Multilingual Knowledge Graphs: Challenges and Opportunities

Partha Sarathi Mandal*, Sukumar Mandal**

ARTICLE INFO

Article history:
Online first 20 August, 2024

Keywords:
Accessing Data,
Bilingualism,
Data Integration,
Knowledge Graphs,
Multilingualism,
Semantic Web

ABSTRACT

Multilingual Knowledge Graphs (MKGs) have emerged as a crucial component in various natural language processing tasks, enabling efficient representation and utilization of structured knowledge across multiple languages. One can get data, information, and knowledge from various sectors, like libraries, archives, institutional repositories, etc. Variable quality of metadata, multilingualism, and semantic diversity make it a challenge to create a digital library and multilingual search facility. To accept these challenges, there is a need to design a framework to integrate various structured and unstructured data sources for integration, unification, and sharing databases. These are controlled using linked data and semantic web approaches. In future, multilingual knowledge graph overcomes all the linguistic nuances, technical barriers like semantic interoperability, data harmonization etc and enhance cooperation and collaboration throughout the world. Through a comprehensive analysis of the current state-of-the-art techniques and ongoing research efforts, this paper aims to offer insights into the future directions and potential advancements in the field of Multilingual Knowledge Graphs. This paper deals with a multilingual knowledge graph and how to build up a multilingual knowledge graph. It also focuses on the various challenges and opportunities for designing multilingual knowledge graphs.

1. Introduction

With the advancement of information and technology, the design and implementation of multilingual knowledge graph come out as an interesting area of researches. A knowledge graph is a structured representation of information that holds relationships and dependencies among various nodes. It integrates multiple languages based on knowledge in a layer of complexity. There are many challenges in designing multilingual knowledge graph like linguistic variance, data quality and consistency, resource intensiveness, semantic variability, security concerns, cross-lingual entity linking, maintenance

* PhD Research Scholar, Department of Library and Information Science The University of Burdwan, Burdwan, 713104. (mandalpsm@gmail.com) (First Author, Corresponding Author)
** Assistant Professor, Department of Library and Information Science The University of Burdwan, Burdwan, 713104. (sukumar.mandal5@gmail.com) (Co-author)

and up-to-date etc. To face these challenges, it requires overcoming the limitations of data quality, semantic and syntactic interoperability. On the other hand, cross lingual linking and evolution also make an ongoing challenge in multilingual knowledge graph. Despite of all these challenges, there are some opportunities for invention, innovation and advancement. This paper proffers the challenges and chances inherent in the design of MKGs (Multilingual Knowledge Graphs). It also sheds light on the intricate balance between challenges and chances in the pursuit of designing robust and inclusive multilingual knowledge representation systems.

2. Definition of the Key Terms

Monolingual, bilingualism, multilingualism, knowledge graph and multilingual knowledge graphs are the key terms in this study. Here, knowledge graph is a structural representation of knowledge. It represents the relationship between entities in a specific domain. Entities are represented as nodes and relationships are represented as edges or links. Each and every edges and nodes are associated with attributes or properties. The characteristic of Knowledge graphs include - entities, relationships, nodes and edge and attributes and properties, semantic interoperability, link data, scalability and extensibility. It enhances knowledge representation, discovery and utilization in various applications. A multilingual graph is a graphical representation of edges and nodes which are associated with multiple levels or attributes defined in multiple languages. Graph structure, language-dependent information, cross language connectivity, semantic interoperability and machine translation integration are the key element of multilingual knowledge graph. It gives an opportunity for developing cross-language representation and understanding. User-centric multilingual graphs provide an opportunity of multi language option for the users. The multilingual knowledge graph is associated with the application of cross-language knowledge representation for globalized information retrieval, natural language processing, and cross-cultural analytics. Though ‘multilingual graph’ and multilingual Knowledge graph have some similarities, they used in different context. Multilingual graphs refer the ability to represent graph element with multiple labels in different languages. It focuses on the linguistic diversity to represent the information in different languages. The relationship and semantics within the graphs are not designed systematically. On the other hand ‘multilingual knowledge graph not only represent entities and relationship but also integrate information in multiple languages. It highlights the structural representation of knowledge semantically. In it, entities, relationship and attributes are well designed and it is associated with ontology which is the representation of a specific domain. It focuses on the integration of knowledge for reasoning and getting information in different languages. The distinction lies on its semantic representation and the emphasis on the knowledge.

