/Deerwester_1990.pdf
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). "Latent Dirichlet Allocation." Journal of Machine Learning Research, 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
3. Grootendorst, M. (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." https://arxiv.org/pdf/2203.05794.pdf
4. "Topic Modelling Using ChatGPT API." Towards Data Science. https://towardsdatascience.com/topic-modelling-using-chatgpt-api-8775b0891d16
5. Abhinandan Udupa; K N Adarsh; Anvitha Aravinda; Neelam H Godihal; N Kayarvizhy .(2022) "An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling" Publisher: IEEE Conferences -2022 Fourth International Conference. https://ieeexplore.ieee.org/abstract/document/10058687.
6. Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., & Kaymak, U. (2023) "Towards Interpreting Topic Models with ChatGPT" Paper presented at The 20th World Congress of the International Fuzzy Systems Association, Daegu, Korea, Republic of https://pure.tue.nl/ws/portalfiles/portal/300364784/IFSA_InterpretingTopicModelsWithChatGPT.pdf
7. **David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin(2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**:Describes BERT, a method for pre-training language representations that has significantly advanced the field of NLP, including topic modeling. https://arxiv.org/pdf/1810.04805.pdf
8. **Newman, D., et al. (2010): David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. (2010).** Automatic Evaluation of Topic Coherence: This paper is pivotal in the field of topic modeling as it introduces methods for the automatic evaluation of topic coherence. The authors develop a framework for assessing the interpretability of topics generated by topic models, providing a quantitative approach to evaluate the coherence of

topics, which is crucial for determining the quality and usability of topic models in various applications. https://aclanthology.org/N10-1012.pdf

9. **Latent Dirichlet Allocation (LDA) and topic survey modeling:** models, application and survey :Describes topic modeling using Latent Dirichlet Allocation (LDA) from 2003 to 2016, focusing on its development, trends, and applications in various fields. https://www.researchgate.net/publication/321069759_Latent_Dirichlet_Allocation_LDA_and_Topic_modeling_models_applications_a_survey
10. **Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification:** Recurrent Convolutional Neural Networks for Text Classification: Discusses RCNNs for text classification, providing insights into deep learning approaches applicable in our topic modeling tasks. https://ojs.aaai.org/index.php/AAAI/article/view/951
11. **Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis**. https://www.researchgate.net/publication/200045222_An_Introduction_to_Latent_Semantic_Analysis
12. **Hofmann, T. (1999). Probabilistic Latent Semantic Analysis: https://arxiv.org/abs/1301.6705**
13. **Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. https://scholar.google.com/scholar_url?url=https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.pdf&hl=en&sa=X&ei=tTp9Zd-IJNyWy9YP7Omb4AU&scisig=AFWwaeaz6tEmYYzBJjG_vPRPlVZP&oi=scholarr**
14. **Griffiths, T. L., & Steyvers, M. (2004). Finding Scientific Topics. https://www.pnas.org/doi/full/10.1073/pnas.0307752101**
15. **Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. https://repozitar.cz/publication/15725/?lang=cs;kod=S530**
16. **Zhao, W. X., et al. (2011). Comparing Twitter and Traditional Media Using Topic Models. https://scholar.google.com/scholar_url?url=https://ink.library.smu.edu.sg/cgi/viewcontent.cgi%3Farticle%3D2374%26context%3Dsis_research&hl=en&sa=X&ei=YTt9ZZjlBZesy9YPtMyBqAk&scisig=AFWwaeYO_jLty32m-lE_OgOyH7mD&oi=scholarr**
17. **Quan, X., et al. (2015). Short and Sparse Text Topic Modeling via Self-Aggregationhttps://scholar.google.com/scholar_url?url=https://scholars.cityu.edu.hk/files/86898031/321.pdf&hl=en&sa=X&ei=kTt9ZcbYMOrBy9YPpLCfqAU&scisig=AFWwaeYgKAbM_HhXIZ4DjuTiWkgC&oi=scholarr**
18. **Nguyen, D. Q., et al. (2015). Improving Topic Models with Latent Feature Word Representations.Published in *Transactions of the Association for Computational Linguistics* (2015) https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00140/43288/Improving-Topic-Models-with-Latent-Feature-Word**

19. **Qiang, J., et al. (2017). Topic Modeling Over Short Texts by Incorporating Word Embeddings. https://link.springer.com/chapter/10.1007/978-3-319-57529-2_29.**
20. **Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, Philip Resnik *"Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence"* in 35th Conference on Neural Information Processing Systems,2021.**
https://users.umiacs.umd.edu/~jbg/docs/2021_neurips_incoherence.pdf