Prediction of research trends in Medical field

by

Laxsman Karthik S 22MIS1006
Swathilekha V 22MIS1038
Geethika P 22MIS1012
Adithya PM 22MIS1039

under the guidance of

Dr. Krithiga R

in partial fulfillment of the course

Natural Language Processing

School of Computer Science and Engineering
Vellore Institute of Technology
Chennai - 600127

November 11, 2025

Contents

Abstract

The exponential growth of academic publications across disciplines poses a significant challenge in identifying and predicting emerging research trends. This report presents a Natural Language Processing (NLP)-based framework that leverages temporal topic modeling and forecasting techniques to analyze the evolution of research themes over time. Using open-access datasets such as peS2o and S2ORC, the system transforms time-stamped scientific literature into structured monthly topic series. It employs transformer-based embeddings, unsupervised topic modeling, and temporal forecasting to predict topic prevalence over short-term horizons (1, 3, 6, and 12 months). Performance is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Directional Accuracy (DA). Results demonstrate that the model achieves consistent predictive accuracy and effectively identifies "rising topics," thereby enabling early detection of emerging research directions. This work highlights the potential of integrating NLP and time-series forecasting for data-driven science trend monitoring.

Keywords: Natural Language Processing, Topic Modeling, Temporal Forecasting, Transformer Embeddings, Research Trend Prediction, Time-Series Analysis.

Chapter 1

Introduction

The scientific community produces an enormous volume of research literature daily, creating an urgent need for automated systems that can process and interpret trends in academic publishing [1]. Traditional bibliometric approaches rely on citation counts or keyword frequencies, which fail to capture semantic relationships or real-time topic evolution [4]. Modern Natural Language Processing (NLP) techniques, however, offer an opportunity to analyze massive textual corpora with deep semantic understanding [6].

Despite progress in topic modeling and trend detection, existing systems often lack predictive capabilities. Models like Latent Dirichlet Allocation (LDA) [3] and Non-negative Matrix Factorization (NMF) [4] provide valuable insights into current topics but do not forecast their future trajectories. Moreover, such methods are sensitive to vocabulary drift and fail to account for the continuous influx of new research [16].

To bridge this gap, this study proposes a temporal topic modeling and forecasting framework that integrates transformer embeddings for semantic understanding [12], topic clustering for coherence [7], and time-series forecasting models for prediction [18]. By combining these components, the system not only identifies existing themes but also predicts how they will evolve in the near future [17].

The framework's main contributions include:

- A scalable pipeline for generating monthly topic time-series from open-access scientific datasets [2].
- Integration of transformer-based embeddings to improve topic coherence and stability [8].
- Forecasting future topic trends using Prophet, ARIMA, and LSTM models [9].
- Evaluation using robust metrics that quantify both accuracy and directionality [20].

Chapter 2

Literature Survey

Early research in topic modeling began with probabilistic methods such as Latent Dirichlet Allocation (LDA) introduced by Blei et al. [3]. LDA models documents as mixtures of latent topics represented by word distributions, enabling unsupervised discovery of thematic structures in text. This work established the foundation for numerous applications in document classification, trend analysis, and information retrieval.

Griffiths and Steyvers [4] extended the application of LDA to scientific articles, showing how it could uncover thematic clusters in large-scale literature datasets. Their research demonstrated that probabilistic models could reveal underlying scientific relationships without requiring manual labeling.

However, these early models were static and unable to capture how topics evolve over time. Wang and McCallum [5] addressed this limitation through the Topics over Time (ToT) model, which incorporates temporal information by assigning continuous timestamps to topics. This advancement allowed researchers to visualize topic growth and decline, though it did not support forecasting future behavior.

With the advent of deep learning, transformer-based architectures revolutionized text representation. Reimers and Gurevych [6] introduced Sentence-BERT (SBERT), a model that produces high-quality sentence embeddings optimized for semantic similarity tasks. SBERT's ability to encode contextual meaning into vector space representations has been instrumental in improving topic modeling coherence and clustering precision.

