Kumar, Deepak; Weissenberger-Eibl, Marion

Conference Paper — Manuscript Version (Preprint)

Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights

Suggested Citation: Kumar, Deepak; Weissenberger-Eibl, Marion (2024) : Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights, R&D Management Conference 2024, 17-19 June, 2024, Stockholm, KTH Royal Institute of Technology, Fraunhofer-Gesellschaft, München, https://doi.org/10.24406/publica-3456

Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights

1st M.Sc. Deepak Kumar
Joint Innovation Hub
Fraunhofer ISI
Heilbronn, Germany
deepak.kumar@isi.fraunhofer.de

2rd Univ.-Prof. Dr. Marion Weissenberger-Eibl
Joint Innovation Hub
Fraunhofer ISI
Heilbronn, Germany
Weissenberger-Eibl@isi.fraunhofer.de

***Abstract*—In the fast-paced realm of technological evolution, accurately forecasting emerging trends is critical for both academic inquiry and industry application. Traditional trend analysis methodologies, while valuable, struggle to efficiently process and interpret the vast datasets of today's information age. This paper introduces a novel approach that synergizes Generative AI and Bidirectional Encoder Representations from Transformers (BERT) for semantic insights and trend forecasting, leveraging the power of Retrieval-Augmented Generation (RAG) and the analytical prowess of BERT topic modeling. By automating the analysis of extensive datasets from publications and patents, the presented methodology not only expedites the discovery of emergent trends but also enhances the precision of these findings by generating a short summary for found emergent trends. For validation, three technologies—reinforcement learning, quantum machine learning, and Cryptocurrencies— were analysed prior to their first appearance in the Gartner Hype Cycle. Research highlights the integration of advanced AI techniques in trend forecasting, providing a scalable and accurate tool for strategic planning and innovation management. Results demonstrated a significant correlation between model's predictions and the technologies' appearances in the Hype Cycle, underscoring the potential of this methodology in anticipating technological shifts across various sectors.**

***Index Terms*—BERT, Topic modelling, RAG, Gartner Hype Cycle, LLM, BERTopic.**

I. Introduction

In an era characterized by rapid technological advancements and an ever-increasing volume of data, the ability to forecast technological trends accurately is more crucial than ever for both academia and the industrial sector. Traditional trend analysis methodologies, primarily manual or semi-automated, have been integral in understanding past and current technological trajectories, discussed further in section II. However, these methods are increasingly inadequate in the face of the exponential growth of scientific and patent literature, struggling to process vast datasets and often missing nuanced semantic relationships. The need for innovation in trend forecasting methodologies is clear, not only to keep pace with the speed of technological innovation but also to provide strategic insights that can inform proactive decision-making and innovation management.

The advent of advanced natural language processing (NLP) technologies and the emergence of Generative AI offer promising avenues for AI based trend analysis. Particularly, the integration of Generative AI with Bidirectional Encoder Representations from Transformers (BERT) represents a significant leap forward. Section III, presents a pioneering approach to trend forecasting that leverages the semantic analysis capabilities of BERT and the dynamic, contextually aware text generation of Retrieval-Augmented Generation (RAG) models. By automating the extraction and analysis of data from the Dimensions and arXiv database—encompassing publications and patents, the presented methodology not only streamlines the identification of emerging trends but also enriches the interpretability and applicability of these insights. The methodology is validated through a comparative analysis with the Gartner Hype Cycle, highlighting its effectiveness in predicting the emergence and trajectory of key technological trends.

This paper systematically explores the research conducted. Section II reviews the state-of-the-art in trend analysis methodologies, highlighting the evolution and pinpointing the gaps addressed by this study. Section III details the proposed methodology for trend forecasting, emphasizing the integration of BERT and Generative AI technologies for enhanced trend identification and analysis. In Section IV, the process of dataset compilation is described, focusing on selection criteria and preprocessing steps to ensure data integrity and relevance. Section V presents the study's findings, demonstrating the methodology's application to real-world data and its effectiveness in uncovering emergent trends. The discussion in Section VI reflects on the study's implications and contributions to both academia and industry. Finally, Section VII suggests future research directions, aiming to improve the precision and applicability of AI-driven trend forecasting.

II. State-of-the-art Trend Analysis Methodologies

Trend analysis methodologies have significantly evolved, transitioning from manual to automated processes to cater to the growing complexities of data and information in various

domains. Initially, manual trend analysis, reliant on expert intuition and labor-intensive data sorting, provided foundational insights but was hampered by scalability and subjectivity limitations. The advent of automated techniques, fueled by advances in computing power and artificial intelligence, marked a paradigm shift. These methods employ sophisticated algorithms to sift through vast datasets, identifying patterns and trends with greater speed, accuracy, and objectivity than ever before. Thus, manual and automated methods form the contemporary pinnacle of trend analysis, each reflecting the technological and methodological advancements of their era and setting the stage for future innovations.

