Asian J Biol Life Sci., 2026; 15(1):143-149.
https://ajbls.com

Research Article

AMR Curator-An Interactive Platform for Analysing AMR Literature Using NLP

Umesh Rani, Naval Singh, Deepshikha Kaushik*

Centre for Bioinformatics, Maharshi Dayanand University, Rohtak, Haryana, INDIA.

ABSTRACT

Background: The rising issue of Antimicrobial Resistance (AMR) poses a major threat to global health. **Objectives:** This study introduces a Streamlit application that helps analyze and visualize trends in AMR literature using natural language processing and interactive data visualization. **Materials and Methods:** We curated a PubMed dataset comprising abstracts published from 2015 to 2025 and used advanced techniques such as topic modeling, named entity recognition, and keyword analysis. We processed 154608 AMR-related abstracts, identified five main research themes by topic modelling, categorized biomedically relevant entities into key pathogens, antimicrobial agents, and geographical entities using NER. **Results:** Our user-friendly interface represents literature-based AMR analysis tools that combines multiple analytical methods into a single platform. The platform is freely accessible at [https://d-k-02-amr-dashboard-app-qdprm0.streamlit.app/]. **Conclusion:** This tool offers researchers, epidemiologists, and policymakers an interactive resource for tracking research trends and making informed decisions in antimicrobial stewardship.

Keywords: Antimicrobial resistance, Interactive data visualization, Named entity recognition, Natural language processing, Topic modeling.

Correspondence:
Dr. Deepshikha Kaushik
Centre for Bioinformatics, Maharshi Dayanand University, Rohtak-124001, Haryana, INDIA.
Email: drdeepshikha.rp.cbinf@mdurohtak.ac.in

Received: 24-12-2025;
Revised: 09-02-2026;
Accepted: 18-04-2026.

INTRODUCTION

Antimicrobial Resistance (AMR) stands as one of the most urgent healthcare challenges worldwide today. It happens when harmful microorganisms develop resistance to antimicrobial drugs, threatening years of medical progress (World Health Organization, 2020). Contributing factors include misuse of antimicrobials, environmental impacts, and the natural adaptability of pathogens, all of which have triggered the quick emergence of resistance across different microbial species.

The severity of the AMR crisis is underscored by projections suggesting that without effective intervention strategies, antimicrobial-resistant infections could result in 10 million deaths annually by 2050 (O'Neill, 2016). This alarming projection has galvanized the global health community to intensify research efforts, resulting in an exponential increase in AMR-related publications across scientific literature.

While the increase in AMR research is a positive step in tackling this global issue, it also creates major challenges for researchers and policymakers trying to synthesize and derive useful insights from the extensive body of literature. The large volume of published studies can make it hard to spot emerging patterns, track how resistance mechanisms develop, and understand the geographic distribution of AMR threats.

Current resources such as the Comprehensive Antibiotic Resistance Database (CARD) and the National Database of Antibiotic Resistant Organisms (NDARO) provide essential sequence and phenotypic data respectively (Alcock *et al.*, 2020; Zankari *et al.*, 2012). However, these databases have limitations in their ability to facilitate literature exploration through interactive visualizations, and they lack advanced natural language processing features that could enhance understanding of research trends and patterns.

To address these limitations, this study introduces a new web-based platform (Figure 1) designed specifically for microbiologists, epidemiologists, and data scientists to monitor, analyze, and visualize research trends in antimicrobial resistance. The platform utilizes advanced natural language processing techniques and interactive visualization tools to turn raw literature data into actionable insights for the scientific community.

MATERIALS AND METHODS

Data Collection

The foundation of this research platform was built upon a comprehensive collection of scientific abstracts retrieved from PubMed, the premier biomedical literature database. Data

OPEN ACCESS

Manuscript TECHNOMEDIA

DOI: 10.5530/ajbls.20260077

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

collection was conducted using a Python-based script that interfaced with the National Centre for Biotechnology Information (NCBI) Entrez API. To overcome the inherent limitations of the NCBI record retrieval system, the data collection process was strategically segmented into 4-month intervals for a period of 10 years, 2015 to 2025.

