SGKPS: A Semantic AI-Driven Strategy for Knowledge Graph Population for Sustainable Journalism as a domain of choice

Gokuldas Vedant Raikar [1,†] and Gerard Deepak [2,*,†]

[1] Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal 560064, India
[2] BMS Institute of Technology and Management, Yelahanka, Bengaluru,

Abstract

A framework that complies with Web 3.0 is strategically necessary which is semantically inclined for knowledge graph population specifically for domains which are of importance like sustainable journalism. Sustainable journalism is a intersection of sustainability and journalism. This paper uses a incremental knowledge addition model from the perspective of the dataset for extraction of terms and categories from the dataset which is further nicened with topic modelling framework like the LDA and the standard repositories like googles knowledge graph repositories is used to obtain the initial knowledge graphs and subgraphs. Strong LLM like the BLOOM helps in caption generation with the LDA enriched terms. Journalism knowledge stack comprising of eBooks , glossary term make index terms along with the web 3.0 meta data is also a important source of augmenting knowledge into the proposed framework, the Bi-LSTM based classifier classifies the journalism knowledge stack and includes a strong deep learning model into the framework, the presence of jiang conratt index , Sim Rank on differential step deviance measures and the optimization which is achieved using the invasive weed optimization gives the optimal solution set that is best in its class which is further merged with the existing knowledge graph through semantics relatedness computation which leads to the best in class model which integrates a robust deep learning model, large language model, and topic model namely the LDA and a series if semantic similarity measures with differential threshold along with optimization gives a hybrid model for knowledge graph generation and population with overall precision of 96.77% , recall of 97.18% and FDR rate of 0.04 is for These properties of the suggested framework makes the model the best in class model for knowledge graph population.

Keywords

Knowledge graph, Knowledge graph population, sustainable Journalism, deep learning classifier, large language model

1. Introduction

Real-world items, occasions, concepts, and the connections between them are represented in an organized manner in a knowledge graph. It is arranged according to theoretical constructs called organizing principles and is stored in a graph database, utilizing edges to define relationships between entities and nodes to represent entities. The process of populating and enriching knowledge graphs by systematic extraction, integration, and organization of data from many sources is known as knowledge graph generation and population. Through this process, unstructured data is converted into structured entities and relationships, creating a comprehensive network that represents concepts found in the actual world and their linkages.

AISD-2024: Second International Workshop on Artificial Intelligence: Empowering Sustainable Development, October 2, 2024, co-located with the Second International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2024), Virtual Event, Lucknow, India.

[*] Corresponding author.

[†] These authors contributed equally.

gokuldas.mitblr2022@learner.manipal.edu (G. V. Raikar); *gerard.deepak.christuni@gmail.com* (G. Deepak)
0009-0009-8256-4638 (G. V. Raikar); 0000-0003-0466-2143 (G. Deepak)

Knowledge graphs are constructed and enlarged dynamically using methods including entity identification, relationship extraction, and natural language processing. This enhances applications like search engines, recommendation systems, and AI-driven insights and makes it possible to represent data more accurately. A well-functioning knowledge graph population guarantees the precision and applicability of the data while also augmenting the graph's capacity to produce useful insights and facilitate sophisticated data-driven applications. Press coverage is essential for educating the public, establishing opinions, and keeping authority figures responsible.

It entails obtaining, evaluating, and disseminating news and information to various audiences while making sure that the truth is appropriately and honestly conveyed. Print, internet, and broadcast media are just a few of the platforms on which modern journalism operates. It constantly adapts to the fast-paced developments in technology and the changing needs of its audience. Journalism's fundamental goal is still to empower people and promote an informed public by disseminating accurate, timely, and transparent information. The goal of journalism sustainability is to make sure that news organizations run in a way that promotes long-term social, economic, and environmental well-being. This entails implementing strategies to lessen carbon footprints, like favoring digital distribution over print, and encouraging moral reporting that upholds social justice and openness.

Journalism can safeguard its own future in an increasingly digital and resource-conscious world while also advancing larger societal objectives like social justice and environmental conservation by placing a high priority on sustainability. In sustainable journalism, knowledge representation and reasoning, or KRR, is the process of using sophisticated analytical methods and structured data to improve the impact, accuracy, and relevance of news reporting. Journalists can methodically arrange and evaluate data on social and environmental concerns by utilizing KRR frameworks, which makes it possible to provide more intelligent and nuanced reporting. This strategy encourages the creation of material that educates readers about sustainability issues while simultaneously endorsing evidence-based remedies and promoting thoughtful conversation. The creation of transparent, responsible, and powerful narratives that encourage improvement in sustainability practices and policy is facilitated by the incorporation of KRR into journalism. By combining many datasets, including policy and environmental data, knowledge graph population is revolutionizing sustainability reporting in today's news. This technique improves the quality and breadth of reporting by enabling journalists to develop intricate, interrelated perspectives on sustainability-related issues. News organizations can use knowledge graphs to provide interactive, real-time content that tracks progress and emphasizes trends, promoting more informed public debates and efficient coverage of social and environmental issues. With deep learning providing strong capabilities for extracting and integrating complicated data from several sources, knowledge graph population for sustainable journalism is using deep learning more and more.