A. Manual

Before the advent of sophisticated computational tools and methodologies, trend analysis was primarily a manual process, deeply reliant on human intuition, experience, and analytical skills. As highlighted by the National Research Council (2010), in the era of manual analysis, experts and analysts would sift through data sets, often presented in tables, charts, or graphs, to identify patterns, shifts, and emerging trends in various domains such as finance, market behaviors, and technological advancements [1].

1) Challenges of Manual Trend Analysis: The effectiveness of manual trend analysis, as noted by the National Research Council (2010), diminished with the advent of the digital age, due to challenges such as scalability, subjectivity, time consumption, and limited data comprehensiveness [1]. Scalability issues arose as the volume of data outpaced manual processing capabilities. The reliance on human judgment introduced subjectivity, while the labor-intensive nature of the analysis proved time-consuming and inefficient [1]. Moreover, the limited capacity for data processing often led to incomplete trend predictions, as analysts could only work with manageable subsets of data.

These challenges highlighted the inadequacies of manual trend analysis in the face of burgeoning datasets and the complex dynamics of global markets and technologies. Consequently, the drive towards automated, sophisticated analytical methods grew, aiming to overcome these limitations by leveraging computational power and artificial intelligence for more scalable, objective, and efficient trend analysis.

B. Automated

Automated trend analysis represents a significant shift from traditional manual methods, tapping into the capabilities of advanced artificial intelligence (AI) and machine learning techniques to efficiently and accurately identify emerging trends within large datasets. As Wang (2017) demonstrates, these automated methodologies utilize AI-driven data mining models, with Latent Dirichlet Allocation (LDA) being one prominent example [2].

Negara et al. (2019) further illustrate the application of LDA showcasing its utility in extracting relevant topics from social media conversations [3]. LDA treats each document as a mixture of topics, identifying prevalent themes through a generative process. This method conceptualizes documents as outcomes of random processes involving hidden topics, where each topic is characterized by a distribution over words.

Blei et al. (2003) provided a foundational framework for LDA, detailing a generative process that includes choosing a distribution over topics from a Dirichlet distribution for each document and selecting topics and words for each word in the document from multinomial distributions [4]. This mathematical formulation underpins the approach that enables the identification of thematic structures across large text corpora.

The generative process of LDA, as defined by Blei et al. (2003), starts with each document determining its number of words N, chosen from a Poisson distribution, followed by the selection of a topic distribution θ from a Dirichlet distribution [4]. Each word in the document is then generated by first picking a topic from θ and subsequently a word from that topic's specific word distribution (Blei et al., 2003).

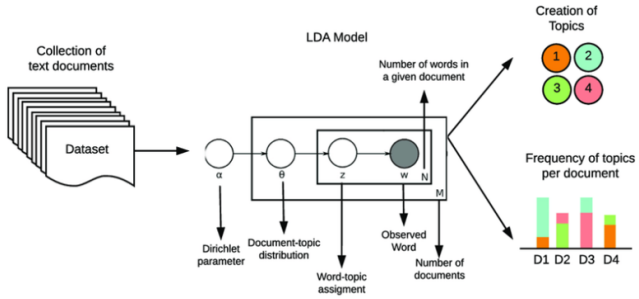

Fig. 1. Topic Modeling workflow of LDA [10]

Mathematically, the generative process for a document w in a corpus D under the LDA model is as follows:

1) Each document is presumed to be generated by first deciding on the number of words it will contain $N \sim \text{Poisson}(\xi)$, drawn from a Poisson distribution.
2) A distribution over topics within a document, denoted by θ, is chosen from a Dirichlet distribution, symbolized as $\theta \sim \text{Dirichlet}(\alpha)$. Here, α represents the Dirichlet prior parameter that influences the distribution of topics θ across the document.
3) For each word in the document, the process is as follows:
   a) A topic z_n is chosen for each word w_n from a multinomial distribution parameterized by the topic distribution θ for the document, denoted as $z_n \sim \text{Multinomial}(\theta)$.
   b) Subsequently, a word w_n is selected based on the chosen topic's distribution over words, parameterized by β, represented as w_n from $p(w_n|z_n, \beta)$. Here, β characterizes the distribution of words for the topic z_n.