This temporal segmentation approach was implemented to circumvent NCBI's record limit restrictions while ensuring comprehensive coverage of the target period. Search queries were systematically executed and results were saved as individual CSV files to facilitate batch processing and maintain data integrity throughout the collection process. Following the completion of data collection across all temporal segments, the individual CSV files were merged into a unified dataset for subsequent analysis.

Data Preprocessing and Enrichment

The raw dataset underwent comprehensive preprocessing and enrichment procedures implemented through custom Python scripts. Tabular data loading, cleaning, and preprocessing were performed in Python using the Pandas library (Mckinney, 2010). The preprocessing pipeline included several critical steps designed to enhance data quality and extract meaningful insights from the textual content.

Missing values in the dataset were systematically identified and corrected using appropriate cleaning methods to ensure data completeness and analysis reliability. The text preprocessing involved extracting top keywords with Term Frequency-Inverse Document Frequency (TF-IDF) analysis, which highlighted the most discriminative terms within the corpus.

Topic modeling was performed using Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003), implemented via the Gensim library (Rehurek and Sojka, 2010), revealing underlying thematic structures in the literature. Prior to modeling, each abstract underwent natural language processing using spaCy pipeline, that was comprised of tokenization, lowercasing, removal of punctuation, filtering of non-alphabetic tokens and elimination of common stopwords (Honnibal, 2017). The clean tokens were then used to construct a document- term matrix using the Gensim library. This unsupervised machine learning approach enabled the identification of distinct research themes and their relative significance across the dataset.

Named Entity Recognition (NER) was performed using the spaCy natural language processing library (Honnibal, 2017), which facilitated the automatic extraction of relevant biomedical entities, including pathogen names, antimicrobial agents, and geographical locations. Additionally, word clouds were generated for Medical Subject Headings (MeSH) terms to provide visual representations of the most frequently occurring standardized medical terminology, using matplotlib (Hunter, 2007).

Data Storage and Access Infrastructure

The enriched dataset was stored using MongoDB, a NoSQL database system that provides the flexibility and scalability required for handling large volumes of semi-structured textual data. The implementation of MongoDB was achieved through custom Python scripts that established the necessary database connections and data insertion protocols.

This database architecture was chosen to support dynamic querying and quick data retrieval within the dashboard. The NoSQL model offers the flexibility to handle various data structures in bibliographic records while ensuring efficient performance for complex queries.

Interface Design and User Experience

The user interface was developed using Streamlit, a Python-based framework that enables the rapid development of interactive web applications for data science and machine learning projects (Streamlit Inc, 2024). The dashboard features a clean, intuitive design as organized into four primary navigation tabs, each serving distinct analytical purposes. Figure 1 shows a general search result of AMR Curator with *E. coli* as microorganism and 2025 as search year.

The Home tab offers users an overview of AMR-related content and features search options for publications by year and pathogen. The Statistics tab provides detailed visualizations of keyword frequency distributions, topic modeling results, and MeSH term word clouds. The About tab includes comprehensive information about the tool's purpose, features, and usage instructions to support user engagement and effective use. The Contact tab enables users to reach out with feedback and queries.

RESULTS

Dataset Characteristics and Overview

The final corpus includes a large collection of abstracts from a wide variety of journals in biomedical and clinical research. Each record provides detailed metadata such as publication year, journal details, MeSH terms, abstract content, and author affiliations. This comprehensive metadata setup facilitates diverse analysis methods and helps answer different research questions about AMR trends and patterns.

Keyword Analysis and Trend Identification

The TF-IDF analysis revealed several key terms that dominate the AMR literature landscape. The most prominent keywords identified include "resistance," "*E. coli*," "cephalosporin," and "multi-drug," which align closely with well-documented global AMR hotspots and concerns. These findings demonstrate the platform's capability to identify and highlight the most significant themes within the literature, providing researchers

with immediate insights into the primary focus areas of AMR research.

The prevalence of these specific terms reflects current global priorities in AMR research, with particular emphasis on gram-negative bacterial pathogens and broad-spectrum antimicrobial agents. The identification of "multi-drug" as a top keyword underscores the growing concern regarding multidrug-resistant organisms and their impact on clinical outcomes (Figure 2).

Generation of LDA topic model

A 5-topic LDA model was generated, which valuably provided interpretable thematic clusters comprising major AMR research areas. Although a formal coherence- based model comparison was not conducted in this iteration, the 5-topic structure yielded semantically coherent themes that were found to be suitable for exploratory literature summarization. These topics, each one representing a major research theme, were used to characterize dominant research directions across the AMR publication scenario.