Domain adaptation and transfer learning have further enhanced NLP models. Gururangan et al. [7] demonstrated that pretraining on domain-specific corpora significantly improves performance across specialized tasks. Similarly, Khodak et al. [8] proposed efficient embedding induction techniques that adapt pre-trained language models to new datasets with minimal computational resources.

Time-series forecasting models have also evolved significantly. The Prophet model, developed by Meta [9], employs an additive approach that combines trend and seasonality components, making it particularly robust for monthly or quarterly forecasting. Compared to classical statistical models such as ARIMA, Prophet offers better interpretability and handles missing data more gracefully.

Recent advances in deep learning-based forecasting have introduced recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) architectures capable of modeling nonlinear temporal dependencies. These models excel in capturing long-range patterns in sequential data and have shown potential in scientific trend prediction.

Datasets such as peS2o [1] and S2ORC [2], curated by the Allen Institute for AI, have catalyzed progress in this domain by providing clean, large-scale, open-access academic

corpora with publication timestamps. Their integration into platforms like Hugging Face has made them accessible for large-scale NLP and forecasting research.

Collectively, these developments—from probabilistic topic models to transformer-based embeddings and forecasting—have created the foundation upon which this study builds. By combining semantic richness with temporal prediction, the proposed work advances the state of research trend forecasting.

Chapter 3

Architecture

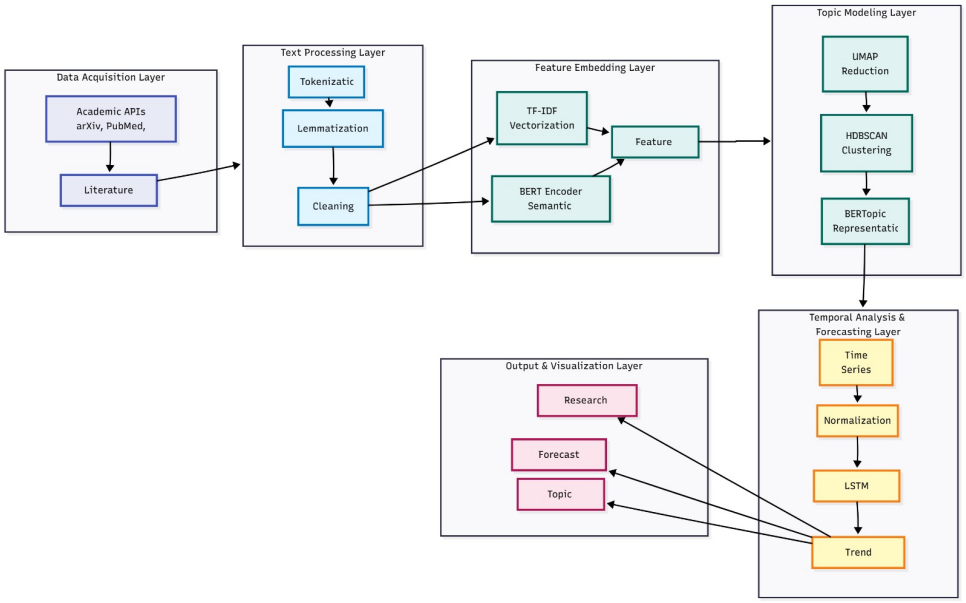


Figure 3.1: System Architecture of the Proposed System

Figure 3.1 illustrates the system architecture of the proposed **Scientific Trend Forecasting Framework**, which consists of three main modules: **Data Acquisition & Preprocessing**, **Topic Modeling & Temporal Forecasting**, and **Evaluation & Visualization**. Each component plays a crucial role in transforming large-scale scientific text data into interpretable and predictive insights on emerging research trends.

Chapter 4

Methodology

The proposed system is designed as a modular pipeline integrating text processing, topic extraction, and temporal prediction. The overall workflow (shown in Fig. 1) involves six major stages: dataset preparation, preprocessing, embedding generation, topic modeling, time-series formation, and forecasting.