The joint distribution of the topic mixture θ, topics z, and words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

The marginal distribution of a document, integrating over θ and summing over z, is obtained as:

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta$$

This robust modeling capability makes LDA invaluable for applications in document classification and the exploration of thematic structures across extensive datasets. However, the effectiveness of LDA hinges on the careful selection of model parameters, which underscores the importance of analytical rigor and domain expertise in deriving meaningful insights (Blei et al., 2003).

1) Limitations of Automated Approaches: Despite the utility, LDA come with limitations that are critical for their application in the topic modeling tasks (Atagün et al., 2021).

1) Context Ignorance: Both models rely on word frequency for topic assignment, neglecting contextual usage. This oversight can lead to ambiguous interpretations of homonyms or polysemes, thereby affecting topic accuracy.
2) Hyperparameter Complexity: Optimal hyperparameter determination, such as LDA's number of topics, challenges users with its need for extensive experimentation or deep domain knowledge.
3) Scalability Issues: The models' applicability to large-scale datasets is hampered by their computational complexity and memory demands, limiting their use in big data contexts.
4) Short Text Challenges: LDA efficacy decreases with short text corpora, where sparse data complicates accurate topic assignment.

In summary, LDA limitations highlight the necessity for advancements that address word context sensitivity, simplify hyperparameter tuning, enhance scalability, and improve short text analysis. These developments are essential for evolving more effective, accurate, and adaptable topic modeling techniques suitable for contemporary text analysis demands.

III. Introduction to BERT and RAG

Building on traditional topic modeling techniques, the integration of Bidirectional Encoder Representations from Transformers (BERT) and Retrieval-Augmented Generation (RAG) presents a transformative approach. This methodology benefits from BERT's deep language understanding for nuanced semantic analysis and leverages RAG for dynamic knowledge retrieval. This combination enhances trend forecasting's precision and comprehensiveness by addressing limitations observed in method like LDA.

A. Bidirectional Encoder Representations from Transformers (BERT)

BERT Topic Modeling advances text analysis by employing its bidirectional nature to fully grasp the context of words, setting a new standard in natural language processing (NLP). Developed by Devlin et al. (2019), BERT has revolutionized how machines understand human language by pre-training on extensive text corpora and subsequent fine-tuning, adapting to various topics and texts with superior performance over traditional models [6]. Koroteev (2021) further explores the applications of BERT in NLP, demonstrating its ability to capture language nuances across different contexts [7].

B. Advantages of BERT in Topic Modeling

The BERT framework introduces several advancements in topic modeling as highlighted by Grootendorst (2022) [11]:

1) Deep Contextual Learning: BERT's architecture allows for an in-depth understanding of context, improving over models that miss context-dependent meanings (Atagün et al., 2021).
2) Dynamic Word Embeddings: Unlike static models, BERT provides context-sensitive word embeddings that effectively capture the subtleties of language.
3) Short Text Analysis: Its robust contextual comprehension aids in analyzing short texts, overcoming the sparsity issues faced by models like LDA.

C. The BERTopic Process

The BERTopic model, as outlined by Atagün et al. (2021), employs a series of sophisticated steps, combining transformer models with class-based TF-IDF, to generate interpretable topic clusters [8]:

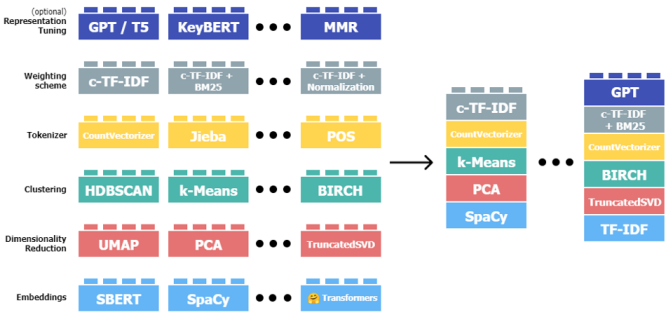

Fig. 2. Sub Steps of BERTopic [11]

The BERTopic process includes:

1) Embedding: Documents are embedded into a numerical space using sentence-transformers, facilitating clustering based on semantic similarity.
2) Dimensionality Reduction: UMAP is utilized to reduce the high-dimensional space, preserving semantic relationships for meaningful clustering.
3) Clustering: HDBSCAN clusters embeddings into topics, effectively identifying distinct groups and outliers.
4) Tokenizer: A bag-of-words model for each cluster enables frequency analysis of terms, enhancing topic representation.
5) Weight Scheming: A class-based TF-IDF transformation discerns word distinctiveness within clusters, providing a refined description of topics.