The most prevalent topic was Bacterial activity and biofilm resistance (*n*=37,420), which indicated a strong emphasis (32.71%) on microbial mechanisms and resistance development at the cellular and community levels.

The second most prominent (24.67%) topic, Genomic studies of resistant isolates (*n*=28,231), highlights the increasing application of genomic and molecular tools to characterize resistance determinants and evolutionary patterns. AMR policy and public health (*n*=22,136), another major topic (19.35%), underscores the role of surveillance, stewardship, and policy-driven interventions in addressing AMR at the population level.

Research related to Clinical and hospital-acquired infections (*n*=16,023) reflects ongoing concerns (14%) regarding resistant pathogens in healthcare settings and infection control practices. Finally, Tuberculosis/HIV drug resistance and treatment (*n*=10,603) formed a distinct topic (9.27%), capturing studies focused on resistance management in high-burden infectious diseases.

Collectively, these five LDA-derived topics demonstrate a comprehensive research landscape encompassing molecular mechanisms, clinical implications, and public health strategies for combating antimicrobial resistance.

The temporal analysis (Figure 3) reveals distinct trends in topic prevalence over the study period, with novel antimicrobial strategies showing the most significant growth trajectory.

Named Entity Recognition Findings

The NER analysis successfully identified (Figure 4) and categorized numerous biomedically relevant entities across three primary categories: drugs, locations, and pathogens, reflecting the thematic emphasis of the Antimicrobial Resistance (AMR) literature.

Among drug entities, carbapenems were the most frequently mentioned drugs (*n*=181), which indicates their important role in discussions of multidrug-resistant infections. The prominence of Colistin (*n*=89) and vancomycin (*n*=36) reflects their importance as last-line or critical therapeutic agents. Mentions of penicillin (*n*=10) and azithromycin (*n*=9) occurred less frequently, which suggests a comparatively smaller focus on these agents in resistance-related contexts.

For geographic entities, Europe (*n*=48) was the most prominent referenced region, followed by China and the United States (each *n*=18). Presence of the United Kingdom (*n*=9) and Africa (*n*=7) indicate broader global coverage, although they appear with uneven regional representation.

Within the pathogen category, *Staphylococcus* (*n*=74) was the most commonly identified genus, followed by *Escherichia* (*n*=63) and *Pseudomonas* (*n*=56), which highlights their significance in AMR research. *Acinetobacter* (*n*=37) and *Klebsiella* (*n*=21) were also frequently observed, underscoring their roles as clinically important resistant pathogens.

Overall, the NER results complement the LDA findings by linking dominant research topics to specific antimicrobial agents, pathogens, and geographic regions, thereby providing a more granular understanding of the AMR research landscape.

ETHICAL STATEMENT

This study involved the analysis of textual data and did not include human participants, animals, or clinical interventions. All data used in this study were obtained from publicly available

Table 1: Comparison of AMR curator with existing platforms.

Feature	CARD	NDARO	AMR Curator
Data Type	Genomic sequences	Phenotypic data	Literature
Search Functionality	Gene/protein based	Organism based	Text /NLP based
NLP Integration	No	No	Yes
Topic Modeling	No	No	Yes
Literature Analysis	No	No	Yes
User Interface	Web-based	Interactive browser	Interactive Web-based