Deep learning improves the capacity to interpret unstructured data, including news articles, reports, and social media information, and converts it into organized, useful insights by utilizing neural networks and sophisticated algorithms. With this method, journalists can create in-depth knowledge graphs that depict intricate connections and patterns in sustainability. Consequently, news outlets are able to present more complex and data-driven reporting, reveal previously unnoticed trends, and facilitate better-informed public discussion on social and environmental issues. Knowledge graph generation and improvement across the whole development process are being transformed by deep learning and generative AI. Deep learning models can automatically mine enormous volumes of text and data to find pertinent entities and relationships at the beginning of the knowledge extraction process. Then, with the aid of generative AI, this data is structured and arranged into logical and practical formats. These solutions improve the alignment of heterogeneous data sources during the integration phase, guaranteeing completeness and consistency in the knowledge graph. Lastly, by learning from fresh data and user interactions, deep learning algorithms and generative AI tools enable ongoing updates and expansions, maintaining the accuracy and applicability of the knowledge network. When combined, these cutting-edge methods simplify and improve the process of creating vibrant, perceptive knowledge graphs.

1.1 Motivation

The primary motivation is the lack of strategic knowledge graph generation and population mechanism for sustainable journalism which is a intersection of sustainability and journalism in the present day era and most importantly the web 3.0 cannot interpret the traditional models that doesn't include semantic intelligence and interpreted knowledge addition is definitely absent, a framework that hybridizes LLMs with strong deep learning models and a system of semantics symmetric measures is the motivation of this paper.

1.2 Contributions

The principal contribution of the suggested architecture includes the presence of topic modelling based LDA and google knowledge graph repository to fill up the initial knowledge graphs and subgraphs from the dataset extracted terms and categories. LDA plays a role in caption generation through enrichment using BLOOM large language model the presence of generalized knowledge stack which comprises of eBooks , Glossary , index terms and web 3.0 metadata which need the high density auxiliary knowledge which is domain centric and it's classification using a powerful deep learning model is an additional framework addition , the inclusion of jiang conratt index and SimRank at various points in the suggested framework pipeline with strong metaheuristic like invasive weed optimization is also a contribution to the paper which will create a best in class for populations of knowledge graph.

1.3 Organization

This paper is divided into the following sections, section 1 contains the introduction to the paper, section 2 structed to encompass the related works, section 3 shows the proposed methodology, the results are contained in section 4 and section 5 contains the conclusion of the paper.

2. Related Works

This section contains the related works for knowledge graph population for sustainable journalism as a domain of choice.

2.1. knowledge graph population and completion

The paper by Jay Pujara et al.,[1] presents a technique that uses probabilistic soft logic (PSL) to convert ambiguous extractions about entities and their relationships into a knowledge graph. This method incorporates ontological constraints, removes noise, deduces missing data, and decides which candidate facts to include by discussing candidate facts and their extraction confidences in concert. Examined on both artificial and real-world datasets, the approach demonstrates enhanced AUC and F1 scores with noticeably shorter running times than previous approaches. In the publication by Baoxu Shi et al., [2] ConMask is presented for open-world Knowledge Graph Completion (KGC), which overcomes the drawbacks of conventional closed-world techniques. ConMask outperforms existing models in both open and closed-world KGC tasks by using entity name embeddings, text descriptions, and a convolutional neural network with content masking to link unseen entities to the Knowledge Graph.

The research by Tim Weninger et al., [3] presents ProjE, a neural network model that, using a smaller parameter size, performs 37% better than the most advanced methods currently available. ProjE learns joint embeddings and makes minor adjustments to the loss function to improve knowledge graph completion and fact-checking. The study by Guoliang Ji et al.,[4] provides introduction to TranSparse, a framework that achieves state-of-the-art performance and significantly surpasses earlier models in terms of Trans (E, H, R, and D) in triplet classification and prediction of links tasks. TranSparse addresses the disparities and diversity in knowledge graphs by using adaptive sparse matrices for transfer operations. To achieve best in class performance on temporal facts in dynamic environments, the paper by Tingsong Jiang et al. [5] presents a novel time-aware knowledge graph completion model that incorporates

temporal information using a time-aware embedding model and a joint inference model based on Integer Linear Programming.
The study by Théo Trouillon et al, [6] demonstrates that all real square matrices can be represented as the real part of unitarily diagonalizable matrices and presents a scalable, complex-valued embedding approach for knowledge graph completion that strikes a compromise between expressiveness and complexity. With linear space and time complexity, this approach—which uses a Hermitian dot product—performs better than current models on link prediction benchmarks.