6) Fine-Tune Representation: Optionally, further fine-tuning of BERT tailors topic representations to specific dataset characteristics.

The BERTopic process effectively combines the state-of-the-art NLP capabilities of BERT with advanced clustering and dimensionality reduction techniques to produce coherent and relevant topics, enhancing the interpretability and usability of the results for trend forecasting and other applications.

D. Retrieval-Augmented Generation (RAG) for Enhanced Analysis

Retrieval-Augmented Generation (RAG) introduces a novel approach in natural language processing by combining the generative capabilities of language models with strategic information retrieval. Finardi et al. (2024) describe how this method integrates a neural language generation model with an information retrieval system within a neural retriever-generator framework [9]. The framework augments predictive text generation by dynamically incorporating information from an external data repository.

The process is modeled as:

$$p(y|x) = \sum_{d \in D} p(d|x)p(y|x,d)$$

where $p(d|x)$ is the probability of retrieving document d based on input x, and $p(y|x,d)$ denotes the probability of generating response y given input x and retrieved document d. The retriever selects documents that maximize relevance to the query, thereby enhancing the information quality for the generator.

Then, a transformer-based generator model incorporates these documents, merging pre-trained knowledge with the retrieved data to produce the final output. This mechanism enables RAG to generate content that is not only coherent and contextually aware but also enriched with domain-specific knowledge beyond the model's initial training data.

RAG significantly advances the generative capabilities of AI, facilitating the creation of informative, accurate content by leveraging both the model's inherent knowledge and external, relevant information. Finardi et al. (2024) contribution underscores the versatility and power of integrating retrieval mechanisms within generative models, pushing the boundaries of how machines understand and generate human language [9].

IV. Dataset

The integrity and relevance of the dataset are fundamental in topic modeling, influencing both the accuracy of topic detection and the depth of insights. The process initiates with an expert-led selection of keywords, ensuring data collection is aligned with emergent trends of interest. This strategic keyword selection and targeted retrieval of content are pivotal, setting the groundwork for substantive analysis.

A. Sources

The dataset consist of publications and patents from Dimensions AI and the arXiv database, recognized for their broad coverage across various disciplines. A keyword-based extraction method is employed to search these repositories for documents featuring the selected keywords in their titles and abstracts.

B. Preprocessing

Preprocessing is vital for text preparation, involving the removal of URLs, punctuation, and normalization to lowercase, thereby reducing noise and standardizing the dataset. Additionally, a stop-word removal process eliminates non-informative words, focusing the analysis on substantive content. This phase also verifies the text's alignment with identified thematic areas, resulting in a dataset primed for advanced topic modeling, facilitating the uncovering of significant thematic structures.

V. Methodology

This research methodology integrates sophisticated AI-driven models for trend forecasting, utilizing a dataset comprising research papers and patents described in figure 3. The approach unfolds in a sequential manner:

Input: Keywords **K** are selected by experts.
Output: **BERT** modelled topics and their explanation using **RAG**.
1 The Database **DB** is scanned with **K** to collect relevant documents (patents and publications).
2 Initialize Dataset **D** as empty.
3 For each document in the database **DB** containing **K**:
4 Add document to Dataset **D**.
5 Dataset **D** is generated.
6 Titles and abstracts are extracted from **D**.
7 The collected dataset **D** undergoes preprocessing.
8 Text data is cleaned and normalized.
9 Stop words, URL, punctuations are removed, and stemming is performed.
10 **BERT** topic modeling is applied to generate topics **T**.
11 Dataset **D** is processed through the **BERT** model.
12 A topic **X** is manually selected from **T**.
13 The most relevant topic is selected as topic **X** and reviewed.
14 RAG is applied to topic **X** using dataset **D**.
15 Topic **X** and Dataset **D** are input into the **RAG** model.
16 The content of topic **X** is refined and expanded upon.
end

Fig. 3. Algorithm for BERT modelling and RAG integration

A. Data Collection and Preprocessing

Keywords selected by experts guide the data collection from publications and patents, leveraging sources as described in Section IV. This strategic data retrieval, followed by preprocessing, as described in Section IV, streamlines the dataset for effective topic modeling.