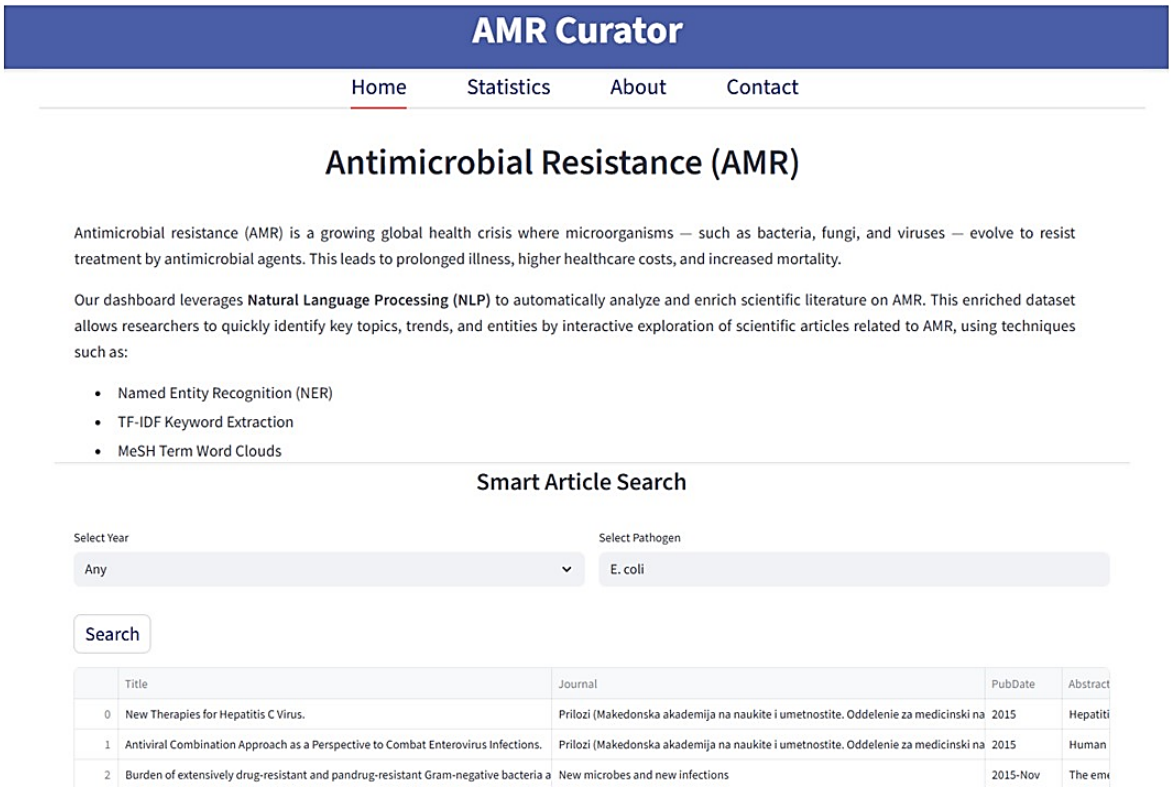

Figure 1: User Interface of AMR Curator with specific search result (*E. coli* as microorganism and 2025 as search year).

sources or anonymized datasets, and no personally identifiable information was collected, accessed, or retained. The study complied with applicable institutional, national, and international ethical guidelines for research using secondary data. As no direct interaction with human subjects occurred and no sensitive personal data were processed, formal ethical approval and informed consent were not required.

DISCUSSION

Platform Innovation and Unique Features

The platform represents a major step forward in literature-based AMR analysis tools by combining multiple analytical methods into a single, user-friendly interface. Unlike static resources, it allows for dynamic and interactive exploration of AMR research through several key innovations: year-wise and pathogen-specific search functions enable users to monitor how research interests have changed over time and to analyze pathogen-specific trends. Real-time keyword and NER insights provide instant access to relevant entities and themes without the need for complicated queries or manual reviews. The MongoDB backend offers fast response times and scalability, supporting future dataset growth and additional features. Its modular architecture facilitates easy expansion, allowing new analytical techniques and visualization methods to be added as needed.

Comparative Analysis with Existing Tools

A systematic comparison of the developed platform with existing AMR-related computational resources reveals distinct functional differences and other capabilities (Table 1). While established databases such as CARD, NDARO provide valuable genomic and phenotypic data, they lack the literature-focused analytical capabilities that characterize the present platform (Pedregosa *et al.*, 2011).The comparative analysis presented above demonstrates that while existing platforms, CARD (Alcock *et al.*, 2020) and NDARO (National Center for Biotechnology Information, NCBI). excel in genomic and phenotypic data management respectively, the developed platform uniquely addresses the literature analysis gap in AMR research infrastructure. The integration of natural language processing capabilities, including named entity recognition and topic modeling, represents a novel approach to AMR knowledge discovery that complements rather than competes with existing genomic resources.

Clinical and Research Implications

The platform's analytical capabilities have significant implications for both clinical practice and research prioritization. The identification of emerging resistance patterns through literature analysis can inform clinical decision-making and antimicrobial stewardship programs. The topic modeling results provide insights into research gaps and areas requiring increased attention from the scientific community.