2.2. Knowledge Graphs in Journalism

The study by Marc Gallofré Ocaña et al.,[7] examines Journalistic Knowledge Platforms (JKPs) utilized in newsrooms that are based on knowledge graphs and emphasizes how they can improve content analysis, automate metadata annotation, and streamline workflows. To lower production costs and enhance information quality while adjusting to new journalism forms and reader demands, future JKPs will require better extraction mechanisms for textual and multimedia news. The work by Ole A. Christensen et al.,[8] explores how to use natural language processing, machine learning, and knowledge graphs to leverage open, social, and other data sources for journalism. It presents the adaptable News Hunter prototype and outlines development plans for the future to meet the needs of digital journalism. The paper by Andon Tchechmedjiev et al., [9] presents ClaimsKG, a semi-automated pipeline that builds a knowledge graph with verified claims with structured metadata from RDF annotations and fact-finding websites. The dataset comprises 6,606,032 triples and 28,383 claims, providing a query able resource for claim analysis and validation.

2.3. Sustainable Journalism as a domain of choice

The work by Stephen J. A. Ward et al., [10] examines the five stages of the development of journalism ethics: the Enlightenment public ethic, the liberal press theory of the 19th century, criticisms of this doctrine in the 20th century, and the development of contemporary journalism ethics in response to challenges posed by digital and global media. The paper by Joaquim Fidalgo et al., [11] highlights how new media forms like citizen journalism and social platforms challenge the traditional roles of professional journalists as it looks at the effects of the digital age on journalism. Eliot Friedson's concepts of "practice ethics" and "institutional ethics" are cited as it discusses the ethical implications of this shift and the need for clearer definitions of journalism and journalistic roles in the changing media landscape. The study by William P et al.,[12] finds that knowledge gaps between groups with higher and lower levels of education are smaller and less pronounced among heavy readers of newspapers and television news than among light users. Nevertheless, neither on-screen nor newspaper use has an impact on voting gaps, even though newspaper use is associated with wider differences in overall political participation.

3. Proposed Architecture and Methodology

The suggested architecture for the semantic artificial intelligence (AI) driven knowledge graph population strategy for sustainable journalism as a preferred domain is presented in this section.

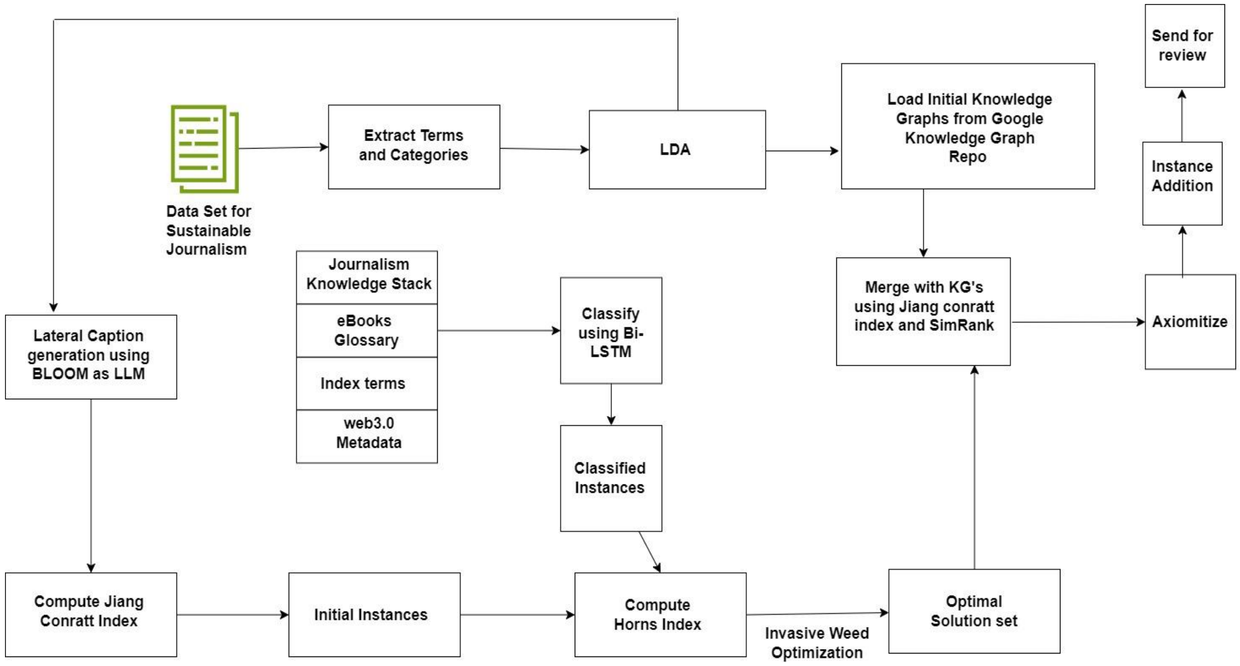
Data Set for Sustainable Journalism
Extract Terms and Categories
LDA
Load Initial Knowledge Graphs from Google Knowledge Graph Repo
Send for review
Instance Addition
Journalism Knowledge Stack
eBooks Glossary
Index terms
web3.0 Metadata
Classify using Bi-LSTM
Merge with KG's using Jiang conratt index and SimRank
Axiomitize
Lateral Caption generation using BLOOM as LLM
Classified Instances
Compute Jiang Conratt Index
Initial Instances
Compute Horns Index
Invasive Weed Optimization
Optimal Solution set