B. BERT Topic Modeling

The methodology leverages SBERT within BERT's transformer architecture for document embedding, encapsulating semantic features essential for clustering. The process involves:

1) Dimensionality Reduction: UMAP is employed to reduce the dimensionality of embeddings, maintaining the integrity of data structures critical for clustering.
2) Clustering: HDBSCAN clusters the embeddings into semantically coherent groups, each representing distinct topics.
3) Topic Representation: The c-TF-IDF algorithm refines topic essence by highlighting term distinctiveness within clusters.

$$\text{c-TF-IDF}_{t,d} = \frac{\text{TF}_{t,d}}{\text{Sum of TFs in document } d} \times \log\left(\frac{N}{\text{DF}_t}\right) \tag{1}$$

where $\text{TF}_{t,d}$ is the term frequency of term t in document d, N is the total number of documents, and DF_t is the document frequency of term t.

C. Topic Selection and Analysis

Experts review the keywords associated with each topic, selecting the most representative and relevant terms, ensuring the topics' applicability and depth.

D. Retrieval-Augmented Generation (RAG)

Upon identifying the topics, the Retrieval-Augmented Generation (RAG) technique, combined with LLaMA2, is employed to generate detailed summaries, applications about each expert reviewed topic from the same dataset employed in topic modeling. This method enriches the comprehension of each topic, yielding profound insights.

E. Benchmarking with Hype Cycle Analysis

Benchmarking the predictive capabilities of our trend forecasting methodology against the Gartner Hype Cycle is essential for validating its effectiveness in identifying emergent trends. The Hype Cycle, a well-regarded model displaying the adoption and maturation of technologies, serves as the benchmark. In this analysis, the capability of the model to anticipate technological trends is retrospectively evaluated by examining historical data extending up to five years prior to the inaugural inclusion of specific technologies within the Hype Cycle. This evaluation centers on three distinct technologies: Cryptocurrencies, first recognized in the Hype Cycle in 2016; Reinforcement Learning, making its entry in 2023; and Quantum Machine Learning, introduced in 2021. Such a selection spans diverse technological arenas, facilitating a rigorous and encompassing validation endeavor. The objective is to ascertain whether the model exhibits the proficiency to identify early signs of technological emergence as outlined in the Hype Cycle.

VI. Results

An empirical evaluation of the proposed AI-driven trend forecasting methodology has been meticulously conducted, focusing on three technologies: Cryptocurrencies, Quantum Machine Learning (Quantum ML), and Reinforcement Learning, each charting distinct paths through the Gartner Hype Cycle. Keywords for querying the Dimensions and arXiv databases, identified by domain experts for these technologies, targeted the titles and abstracts of publications and patents. The data thus acquired were processed using a BERT-based topic modeling pipeline.

The retrieval of documents pertinent to the specified technologies was facilitated by keywords, as delineated in Table I, provided by experts.

TABLE I
KEYWORDS USED FOR TECHNOLOGIES

Technologies in Gartner Hype Cycle	Expert provided Keywords
Cryptocurrencies	alternative currency, digital currency, alternative payment, digital payment, digital money, digital cash, electronic currency, electronic money, internet cash
Quantum ML	quantum learning, quantum insight, quantum predict, smart quantum, quantum optimization, qubit learning, quantum learner, quantum predictions
Reinforcement Learning	adaptive algorithms, policy optimization,sequential decision making, feedback loops, decision making, reward optimization

TABLE II
DATASET DESCRIPTION

Technologies in Gartner Hype Cycle	Data collection Interval	Number of Publications	Number of Patents
Cryptocurrencies	2011-2015	2078	2402
Quantum ML	2017-2020	1575	200
Reinforcement Learning	2018-2022	4079	110

Following the collection of the dataset, a series of preprocessing steps, as detailed in Section IV, were applied to refine the data prior to its analysis through BERT topic modeling. This procedure transformed the textual content into semantically dense embeddings as discussed in section V.

A. Quantitative Analysis

This section provides a quantitative assessment of the topics extracted from the BERT-based topic modeling, offering insights into the contextual relevance and evolution of terms associated with the selected technologies over time.

1) Cryptocurrencies: The qualitative analysis for Cryptocurrencies, as depicted in Table III, reveals a topic that has been successfully identified in publication data. However on the patent data the algorithm was not able to detect the topic directly but indirect words correlating with the topic were identified. The topic modeling process has adeptly highlighted key terms that are intrinsic to Cryptocurrencies as described in table III. The number of documents discussing Cryptocurrencies-related topics has seen a significant rise over the past five years as shown in figure 4.