The geographical analysis capabilities enable researchers and policymakers to understand the global distribution of AMR research activity and identify regions that may require additional research focus or resource allocation. This information is particularly valuable for international health organizations and funding agencies seeking to optimize their AMR-related investments.

Limitations and Future Directions

While the platform represents a significant advancement in AMR literature analysis, several limitations should be acknowledged. The reliance on PubMed abstracts, while comprehensive, may miss significant findings published in journals not indexed by the database. Additionally, the focus on English-language publications may introduce geographical and cultural biases in the analyzed content.

Future development will target key improvements to overcome current limitations and broaden the platform's features. Incorporating BioBERT-driven relation extraction will facilitate more advanced analysis of entity relationships in the literature. Regional filtering options will help users concentrate on particular geographic areas, and temporal trend forecasting will offer predictive insights into future research paths and emerging threats.

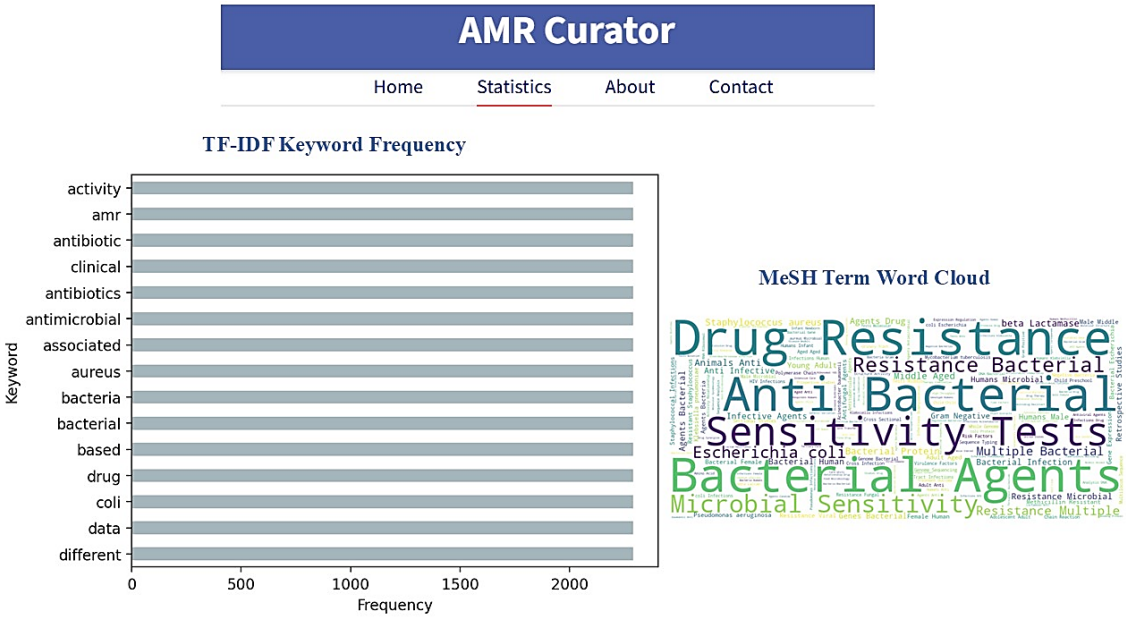

Figure 2: TF-IDF Keyword frequency and MeSH Term Word Cloud generated by AMR Curator.