Figure 1: depicts the proposed architecture of framework for knowledge graph generation for sustainable journalism as a domain of choice.

The strategic framework for knowledge graph populations for sustainable journalism as a preferred domain is illustrated in Figure 1 of the proposed architecture. Prediscussed sustainable journalism is engineered in such a way that it has contents of both sustainability as well as journalism as a strategic domain. From figure 1 it is indicated that the dataset for sustainable journalism is subjected to extraction of terms and categories, terms are directly obtained from the dataset documents itself and the category are obtained directly from the categories of the dataset itself. The terms are derived by using the term frequency and inverse document frequency of TFIDF, which targets the frequent terms and the rare terms both within and throughout the document corpus, to identify keywords of occasionally occurring terms and frequently occurring terms. The terms and categories extracted from the dataset for sustainable journalism is subjected to Latent Dirichlet Allocation (LDA), LDA is an analytical topic modelling framework which Achilles' heel some of the knowledge from the web and web 3.0 laterally reduces the community gap between the knowledge of the web3.0 and the knowledge of the embedded into the localized model.

The reason why Latent Dirichlet Allocation is preferred is because it restricts the number of instances and mainly focusses on the relevance of entities, the terms derived from the LDA undergo loading as knowledge graphs and sub-graphs via Google Knowledge Graph API. The terms extracted from the LDA and all categories are loaded as Google knowledge graphs and subgraphs via Google Knowledge Graph API. These knowledge graphs and subgraphs that emerge from the Google Knowledge Graph following reconstitution are then combined and utilized as the initial knowledge graphs and subgraphs, resulting in the ones with the greatest number of optimal entities emerging from the pipeline. The dataset for sustainable journalism is subsequently subjected to lateral caption generation through the already obtained terms using BLOOM large language model API and The entities emerging from this pipeline are further used into the model, most importantly a CAC of journalism comprising of e-books , glossary terms and index terms from the eBooks and web 3.0 metadata ,a metadata is generated with DSpace acting as a strategic tool of choice and it is ensured that the DSpace is used and configured under agent named Agentspeak where it filters the metadata which is obtained and links it directly with the dataset based entities and this knowledge stack is extensively heavy because it also contains the metric points to the metadata. This is subjected to application using

A robust classifier for deep learning algorithm called the Bi-LSTM classifier, the Bi-LSTM classifier classifies the journalism on its stack and this leads to the classified instances and these classified instances are further used in the model. The instances that come out of the caption generation using the BLOOM LLM is again subjected to the computation of jiang conratt index with the LDA enriched terms and categories at the juncture in order to filter out ill beneficial instances, the jiang conratt index used at a step-deviance of 0.15 the step deviance is empirically large due to the reason and the fact that jiang conratt index is a stronger index and both the lateral captions as well as the LDA both are generated using the provided dataset and henceforth the window for the step deviance measure is little higher, these initial instance which are obtained is further exposed to the horns index calculation with the step deviance at 0.12 thereafter the classified instances of that of the journalism knowledge stack which contains the metadata that the Bi-LSTM classifier has classified.