TABLE III
TOPIC MODELLING RESULTS FOR CRYPTOCURRENCIES

Technologies in Gartner Hype Cycle	**First-time appearance in Gartner Hype Cycle**	**Topic modelling on Patent data**	**Topic modelling on Publication data**
Cryptocurrencies	2016	electronic money + electronic wallet + transaction + virtual currency + digital currency + ledger + electronic + secure + cloud	blockchain + *cryptocurrencies* + cryptocurrency + bitcoin + currency bitcoin + digital currency + virtual currency + ledger

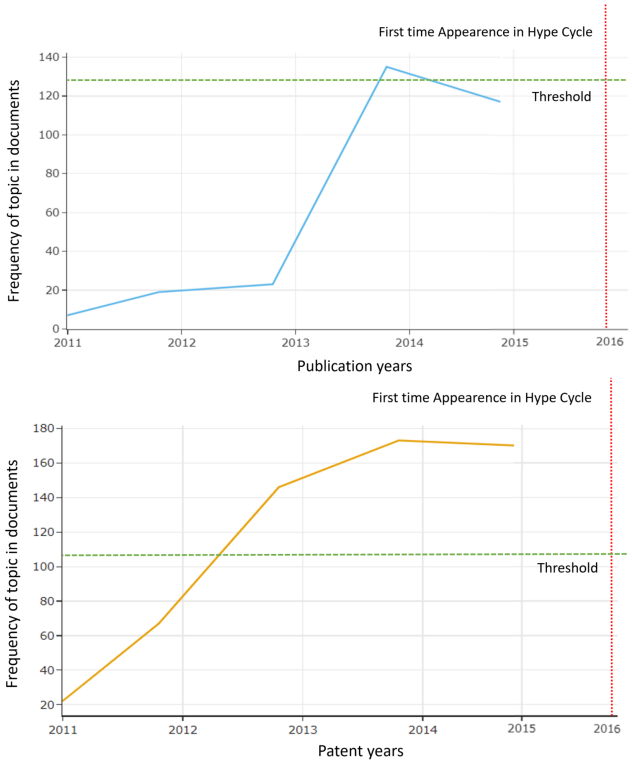

Fig. 4. Cryptocurrencies topic growth over time

2) Quantum ML: In the case of Quantum Machine Learning (Quantum ML), Table IV illustrates Quantum machine learning topic gleaned from the publication and patent data through topic modeling technique. A similar upward trend is observable in the number of documents pertaining to Quantum ML as shown in figure 5.

TABLE IV
TOPIC MODELLING RESULTS FOR QUANTUM ML

Technologies in Gartner Hype Cycle	**First-time appearance in Gartner Hype Cycle**	**Topic modelling on Patent data**	**Topic modelling on Publication data**
Quantum ML	2021	*quantum neural network* + intermediate quantum neural + quantum neural + quantum twin neural + neural network layer + intermediate quantum + multiple qubits	*quantum machine learn* + quantum neural network + quantum neural + quantum compute + quantum machine quantum neuron + quantum speedup

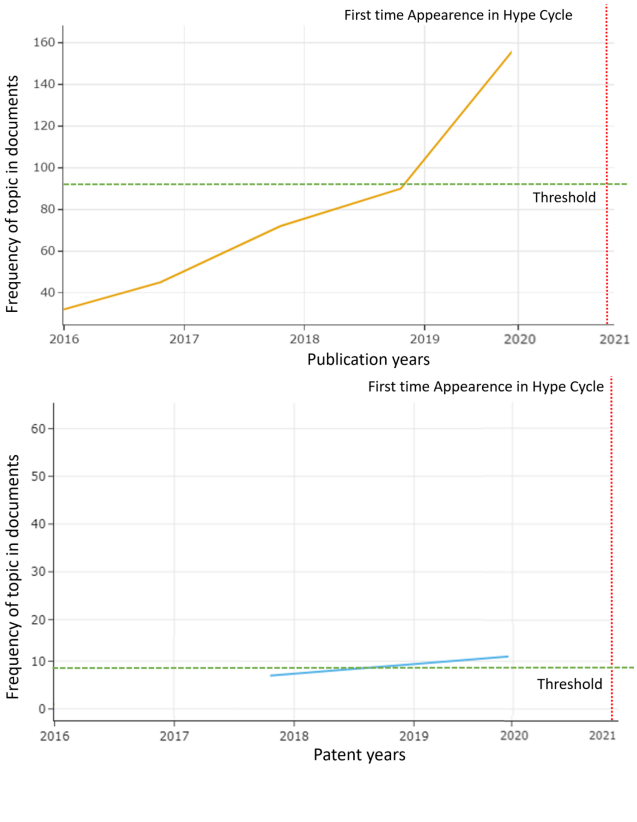

Fig. 5. Quantum ML topic growth over time

3) Reinforcement Learning: For Reinforcement Learning, the topic was also positively identified in analysis, as presented

in Table V. Also, the growth of topic which is increasing over the time for both patents and publication is described in figure 6.