3. Case Study on Multilingual knowledge Graphs

Knowledge graphs not only improve the quality issues but also solve the problems of heterogeneity

of data. They integrate and exchange the data without ambiguity. The best example of Knowledge graph is Google Knowledge graph. It explores various interrelated web resources (Dong et al., 2014). BabelNet4 is an example of multi-lingual graph connects 16 millions entities on 284 languages including geographical name and word notes (Navigli & Ponzetto, 2012). Gabrilovich and Usunier publish many works on knowledge graphs creation and ontology (Bordes & Gabrilovich, 2014). Google provides discovery services for entities and rank the data with their relevance. The nodes represent the entities and edges represent the relation between edges (Cheng et al., 2022). Now, AI system is used for heterogeneous information to build up multilingual knowledge graphs (Ko et al., 2021) in recommender system and question-answering system. These works collectively contribute to the understanding of the challenges and opportunities in developing and utilizing multilingual knowledge graphs, offering insights into effective strategies, methodologies, and applications for handling linguistic diversity in the realm of knowledge representation and retrieval. Peng et al. (2023) focused on the challenges and chances of knowledge graphs in terms of two aspects. One is for building up AI system and other is to application of knowledge Graphs. This paper also discussed different technical challenges like knowledge graph embeddings, knowledge acquisition, knowledge graph completion, knowledge fusion, and knowledge reasoning. Tufchi, Yadav, and Ahmed (2023) discussed on various classification models across a variety of text and image-based datasets related to fake news. They examined on the core evaluation metrics for assessing the accuracy of news authenticity. Various challenges are detected in this research. Evenstein Sigalov (2023) shed light on Wikidata's potential as a lifelong learning process, enabling opportunities for improved Data Literacy and a worldwide social impact. Khan et al. (2024) explored 'Heterogeneous Transfer Learning' process in various disciplines like image and text classification, activity recognition, and cross-project defect prediction. Sorato et al. (2024) used word embeddings to investigate immigrant and refugee stereotypes in a multilingual and diachronic setting. They analyzed the Danish, Dutch, English, and Spanish portions of four different multilingual corpora of political discourse, covering the 1997-2018 periods.

4. Construction of Knowledge Graph

Multilingual Knowledge Graphs are complex in structure. It has various components like entities, relation, attributes and language tags etc. To construct a multilingual knowledge graph, there are several processes of integration of different languages like define the purpose and scope, select knowledge sources, entity recognition and alignment, standardized the data, ontology design, cross-language linking, enrichment of data, representation of data and documentation and accessibility. Data sources are selected through linked open data sets and vocabularies according to its availability, access, size, quality and connectivity. Next step is the integration of data, reconciliation, alignment and curation. For data integration, there are several processes like unification, first come/first serve or most representatives.

5. Accessing Data

Entity collection is accessible via API. There are two methods for API. One is getting HTML or JSON-LD formats and other is using its URI. These methods are chosen for interoperability of web services. URI should be differentiable, unambiguous and immutable. API provides another two methods for resource discovery and information retrieval in the Entity collections. First is the entity auto completion and other is the entity search. With its generic implementation, entity search allows API to formulate complex queries following the solor query syntax. Linked data protocol is used for the presentation of search results. Accessing data from a multilingual knowledge graph involves querying the graph to retrieve relevant information in one or more languages. The process is typically facilitated through query languages, APIs, or specialized interfaces designed to handle multilingual queries. Here is an example of multilingual knowledge graph accessing with WordAtlas. After putting the API Key, one can search one's desire output in multi languages.