The results that come out of this pipeline are subjected to the computation of the horns index and the invasive weed optimization this yields the optimal solution set because the horns index is run and the metaheuristic optimization model which uses nature inspired optimization technique and horns index as a criteria function and the horns index is also set to an empirical threshold which makes it quite optimal and this is where we are using Agentspeak as a strategic domain , this optimal solution set and the initial knowledge graphs etc. are merged using the jiang conratt index and SimRank with jiangs conratt index set to a threshold of 0.12 and similarity index set to a step deviance of 0.12 again this variance in step deviance will lead to participating index of jiang conratt and SimRank and due to the fact that empirical aggregation of auxiliary knowledge with diversification of auxiliary knowledge is required and this merging of entities with the initial generated knowledge graph and subgraphs and by establishing at least a single link between the all the participating entities is the primary requirement and is achieved using Agentspeak this further leads to the schematization is done by using a assertion based agent to correlate the entities which are participating and further inference addition from the corpus of the web3.0 is achieved and finally we are sending it for review by domain experts which have the furnished the participating knowledge graph.

3.1 BLOOM Large Language Model

Large language models (LLMs) have revolutionized AI research by performing a broad range of language tasks based on user instructions. However, the high costs and exclusive access have made it challenging for academics, nonprofits, and smaller research labs to work with these models. To address this, we're introducing BLOOM, the first fully transparent multilingual LLM, developed through an unprecedented global collaboration. BLOOM can produce text in 13 programming languages and 46 natural languages with 176 billion parameters. BLOOM is the first model with more than 100 billion parameters for many of these languages, such as Spanish, French, and Arabic. The accomplishment is the product of a year-long endeavor that involved more than 250 institutions and over 1,000 researchers from more than 70 countries. It was made possible by a €3 million grant from the French research agencies CNRS and GENCI, and it culminated in a 117-day training run on the Jean Zay supercomputer in Paris. Now that BLOOM is freely accessible, researchers can use it to run and study it, thoroughly analyzing its internal operations and overall performance.

Anyone who agrees to the Responsible AI License can use the model locally or via cloud services, thanks to its integration with the Hugging Face ecosystem. For those without access to high-performance hardware, an inference API is in development to facilitate large-scale use. Meanwhile, an early version is available for quick testing and prototyping on the Hugging Face hub. This is just the beginning. We're committed to continually enhancing BLOOM's capabilities, including making it more instructible, adding more languages, and developing more efficient versions. BLOOM represents not just a single model but the foundation of a growing family of models, and we welcome community contributions to expand and refine it further. The BLOOM model's architecture is designed with several advanced features to ensure high performance and stability. It uses the Transformer framework, specifically a causal decoder-only setup, which was found to be the best for handling zero-shot tasks compared to other model types. The research team experimented with various components and settings to optimize the model. BLOOM incorporates ALiBi positional embeddings, which adjust attention scores based on the distance between elements, leading to smoother and more effective training. It also includes an extra layer normalization right after the embedding layer to improve stability, especially with the use of bfloat16 precision. Additionally, the development process involved careful data preprocessing, such as removing duplicates and protecting privacy, and employed multitask prompted finetuning to enhance its ability to generalize across different tasks. These elements reflect a thoughtful blend of innovative and established techniques to maximize the model's performance.

3.2 Latent Dirichlet Allocation (LDA)

A probabilistic generative model called Latent Dirichlet Allocation (LDA) is used to represent documents as mixtures over latent topics, with a word distribution characteristic for each. The following steps are involved in creating each document w in a corpus D: Initially, N, the word count, is selected using a Poisson distribution with parameter ξ . Next, a topic distribution θ is sampled from a distribution of Dirichlet with parameter α . For every word w_n, a topic z_n is selected from an array of multinomials parameterized by θ , and then w_n is computed from a multinomial distribution conditioned on z_n ,with probabilities defined by β as shown in equation 1, 2 and 3 .

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

$$p(w \mid \alpha, \beta) = \int p(\theta \mid \alpha) [\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)] d\theta \tag{2}$$

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left[\prod_{n=1}^{N} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right] d\theta_d. \tag{3}$$

The LDA model's hierarchical structure involves corpus-level parameters α\alphaα and β\betaβ, document-level topic distributions θ_d , and word-level topics z_{dn} and words w_{dn}. Unlike simple Dirichlet-multinomial clustering models, LDA allows each document to be associated with multiple topics, offering a richer representation of document content. The LDA model assumes that the number of topics and their word distributions are fixed but unknown, and it uses Dirichlet distributions to model these topics. The model is hierarchical: it has a corpus-level distribution for topics, document-level distributions for topic mixtures, and word-level distributions conditioned on the topics. Unlike simpler clustering models where each document is associated with a single topic, LDA allows for multiple topics within a document, reflecting a more nuanced and flexible representation of the document's content.