TABLE V
TOPIC MODELLING RESULTS FOR REINFORCEMENT LEARNING

Technologies in Gartner Hype Cycle	First-time appearance in Gartner Hype Cycle	Topic modelling on Patent data	Topic modelling on Publication data
Reinforcement Learning	2023	*reinforcement learn* + reinforcement + model + optimization + robot + qlearning algorithm + strategy + network + training	*reinforcement learning* + deep reinforcement learn + rl + proximal policy optimization + policy optimization + policy gradient + learn algorithm

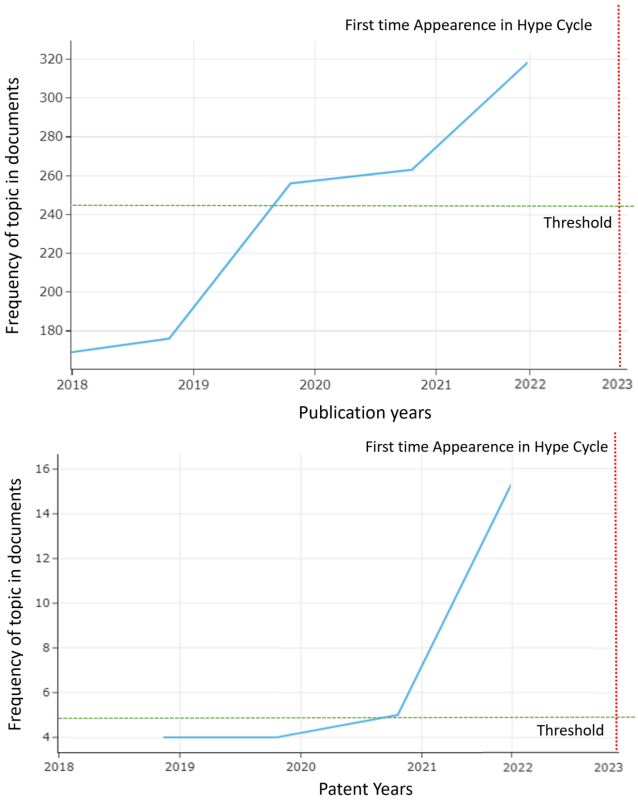

Fig. 6. Reinforcement Learning topic growth over time

The qualitative analysis across all three technologies has been corroborated by the quantitative findings, reinforcing the robustness of the trend forecasting approach that has been implemented. Thresholds for each topic were systematically determined by calculating the percentage of total documents that discuss each of the three technologies, as detailed in Table VI. An average of these percentages has been computed, providing a generalized indicator that is applicable to both publications and patents. These thresholds have been identified as useful tools to distinguish between mere topics of discussion and significant trends. It is suggested that these established thresholds should be applied in ongoing studies to assess their efficacy in trend identification across various datasets. The use of these thresholds could facilitate a more precise distinction between emerging topics and dominant trends, potentially enhancing the predictive capabilities of the trend forecasting methodology.

TABLE VI
THRESHOLD OF TOPIC TO BECOME TREND

	Threshold for Publications	Threshold for Patents
Average	6.1%	4.5%

B. Alignment with Gartner Hype Cycle

The outcome of topic modeling with respect to Gartner Hype cycle is summarized in Table VII, which aligns the results of analysis with the respective technologies' first-time appearances in the Gartner Hype Cycle.

TABLE VII
RESULT ALIGNMENT WITH GARTNER HYPE CYCLE

Technologies in Gartner Hype Cycle	First-time appearance in Hype Cycle	Publication Detection in Gartner Hype cycle based on Threshold Year	Patent Detection in Gartner Hype cycle based on Threshold Year
Cryptocurrencies	2016	Yes (2014)	No
Quantum ML	2021	Yes (2019)	Yes (2019)
Reinforcement Learning	2023	Yes (2020)	Yes (2020)

C. Integration of RAG

Retrieval-Augmented Generation (RAG) was implemented as an extension to topic modeling, which traditionally yields only keywords pertinent to each topic. RAG, utilizing LLaMA2 as the underlying language model, was instrumental in generating comprehensive description that delve beyond keywords to elucidate topics. This step was crucial for articulating detailed description and outlining potential applications. The same dataset that facilitated topic modeling also served as the basis for RAG's application, ensuring consistency in data analysis. The effectiveness of RAG was exemplified through the case study example of cryptocurrencies, demonstrating its capacity to enhance topic understanding significantly.