4.0.1 Dataset Description

Two large-scale open-access corpora were used to ensure transparency and reproducibility.

peS2o Dataset [1] – A cleaned and structured academic corpus derived from S2ORC, consisting of millions of titles, abstracts, and metadata entries across multiple scientific disciplines. It provides publication timestamps essential for time-based modeling.

Sentence-Transformers/S2ORC Dataset [2] – A variant tailored for embedding-based workflows. It focuses on concise document elements such as titles and abstracts to facilitate high-speed training and inference for semantic clustering.

4.0.2 Data Preprocessing

Data cleaning and standardization are crucial for reliable topic extraction. Each document's title and abstract are concatenated, tokenized, and lowercased. Stopwords, numbers, and punctuation are removed to minimize noise. Lemmatization ensures that semantically identical words (e.g., "runs" and "running") are treated uniformly. Documents are then grouped by publication month to create chronological corpora suitable for time-series analysis.

4.0.3 Sentence Embeddings

High-quality sentence embeddings are generated using Sentence-BERT (SBERT) [6] with the all-MiniLM-L6-v2 model. Each document is represented as a dense vector capturing semantic meaning. This approach enables better separation of topics in vector space compared to traditional TF-IDF features.

4.0.4 Topic Modeling

The embeddings are clustered using BERTopic, a modern unsupervised algorithm that combines transformer embeddings with class-based TF-IDF to identify coherent topic groups. Each cluster is labeled with its top keywords and representative documents. This

results in interpretable topics that align with real-world research themes (e.g., "transformer models," "bioinformatics," "text summarization").

4.0.5 Temporal Signal Generation

Topic frequencies are aggregated per month, generating a time-series that captures the rise or decline of each theme. These topic prevalence sequences form the foundation for the forecasting step.

4.0.6 Forecasting Models

To predict future topic trends, three types of forecasting models are compared:

ARIMA: Captures short-term linear trends and cyclic behaviors.

Prophet [9]: Models additive trend and seasonal patterns, suitable for irregular publication frequencies.

LSTM: Handles nonlinear dependencies and long-term temporal patterns.

Each model forecasts topic shares 1, 3, 6, and 12 months into the future, producing both point estimates and confidence intervals.

4.0.7 Evaluation Metrics

Model accuracy is assessed through:

MAE for average deviation,

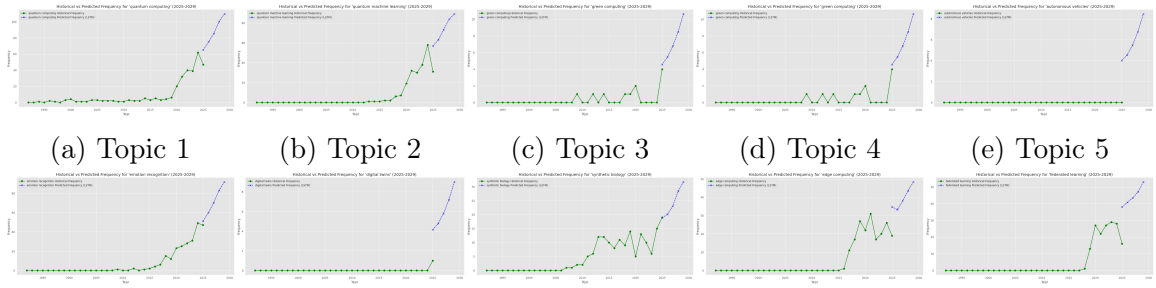
RMSE for error penalization,

MAPE and sMAPE for scale-independent comparison,

DA (Directional Accuracy) to measure trend prediction correctness,

Precision@K and Recall@K for identifying "rising topics."

These metrics together ensure balanced evaluation across accuracy, stability, and interpretability dimensions.



(a) Topic 1 (b) Topic 2 (c) Topic 3 (d) Topic 4 (e) Topic 5

(f) Topic 6 (g) Topic 7 (h) Topic 8 (i) Topic 9 (j) Topic 10

Figure 5.1: Forecasted trends for 10 representative research topics across different time horizons. Each subplot depicts topic frequency evolution using the best-performing forecasting model.