3.3 Horns Index

Horn's Index, also known as the Horn-Morisita Index, is a measure used to quantify the similarity between two sets of count data. It builds on the Morisita index, which calculates how likely it is that two randomly selected individuals from two different datasets will belong to the same species. Horn's Index refines this approach by adapting the Morisita index for use with proportions or transformed count data. The formula involves computing an adjusted measure of overlap, which accounts for sampling with replacement, making it versatile for various types of data. This index is particularly useful in ecological studies and recommendation systems where comparing the similarity between different datasets or user preferences is essential. It is calculated as shown in equation 6.

$$d(x, y) = 1 - \frac{\frac{\sum_i x_i^2 \sum_i u_i^2}{N_x^2 \quad y}}{(\lambda_x + \lambda_y)} \tag{4}$$

$$\lambda_x = \frac{\sum_i x_i^2}{N_x^2} \tag{5}$$

$$\lambda_y = \frac{\sum_i y_i^2}{N_y^2} \tag{6}$$

Where λ_x and λ_y as shown in equation 5 and 6 respectively are computed based on proportions, making the index applicable to a wider range of data types. These formulas provide a quantitative measure of overlap or similarity between two datasets, useful in various fields such as ecology and data analysis.

3.4 DBSCAN Clustering

The density of data points is used by the clustering algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to identify clusters, separating regions with high point density from those with low density. Two primary parameters are used by the algorithm: minPts and ϵ. A point x_i is classified as a Central Idea if it has ideally minPts neighbours within a distance ϵ, which is mathematically expressed as shown in equation 7:

$$| \{x_j \ | \ d(x_i , x_j) \leq \epsilon\} | \geq minPts \tag{7}$$

where $d(x_i , x_j)$ is the distance between points x_i and x_j, and ∣⋅∣ denotes the number of points within ϵ of x_i. Border Points are those within ϵ distance of a Core Point but do not meet the Core Point criteria themselves. Noise points don't belong in the core or border category and do not meet the density criteria of any Core Point. The algorithm works by selecting an arbitrary point and checking if it meets the Core Point criteria. If it does, it expands the cluster by recursively including all reachable points within ϵ\epsilonϵ. For parameter estimation, minPts are typically set to at least D+1, with 2×D often used for larger or noisier datasets. The ϵ value is determined using a k-distance graph, where k = minPts−1 . The optimal ϵ is identified by locating an "elbow" in the graph, which helps balance between too many clusters (high ϵ) and too few (low ϵ). The choice of distance function $d(x_i, x_j)$ also affects clustering results and should be tailored to the dataset's characteristics.

3.5 Invasive Weed Optimization

The algorithm known as Invasive Weed Optimization (IWO) simulates the behavior of weed invasion to solve optimization problems through a stochastic approach. The process involves four key stages: initialization, seed propagation, spatial diffusion, and competitive exclusion. Initially, parameters such as the population size N_0, maximum population N_{max}, dimension D, maximum iterations max iter, seed limits max seed and min seed, and standard deviation bounds $\sigma_{\ init}$ and $\sigma_{\ max}$ are set. The population is initialized uniformly across the search space to ensure diversityA weed's ability to produce seeds during seed propagation is correlated with how fit it is. Specifically, if f represents the fitness of a weed, then the count of seeds m_{seed} is determined by the formula shown in equation 8:

$$m_{seed} = min\ seed\ + \frac{(f-fmin)}{(fmax-fmin)} \times\ (max\ seed - min\ seed) \tag{8}$$

where $fmax$ and $fmin$ are the highest and lowest values of fitness values in the current population, respectively. For spatial diffusion, seeds follow a normal distribution with a standard deviation σ and a zero mean , which is updated according to equation 9:

$$\sigma(iter) = \sigma_{\ init} - (\frac{\sigma_{\ init} - \sigma_{\ final}}{max\ iter}) \times iter \tag{9}$$

where iter is the current iteration, and n is a nonlinear harmonic factor typically set to 3. Finally, competitive exclusion is applied when the population reaches Nmax, where weeds are ranked by fitness, and only the best-performing ones are retained while others are discarded to maintain population size and improve the robustness of the search.

3.6 Jaccard Similarity

Jaccard Similarity is a popular metric for determining how similar two sets or binary vectors are to one another and is particularly valuable in various data science applications. For binary vectors, it evaluates the proportion of attributes that are present in both vectors relative to those that are present in at least one vector. Mathematically, for two binary vectors x and y, Jaccard Similarity Jaccard(x,y) is defined as shown in equation 10:

$$Jaccard(x,y) = \frac{|x \cap y|}{|x \cup y|} \tag{10}$$

where $|\ x \cup y\ |$ represents the count of attributes where both vectors have a value of 1, and $|\ x \cup y\ |$ denotes the total number of attributes where at least one vector has a value of 1. This metric is particularly helpful in situations where an attribute's presence or absence is the only important factor, such as in text mining or customer behavior analysis. In the context of sets, Jaccard Similarity assesses the overlap between two sets relative to their combined size. It is calculated as shown in equation 11:

$$Jaccard\ Similarity = \frac{|A \cap B|}{|A \cup B|} \tag{11}$$

where A and B are the two sets in question. This ratio calculates the proportion of shared elements between the two sets to the total number of unique elements in each set. The Jaccard Similarity varies from 0, indicating no shared elements, to 1, indicating complete overlap. This metric is instrumental in tasks such as comparing documents or customer profiles, providing a clear indicator of similarity or dissimilarity based on shared features.