The results from the RAG is described in the below two paragraphs covering the description of the trend cryptocurrencies and application of cryptocurrencies based on the dataset.

Description of Cryptocurrencies: *Cryptocurrencies are a new type of money that uses advanced technology to solve longstanding problems with digital cash. They are neither commodity money nor fiat money - they are a new, experimental kind of money. The cryptocurrency experiment may or may not ultimately succeed, but it offers a new mix of technical and monetary characteristics that raise different economic questions than other kinds of currency.*

Application of Cryptocurrencies:

1) *Online transactions: Cryptocurrencies like Bitcoin enable secure and fast online transactions without intermediaries like banks or payment processors.*
2) *Cross-border payments: Cryptocurrencies can facilitate cross-border payments without the need for conversion or high fees associated with traditional currencies.*
3) *Decentralized finance (DeFi): Cryptocurrencies like Ethereum support DeFi, which includes lending, borrowing, and trading decentralized assets.*
4) *Smart contracts: Cryptocurrencies like Ethereum allow for the execution of smart contracts, automating various processes such as supply chain management and insurance claims.*
5) *Privacy and anonymity: Some cryptocurrencies, like Monero and Zcash, offer enhanced privacy features, making them ideal for individuals who value their financial data security.*
6) *Investment: Cryptocurrencies like Bitcoin and altcoins can be used as investment vehicles, providing potential returns through price appreciation.*
7) *Non-fungible tokens (NFTs): Certain cryptocurrencies, such as Ethereum, support NFTs, allowing artists and content creators to monetize unique digital assets.*
8) *Gaming: Cryptocurrencies like Roblox and Decentraland have found use in gaming, enabling players to purchase virtual items and experiences using cryptocurrencies.*
9) *Charitable giving: Cryptocurrencies like BitGive enable charitable donations and fundraising campaigns, leveraging blockchain technology for transparency and accountability.*

VII. Conclusion

The study introduces an AI-based methodology that synergistically combines BERT and Generative AI for effective trend forecasting in technology, validated by its alignment with the Gartner Hype Cycle for technologies like Cryptocurrencies, Quantum Machine Learning, and Reinforcement Learning.

VIII. Outlook

The results obtained from this study highlight the methodology's effectiveness in forecasting technological advancements. A substantial correlation with the Gartner Hype Cycle confirms the model's validity as a robust tool for early trend detection. Future research will focus on expanding the scope of analysis through the utilization of an enlarged dataset and a more detailed examination of each topic's intricacies, aiming to enhance the predictive accuracy and broaden the methodology's applicability across various sectors. Additionally, subsequent efforts will explore assessing the technology readiness levels for each identified topic, further refining the tool's utility in strategic technology planning and implementation.

References

[1] National Research Council, Persistent Forecasting of Disruptive Technologies, Washington, DC: The National Academies Press, 2010. [Online]. Available: https://doi.org/10.17226/12557 .

[2] Q. Wang, "A Bibliometric Model for Identifying Emerging Research Topics," *Journal of the Association for Information Science and Technology*, vol. 69, no. 2, pp. 290–304, 2017.

[3] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Batam, Indonesia, 2019, pp. 386–390, doi: 10.1109/ICECOS47637.2019.8984523.

[4] D. M. Blei, A. Ng, M. Jordan, and J. Lafferty, "Latent Dirichlet allocation," Journal of Machine Learning Research (JMLR), vol. 3, pp. 993–1022, 2003.

[5] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles, "Detecting Topic Evolution in Scientific Literature: How Can Citations Help?" in Proc. of the 18th International Conference on Information and Knowledge Management (CIKM'09), pp. 957–966, 2009.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019, arXiv preprint arXiv:1810.04805.

[7] M. V. Koroteev, "BERT: A Review of Applications in Natural Language Processing and Understanding," 2021, arXiv preprint arXiv:2103.11943.

[8] E. Atagün, B. Hartoka, and A. Albayrak, "Topic Modeling Using LDA and BERT Techniques: Teknofest Example," in Proc. 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 660–664, doi: 10.1109/UBMK52708.2021.9558988.

[9] P. Finardi, L. Avila, R. Castaldoni, P. Gengo, C. Larcher, M. Piau, P. Costa, and V. Caridá, "The Chronicles of RAG: The Retriever, the Chunk and the Generator," 2024, arXiv preprint arXiv:2401.07883.

[10] N. Seth, "Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn," Analytics Vidhya, 26 Aug. 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/.

[11] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv preprint arXiv:2203.05794, 2022.