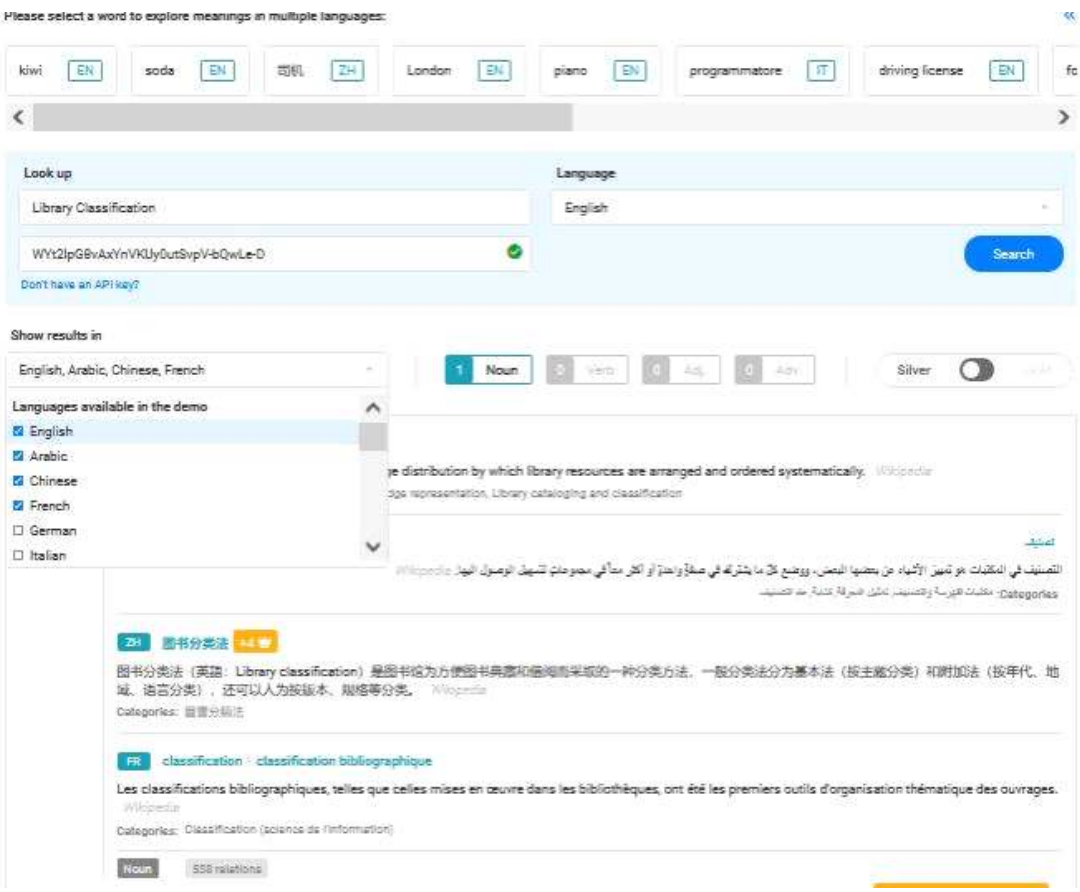

Fig. 1. Search Result of the meaning for 'Library Classification' in different Languages (Source: https://demo.babelscape.com/wordatlas)

6. Beneficiary of Multilingual Knowledge Graphs

For designing and smooth running of the multilingual knowledge graphs, creators, translators, domain experts, software developers, end users, privacy and security experts are responsible. Researchers, academicians, entrepreneur, administrators, technical developers, medical practitioners and content creators are the beneficiary of it. Researchers and academician can analyze the data or information from various languages. It opens a vista of knowledge to him to represent and to visualize the interdisciplinary subjects. Entrepreneur can investigate various linguistic markets which gives a vision on market trends and customers' feedback. It helps to promote customer supports worldwide. Governments and authorities can investigate various linguistic markets to understand market trends and customers views which help them to make a decision in multilingual and multi-cultural context. As technical developers are working with natural language processing (NLP) and machine language (ML), they have needed information and training data from various sources. Cross-lingual information retrieval system helps them to get information from cross linguistic platform. Content creators can use multilingual; knowledge graphs to get information from various linguistic and cultural audience for content creation, integration and curation. From multilingual knowledge graph, medical professionals and researchers can get health information across linguistic barriers. It helps them not only to communicate effectively but also foster global co-operation and collaboration in medical research.

7. Strength of Multilingual Knowledge Graphs

Multilingual knowledge graphs represent the knowledge in multiple languages. It is represented entities with nodes, edges and labels. Nodes represent any person, object, location or event. Edges are called links or lines. It connects two vertices symmetrically or asymmetrically. Label is a sort name which is applied to a node in graph. In graph database, it is used to shape the domain by grouping nodes into the same sets. Some strength of designing multilingual knowledge graphs are in the followings:

- Availability of multilingual content creation - The availability of multilingual content creation has significantly expanded, fostering global communication and cultural exchange. With the rise of advanced language processing technologies and user-friendly tools, individuals and businesses can effortlessly generate content in multiple languages. This accessibility promotes inclusivity, enabling diverse audiences to engage with information in their preferred language. From automated translation services to versatile content creation platforms, the digital landscape now accommodates linguistic diversity. This trend not only facilitates effective communication across borders but also empowers content creators to reach broader audiences, fostering a more interconnected and culturally rich online environment.
- Cross language information integration - Cross-language information integration in multilingual knowledge graphs is a pivotal advancement for seamless global knowledge sharing. It involves harmonizing data across diverse languages within a unified knowledge framework. By linking

entities and relationships across linguistic boundaries, these graphs enhance information interoperability. This integration ensures that users can access and navigate knowledge resources in their preferred language, fostering a more inclusive and accessible digital ecosystem. Through sophisticated algorithms and semantic mappings, cross-language information integration not only breaks down language barriers but also promotes a holistic understanding of interconnected global knowledge, facilitating collaborative research, cultural exchange, and effective communication on a multilingual scale.

- Enriched data analysis - Enriched data analysis within multilingual knowledge graphs signifies a sophisticated approach to extracting insights from diverse linguistic datasets. This advanced technique involves augmenting raw data with additional context, semantics, and language-specific nuances, amplifying the depth and accuracy of analysis. By integrating various languages seamlessly, enriched data analysis ensures a comprehensive understanding of global information trends, fostering more nuanced decision-making. Leveraging advanced algorithms, it enables businesses and researchers to uncover hidden patterns, cultural insights, and market trends across linguistic boundaries. This process not only refines data interpretation but also contributes to a more refined and inclusive approach to knowledge discovery in the realm of multilingual information.
- Adaptability to linguistic changes - Adaptability to linguistic changes is crucial in maintaining the relevance and effectiveness of multilingual knowledge graphs. As languages evolve, these graphs dynamically adjust to incorporate new terms, expressions, and linguistic nuances. Advanced algorithms continuously analyze linguistic shifts, ensuring real-time updates for accurate information representation. This adaptability not only preserves the integrity of the knowledge graph but also enhances its responsiveness to evolving user needs. By staying attuned to linguistic changes, these graphs remain valuable tools for cross-cultural communication, reflecting the fluid nature of languages and guaranteeing that users can access up-to-date, culturally relevant information in their preferred language.
- Multi-search and discovery facility - The multi-search and discovery facility in multilingual knowledge graphs revolutionizes information exploration by enabling users to seamlessly navigate across linguistic dimensions. This advanced feature integrates diverse search queries, allowing users to discover interconnected insights in multiple languages simultaneously. Through sophisticated algorithms and semantic associations, it streamlines cross-language exploration, providing a comprehensive view of interconnected knowledge. This facilitates efficient research, cross-cultural analysis, and a more profound understanding of global information trends. The multi-search and discovery capability not only enhances user experience but also empowers individuals and organizations to harness the richness of multilingual data for informed decision-making and comprehensive knowledge discovery.
- Support global co-operation - Multilingual knowledge graphs play a pivotal role in supporting global cooperation by breaking language barriers. These interconnected graphs facilitate seamless information sharing, transcending linguistic differences. Enabling collaboration across diverse cultures, they empower individuals and organizations to pool insights, fostering a united approach to problem-solving and innovation on a global scale.

8. Weakness of Multilingual Knowledge Graphs

Multilingual Knowledge Graphs (MLKGs) come with certain weaknesses that can impact their effectiveness and implementation. Here are some weaknesses associated with MLKGs:

- Syntactical Barriers: The accuracy and quality of automated translations can be a significant weakness. Errors in translation may lead to misunderstandings or misinterpretations of information, impacting the overall reliability of the knowledge graph.
- Linguistic Variability: Languages exhibit variations, dialects, and regional nuances. Capturing and managing these variations accurately in an MLKG can be challenging, leading to potential inconsistencies and inaccuracies.
- Data Quality and Consistency: Maintaining high-quality and consistent data across multiple languages is a complex task. Inconsistencies in terminology, data quality, or cultural context can hinder the reliability and usability of the MLKG.
- Limited Language Resources: Some languages may have limited digital resources available, including textual data and natural language processing tools. This can result in uneven coverage and quality of information across different languages.
- Semantic Ambiguity: Different languages may express concepts and relationships with varying levels of ambiguity. Resolving semantic ambiguity across languages requires careful consideration and may not always be straightforward.
- Cultural Sensitivity Issues: MLKGs may struggle to capture cultural nuances accurately. Cultural sensitivity is crucial in certain domains, and inaccuracies in representing cultural context can lead to misunderstandings.
- Resource Intensiveness: Building and maintaining MLKGs require significant resources, including human expertise in multiple languages, linguistic experts, and technology infrastructure. This can be resource-intensive and may limit accessibility for some organizations.
- Complex Query Processing: Querying an MLKG with complex queries across multiple languages can be challenging. Developing efficient and effective query processing algorithms that consider linguistic variations is a non-trivial task.
- Privacy and Security Concerns: Managing privacy and security concerns becomes more complex in a multilingual context. Ensuring compliance with privacy regulations across different languages can be challenging and may require careful attention.
- Integration Challenges: Integrating MLKGs with existing systems and databases may pose challenges. Ensuring compatibility and seamless interaction with diverse systems, especially those designed for monolingual environments, can be a hurdle.
- User Experience Challenges: Users interacting with MLKGs may face challenges related to the user interface, especially if the system needs to support multiple languages simultaneously. Providing a seamless and intuitive user experience across languages can be complex.
- Maintenance and Updates: Keeping MLKGs up-to-date with evolving languages and new information is an ongoing challenge. Regular updates are essential to ensure the relevance and accuracy of the knowledge graph over time.

9. Opportunities of Multilingual Knowledge Graphs

Some opportunities related to construct and use of multilingual knowledge graphs (MLKGs) are discussed here:

- Global Information Access: MLKGs enable global access to information by breaking down language barriers. Individuals from diverse linguistic backgrounds can benefit from a more inclusive and accessible knowledge base.
- Cross-Cultural Understanding: MLKGs provide an opportunity to enhance cross-cultural understanding by capturing and representing cultural nuances in different languages. This can facilitate research, collaboration, and communication across cultures.
- Multilingual Content Creation: Content creation for multilingual is valuable for websites, documentation, and other materials intended for audiences with diverse language preferences.
- Improved Search and Discovery: MLKGs enhance search and discovery capabilities across languages. Users can retrieve information more effectively, leading to a more efficient exploration of knowledge in a multilingual context.
- Facilitation of Multilingual Applications: MLKGs serve as a foundation for developing applications that operate in multiple languages. This is beneficial for natural language processing applications, chatbots, virtual assistants, and other language-driven technologies.
- Support for Multilingual Education and Research: MLKGs can support multilingual education by providing a rich knowledge base that spans different languages. Educational materials and resources can be made available to students in their preferred language. The knowledge graph offers a unified platform for exploring and analyzing information across different languages, fostering interdisciplinary studies.
- Enhanced Semantic Interoperability: MLKGs contribute to enhanced semantic interoperability by aligning concepts and relationships across languages. This ensures a consistent representation of knowledge, supporting interoperability with various systems and datasets.
- Preserving Cultural Heritage: MLKGs can contribute to the preservation of cultural heritage by representing knowledge in multiple languages. This is particularly important for domains like history, literature, and anthropology.
- Inclusive AI and Technology: MLKGs support the development of more inclusive artificial intelligence (AI) and technology applications. By incorporating multiple languages, these systems can better serve diverse user populations.

10. Challenges of Multilingual Knowledge Graphs

Though constructing multilingual knowledge graphs have many opportunities there are several challenges in developing and maintaining it. These challenges are in the followings:

- Syntactic Barriers: Language is dynamic and it changes time to time so it is difficult to get

accurate and high-quality translations in multiple languages. It becomes a challenge to represent knowledge graphs. Automated translation tools are trying to face these challenges with idiomatic expressions, cultural nuances, and domain-specific terminology.