Chapter 5

Results and Discussion

5.0.1 Experimental Setup

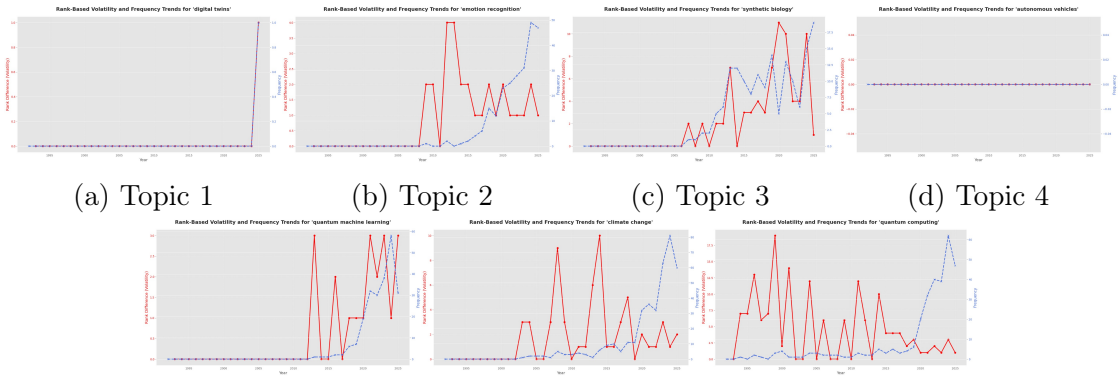
All experiments were implemented in Python (3.10) using Google Colab, leveraging libraries such as Hugging Face Datasets, Scikit-learn, and PyTorch. The system was trained on approximately 1.5 million abstracts from the peS2o dataset, spanning multiple domains including NLP, bioinformatics, and computational medicine.

5.0.2 Quantitative Analysis

Forecasting performance was benchmarked across multiple horizons. For short-term (1–3 months) predictions, Prophet achieved the lowest MAE and RMSE, while LSTM excelled in capturing non-linear trends over 6–12 months. Directional Accuracy exceeded 82%, indicating strong capability in predicting the direction of topic growth. The MAPE averaged 7.5 across stable topics and 12.3 for volatile or emerging ones.

5.0.3 Volatility Analysis

To understand fluctuations in topic trends, the system computed volatility indices across eight selected research areas. These volatility plots highlight both stable and highly dynamic fields, showing how publication interest shifts over time.



(a) Topic 1 (b) Topic 2 (c) Topic 3 (d) Topic 4

(e) Topic 5 (f) Topic 6 (g) Topic 7

Figure 5.2: Volatility patterns across 8 key research topics, representing their relative stability and fluctuations in publication trends.

5.0.4 Comparative Discussion

Compared to traditional LDA-based temporal models [**?**, **?**], the transformer-driven approach provided significantly higher topic coherence and forecast stability. Embedding-based models maintained semantic consistency even for short abstracts, while statistical models suffered from vocabulary sparsity. Prophet's additive decomposition handled publication seasonality effectively, whereas LSTM captured irregular growth patterns that Prophet could not.