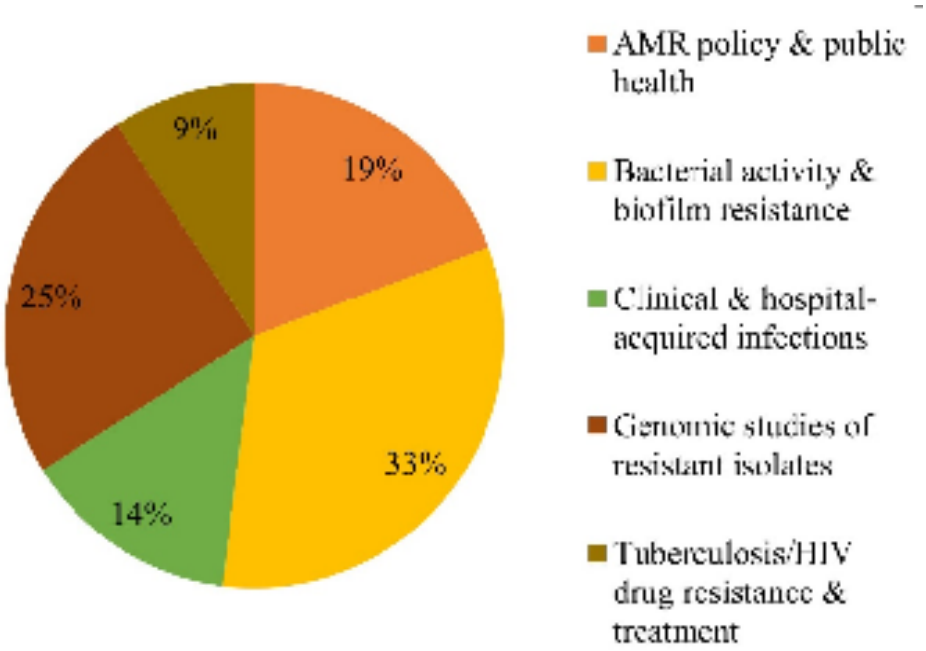

Figure 3: Temporal Analysis: Trends in topic prevalence over study period.

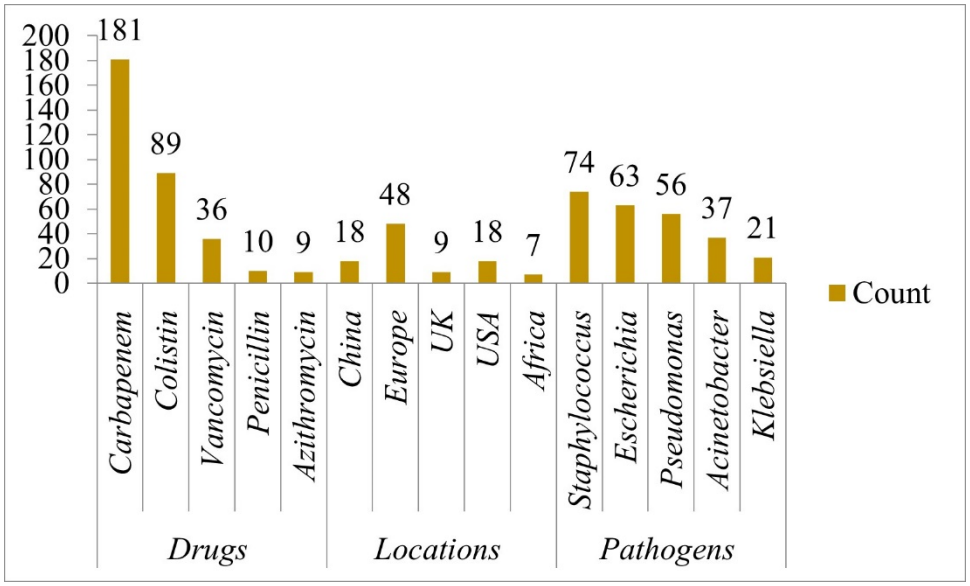

Figure 4: Named Entity Recognition Frequency Graph.

CONCLUSION

This study details the successful creation and deployment of a comprehensive, literature-based analytical platform for antimicrobial resistance research. It effectively merges advanced natural language processing with interactive visualization tools to convert large volumes of scientific literature into actionable insights for researchers, clinicians, and policymakers.

The platform's analytical features have proven useful in identifying key research themes, monitoring temporal trends, and highlighting significant pathogens, antimicrobial agents, and geographical regions in AMR literature. The five main topics identified via LDA modeling offer a broad framework for understanding the current landscape of AMR research. At the same time, the NER results provide detailed insights into the specific entities advancing research in this vital area.

The comparative analysis with existing tools reveals the platform's unique position in the AMR research ecosystem, providing literature-focused capabilities that complement existing genomic and phenotypic resources. The interactive design and scalable architecture ensure that the platform can continue to evolve with the growing body of AMR literature while maintaining optimal performance and user experience.

This work's implications go beyond the research community, reaching clinical practitioners involved in antimicrobial stewardship, public health officials overseeing surveillance and policy, and pharmaceutical researchers working on drug discovery. By offering an accessible, comprehensive analysis of AMR literature trends, the platform facilitates evidence-based decision-making in various areas of AMR research and practice.