3.7 SimRank Similarity

SimRank evaluates how similarly two objects are positioned within a network and is a flexible similarity metric based on graph theory. Two objects are deemed similar if they are related to other similar objects, according to the basic tenet of SimRank. Its recursive formula, which gives the following definition of similarity between two nodes a and b, encapsulates this idea in equation 12:

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_i(b)) \tag{12}$$

Here, I(a) and I(bindicate the respective sets of in-neighbors for nodes a and b, and C represents a decay factor between 0 and 1. The similarity s(a,b) is updated iteratively until convergence, starting from an initial state where s(a,a) = 1 for any node a and s(a,b) = 0 for a ≠ b. The iterative approach involves computing s_{k+1}(a, b) based on the similarity scores from the previous iteration s_k(a,b) refining the similarity measure over successive iterations.

This recursive nature allows SimRank to capture the structural context of nodes in a network, making it applicable to various domains, such as web page similarity or collaborative filtering in recommendation systems. However, while SimRank provides a robust framework for measuring similarity, it may sometimes yield unreasonable similarity scores due to various influencing factors. These issues can be mitigated by integrating evidence weight factors, additional terms, or alternative methods such as PageRank-based adjustments.

3.8 Jiang conratt index

An indicator of the semantic similarity of concepts inside a taxonomy or ontology is the Jiang-Conratt index. Information retrieval and natural language processing benefit greatly from it. Based on the idea of semantic similarity, this index aims to quantify the degree of relatedness between two concepts by assessing how similar they are within the framework of a particular ontology. When considering the taxonomy's depth of concepts and the shortest path between them, the Jiang-Conratt index provides a more nuanced measure of similarity than conventional techniques. It offers a strong foundation for determining how closely linked two concepts are by taking into account both the taxonomy's general structure and specificity, making it an invaluable tool in applications that require a deep understanding of the hierarchical relationships between terms as shown in equation 13.

$$JCI(c1, c2) = max(0, IC(c1) + IC(c2) - 2 \times IC(LCS(c1, c2))) \tag{13}$$

where IC(c) represents the Information Content of concept ccc, and LCS(c1,c2) is their Lowest Common Subsumer. This index measures similarity by assessing how closely related concepts are in terms of their hierarchical position and shared information.

3.9 AgentSpeak

Programming language AgentSpeak is intended for use with agent-oriented applications. It is built upon the belief-desire-intention (BDI) architecture, which models intelligent, self-governing agents. AgentSpeak (formerly known as AgentSpeak(L)), created by Anand Rao in 1996, combines the BDI model with logic programming to make the development and administration of intelligent agents easier. It has influenced the advancement of knowledge about multi-agent systems and offers an organized method for putting beliefs, desires, and intentions into practice. Through a number of formalizations and extensions, AgentSpeak has developed and is still widely used, especially when combined with the Jason platform to create intricate multi-agent systems. Its use in formal verification and advanced agent programming research and competitions is significant.

4. Performance Evaluation and Results

Dataset description: The Implementations are conducted on a single large, related dataset focusing on sustainability in media, journalism and sustainable journalism separately. These datasets were merged together by using a customized annotator to annotate the categories of the dataset and to populate the categories based on the keywords of the dataset and single large dataset was formulated and using the keywords and the annotations and the categories of the dataset, a crawler was devised to crawl the documents in web 3.0 and to add with huge number of annotations and single large datasets comprising of mainly seven major datasets namely, [13] Statista Research Department (2024). Presence of sustainability in media plans in Europe 2023, by stakeholder type [Dataset], [14] Wessler, Hartmut (2023). Sustainable media events? Production and discursive effects of staged global political media events in the area of climate change [Dataset] , [15] Yuriy Syerov (2023). Social Media Big Dataset for Research, Analytics, Prediction, and Understanding the Global Climate Change Trends [Dataset] , [16] Raihan C. A; The Native Tribe; Niketh M. S; Challa Megha Sai (2022). Role of Social Media in Fostering a Sustainable Society through Crowdfunding [Dataset], [17] (2024). Journalism and new Technologies, 2020 - Dataset - B2FIND [Dataset], [18] Work with Data (2024). Journalism-Methodology [Dataset] ,[19] Harvard Dataverse (2022). A method for measuring investigative journalism in local newspapers [Dataset] and a single large dataset was formulated