- Linguistic Variations: Different languages have dialects, and regional differences. These linguistic variations including syntactic and semantic differences, poses become a challenge in constructing MLKG.
- Cultural Sensitivity: It is difficult to translate cultural differences in expression, meaning, and context in a standardized way.
- Resource Intensiveness: Constructing and maintaining multilingual knowledge graphs require important resources including linguistic expertise, translation services, and computational power. This is a barrier for smaller organizations with limited resources.
- Privacy and Security Concerns: Privacy and security concerns become complex in a multilingual context. To protect privacy and security, careful attention and regular supervision are needed.
- Integration Challenges: Integration of MLKGs with existing systems and databases, especially those are designed for monolingual environments, become challenging.
- User Interface Design: Designing a user interface that provides a seamless and intuitive experience across multiple languages can be complex. The interface needs to be user-friendly for different languages.
- Maintenance and Updates: Keeping the MLKG up-to-date with evolving languages, new information, and changes in the knowledge domain is an ongoing challenge. Regular updates are crucial for maintaining the relevance and accuracy of the graph.
- Cross-Lingual Entity Linking: Linking entities across languages (entity alignment) is challenging due to differences in naming conventions, entity structures, and cultural contexts. Ensuring accurate cross-lingual entity linking is a complex task.
- Cross-Lingual Evaluation: Evaluating the performance and effectiveness of MLKGs across different languages is challenging. Metrics and benchmarks for cross-lingual knowledge graph evaluation need to be carefully designed.

11. Conclusion

The challenges are not only help to construct the evolution of multilingual knowledge graph representation but also explore an integrated world of information. In future, multilingual knowledge graph overcomes all the linguistic nuances, technical barriers like semantic interoperability, data harmonization etc and enhance cooperation and collaboration throughout the world. With the advancement of the research, the evolution of multilingual knowledge graph becomes not only a technical endeavor but also it becomes the testament of the power of global knowledge society.

References

Bordes, A., & Gabrilovich, E. (2014). Constructing and mining web-scale knowledge graphs: KDD 2014 tutorial. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1967-1967. New York New York USA: ACM. https://doi.org/10.1145/2623330.2630803

Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition, 121*, 108218. https://doi.org/10.1016/j.patcog.2021.108218

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 601-610. New York New York USA: ACM. https://doi.org/10.1145/2623330.2623623

Evenstein Sigalov, S., & Nachmias, R. (2023). Investigating the potential of the semantic web for education: Exploring Wikidata as a learning platform. *Education and Information Technologies, 28*(10), 12565-12614. https://doi.org/10.1007/s10639-023-11664-1

Khan, S., Yin, P., Guo, Y., Asim, M., & Abd El-Latif, A. A. (2024). Heterogeneous transfer learning: Recent developments, applications, and challenges. *Multimedia Tools and Applications.* https://doi.org/10.1007/s11042-024-18352-3

Ko, H., Witherell, P., Lu, Y., Kim, S., & Rosen, D. W. (2021). Machine learning and knowledge graph based design rule construction for additive manufacturing. *Additive Manufacturing, 37*, 101620. https://doi.org/10.1016/j.addma.2020.101620

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217-250. https://doi.org/10.1016/j.artint.2012.07.001

Oelen, A., Jaradeh, M. Y., Stocker, M., & Auer, S. (2020). Generate fair literature surveys with scholarly knowledge graphs. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 97-106. Virtual Event China: ACM. https://doi.org/10.1145/3383583.3398520

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review, 56*(11), 13071-13102. https://doi.org/10.1007/s10462-023-10465-9

Sorato, D., Lundsteen, M., Ventura, C. C., & Zavala-Rojas, D. (2024). Using word embeddings for immigrant and refugee stereotype quantification in a diachronic and multilingual setting. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-023-00243-6

Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval, 12*(2), 28. https://doi.org/10.1007/s13735-023-00296-3

[About the authors]

Mr. Partha Sarathi Mandal is working as Librarian in Shyamsundar Ramlal Adararsha Vidyalaya. He obtains M.Phil; MLIS and M.A from the University of Burdwan and presently pursuing his PhD in the Department of Library and Information Science, the University of Burdwan. He qualifies UGC-NET in 2019. He has published 19 research articles in various National and International journals. His area of interest includes: Internet of Things, Artificial Intelligence, Webometrics, Multilingual Information Retrieval System, Semantic Web, Linked Open Data, Open Knowledge System etc. Contribution in current study, he selected the research problem, designed the research methodology and collected data for the study.

Dr. Sukumar Mandal is designated as Assistant Professor in the Department of Library and Information Science at The University of Burdwan. He obtains M.Com from the University of Burdwan. He also obtains MLIS and Ph.D from The University of Burdwan. His area of interest are Integrated Library system, Digital Library System, Community Information System and Services, Institutional Digital Repository, Multilingual Information Retrieval System, Semantic Web, Library Administration and Automation, Thesaurus Construction, Visual Vocabulary, Linked Open Data, Open Access, Open Knowledge System etc.