Future updates will enhance the platform's analytical features and address current limitations, maintaining its relevance as the AMR research landscape evolves. Adding more NLP techniques, broadening geographical coverage, and incorporating predictive analytics will make the platform a vital tool for understanding the complex, rapidly changing field of antimicrobial resistance.

ACKNOWLEDGEMENT

We would like to thank Centre for Bioinformatics, Maharshi Dayanand University for providing the necessary facilities and resources.

ABBREVIATIONS

AMR: Antimicrobial resistance; **NLP:** Natural language processing; **NER:** Named entity recognition; **TF-IDF:** Term Frequency-Inverse Document Frequency; **LDA:** Latent Dirichlet Allocation; **MeSH:** Medical Subject Headings; **NCBI:** National Centre for Biotechnology Information; **CARD:** Comprehensive Antibiotic Resistance Database; **NDARO:** National Database of Antibiotic-Resistant Organisms; **HIV:** Human Immunodeficiency Virus; **TB:** Tuberculosis.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

FINANCIAL SUPPORT AND SPONSORSHIP

This research received no specific grant from any funding agency.

AUTHOR CONTRIBUTIONS

Conceptualization: Deepshikha Kaushik; Methodology (NLP pipeline design and analysis framework): Umesh Rani; Data curation and preprocessing: Naval Singh; Software and implementation (TF-IDF, LDA, NER models): Deepshikha Kaushik, Umesh Rani; Formal analysis and interpretation of

results: Deepshikha Kaushik, Umesh Rani; Validation and reproducibility checks: Umesh Rani; Visualization and figure preparation: Deepshikha Kaushik; Writing - original draft: Umesh Rani, Naval Singh; Writing - review and editing: Deepshikha Kaushik, Umesh Rani; Supervision and project administration: Deepshikha Kaushik.

All authors have read and approved the final manuscript.

SUMMARY

A dedicated statistical and computational text analysis pipeline was employed for creating a comprehensive, literature-based analytical platform for antimicrobial resistance research. Data preprocessing steps were applied uniformly across the corpus to reduce noise and improve model performance.

TF-IDF was used to quantify the relative importance of terms within the corpus using scikit-learn (Pedregosa *et al.*, 2011). Term frequency captured frequent occurrence of words within a document, while inverse document frequency down-weighted terms that appeared frequently across many documents. The resulting TF-IDF matrix was used for feature representation and exploratory analysis of key terms.

Topic modeling was performed using LDA which identified latent thematic structures within the literature (Blei *et al.*, 2003). The optimal number of topics was determined empirically through evaluation of topic coherence scores and model interpretability. LDA outputs were used to characterize dominant topics and their relative prevalence across documents.

NER was applied to identify and classify entities such as drugs, locations, and pathogens. The obtained results were used to support qualitative interpretation of the topics identified by LDA and to enhance contextual understanding of the text data.

All analyses were conducted using standard natural language processing libraries, and consistent parameter settings were maintained across experiments to ensure reproducibility.

REFERENCES

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., . (2020). CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research, 48(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Honnibal, M. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (No Title). (2017).

Matplotlib, H. J. D. (2007, May 1). A 2D graphics environment. Computing in Science and Engineering, 9(03), 90–95.

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the Python in Science Conference, 445(1), 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

National Center for Biotechnology Information (NCBI). NDARO: National Database of Antibiotic-Resistant Organisms. Bethesda (MD): National Library of Medicine (US). National Institute of Health. https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/

O'Neill, J. Tackling drug-resistant infections globally: Final report and recommendations.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Řehůřek, R., & Sojka, P. Software framework for topic modelling with large corpora.

Streamlit, Inc. (2024). Streamlit [Internet] (version 1.32.0). Streamlit, Inc. https://streamlit.io

World Health Organization. (2020). Antimicrobial resistance. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., & Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. The Journal of Antimicrobial Chemotherapy, 67(11), 2640–2644. https://doi.org/10.1093/jac/dks261

Cite this article: Rani U, Singh N, Kaushik D. AMR Curator-An Interactive Platform for Analysing AMR Literature Using NLP. Asian J Biol Life Sci. 2026;15(1):143-9.