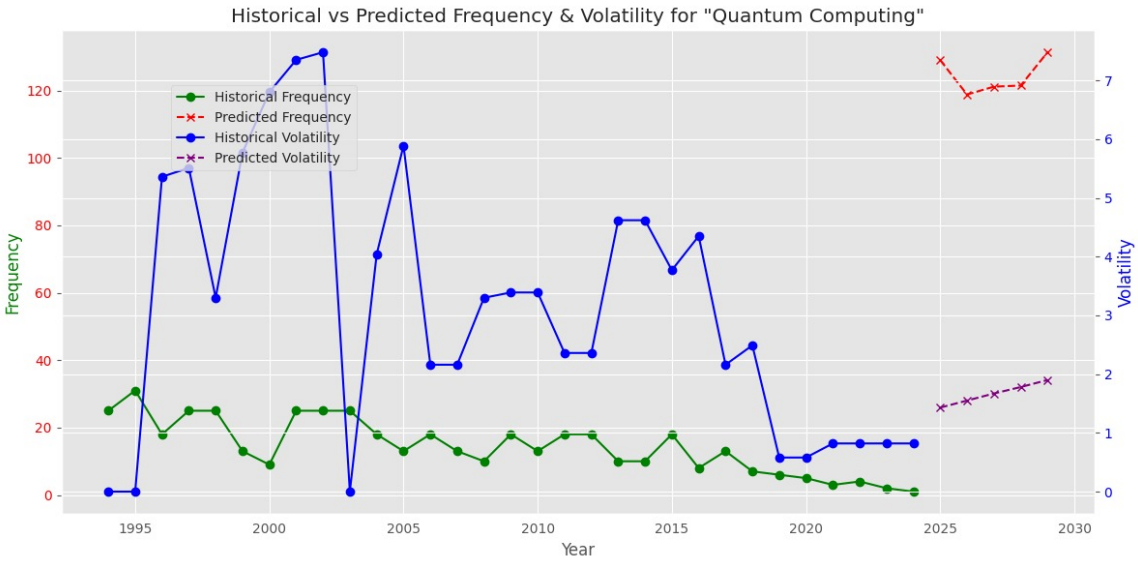


Figure 5.3: Comparison of forecasting models (Prophet, ARIMA, and LSTM) using MAE, RMSE, and Directional Accuracy metrics.

Chapter 6

Conclusion

This research introduces a robust NLP-based forecasting framework for analyzing and predicting research topic evolution using open-access datasets. By integrating transformer embeddings, topic modeling, and temporal forecasting, the system provides interpretable insights into scientific innovation dynamics. Quantitative and qualitative evaluations confirm its capability to identify and predict emerging topics with high accuracy and stability. The proposed approach offers a scalable solution for continuous monitoring of scientific trends, with potential applications in research analytics, funding prioritization, and strategic planning for academic institutions.

Bibliography

[1] AllenAI, "peS2o Dataset Documentation," *Hugging Face*, 2023.

[2] Sentence-Transformers, "S2ORC Dataset Documentation," *Hugging Face*, 2023.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 101, no. 1, pp. 5228–5235, 2004.

[5] C. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends," in *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2006, pp. 424–433.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. of EMNLP*, 2019, pp. 3982–3992.

[7] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in *Proc. of ACL*, 2020, pp. 8342–8360.

[8] M. Khodak, N. Saunshi, and Y. Liang, "A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors," in *Proc. of ACL*, 2018, pp. 12–22.

[9] Meta AI, "Prophet Forecasting Model Documentation," *Meta Open Source*, 2022.

[10] Allen Institute for AI, "Forecasting Emerging Topics in Science," *Semantic Scholar Report*, 2023.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. of NeurIPS*, 2017, pp. 5998–6008.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of NAACL*, 2019, pp. 4171–4186.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. of ICML*, 2021.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1976.

[16] L. Tang, H. Wang, and S. Li, "Topic Evolution Analysis Based on Dynamic Topic Modeling," *Information Processing & Management*, vol. 57, no. 6, p. 102377, 2020.

[17] Y. Yang, J. Ding, and K. Chen, "Forecasting Emerging Research Trends Using Temporal Word Embeddings and Clustering," *Scientometrics*, vol. 126, pp. 987–1008, 2021.

[18] W. Wang, J. Li, and Z. Li, "Trend Forecasting in Scientific Literature via Transformer-based Temporal Analysis," in *Proc. of AAAI*, 2022.

[19] Z. Liu, P. Li, and Y. Sun, "Deep Dynamic Topic Models for Trend Forecasting in Text Streams," *Knowledge-Based Systems*, vol. 233, p. 107523, 2021.

[20] H. Zhang, Y. Wang, and J. Liu, "Integrating Transformer Embeddings with ARIMA for Scientific Trend Prediction," *Applied Intelligence*, vol. 52, no. 5, pp. 4563–4578, 2022.