Table 1.
Comparison of Performance of the proposed SGKPS with other approaches

Model	Average Precision %	Average Recall %	Average Accuracy %	Average F-Measure %	FDR
EKGPS [20]	90.18	92.57	91.37	91.35	0.10
PGKP [21]	92.74	93.84	93.29	93.28	0.08
KGPOKG [22]	94.12	95.81	94.96	94.95	0.06
Proposed SGKPS	96.77	97.18	96.97	96.97	0.04

Precision, recall, accuracy, and F-measure are the preferred metrics for evaluating the performance of the proposed SGKPS, a semantic AI-driven strategy for knowledge graph population in the domain of choice for sustainable journalism. False discovery rate (FDR) is also used as an auxiliary metric. The purpose of the precision, recall, accuracy, and F-measure is to gauge how well the framework performs in terms of outcomes, while the false discovery rate (FDR) measures how many false positives are caused by the suggested framework. The proposed SGKPS has the highest average precision (96.77%), highest average recall (97.18%), highest average accuracy (96.97%), highest average F-measure (96.97%), and lowest false discovery rate (FDR) (0.04), as Table 1 unambiguously demonstrates. Three different baseline models—EKGPS, PGKP, and KGPOKG, respectively—that are used to benchmark and compare the performance of the proposed SGKPS framework were used to assess the performance of the proposed model. Since the suggested SGKPS model is a knowledge graph population model that includes a very potent deep learning classifier to classify the journalism knowledge stack itself, it may be the cause of the model's highest accuracy, recall, precision, F-measure, and lowest FDR percentages. and also, it is coupled with powerful generative AI mechanism for caption generation, the generative AI mechanism is the BLOOM LLM which is being used as a large language model.

Apart from this models like breaking dot allocation and basic initial knowledge graph etc were loaded from the google knowledge graph repository which is a community contributed, community verified knowledge store repository that houses and helps with auxiliary knowledge subsequent to this a journalism knowledge stack comprising of eBooks , glossary and index terms and web 3.0 metadata encompasses stages and stacked knowledge which is subjected to classification using Bi-LSTM model. Further to this the Jiang conratt index , horns index and SimRank add distinct places in the proposed framework and helps in sorting out the irrelevant entities among the most similarly related entities and also set apart the most relevant entities from the relevant entities and this yields the optimal solution and also a strong metaheuristic nature inspired optimization model namely the invasive weed optimization helps in generating the most optimal solution set which is further merged with the existing initial knowledge graphs and subgraphs and is schematized into a knowledge graph repository. Owing to the existence of very strong deep learning model to classify the knowledge stack and the knowledge stack itself which houses the web 3.0 metadata , eBooks , glossary , index terms provides subsequently stitched auxiliary knowledge with high knowledge density and furthermore the semantics oriented computation of relevant entities through SimRank , horns index and jiang conratt index adds to the framework and caption generation using a strong LLM and topic modelling using the Latent Dirichlet Allocation (LDA) helps in providing a very strong ecosystem for hybridization of LLM with deep learning model , semantics driven similarity indexes and a strong metaheuristic optimization model due to these reason proposed model enhances the performance over the other knowledge graph population models and hence is a best in class model for knowledge graph population model.

The reason why the EKGPS which is adding semantic knowledge to the population of online knowledge graphs is not performing as expected although it uses semantic knowledge for knowledge population the model only combines semantic knowledge with NLP based methods from the clustered news events however semantics oriented filtering through quantitative semantic for extremely relevant terms in the ecosystem of much more relevant terms is definitely absent and the presence of strong LLM with a deep learning model and a robust knowledge base made up of the web3.0 metadata and eBooks based dense knowledge entities is definitely absent in the proposed framework hence making the EKGPS framework less efficacious then the suggested framework.

The PGKP, or personal knowledge graph population from user utterances in conversational understanding, is a strategic model that uses relation extraction from database, which is why it is less ideal than the suggested framework and assertion classification is achieved relation detection classification is encompassed and slot filling happens and all of these are really traditional knowledge representation and population techniques combined with the mining of natural language model however semantics oriented interpretation through quantitative semantics analysis and semantic similarity measures **is** definitely absent and also a strong optimization model is absent henceforth in PGKP it only populates the knowledge graph but whether it populates it with the best in class entities is definitely questionable and caption generation using LLM is absent , topic modelling is absent and Because a significant quantity of auxiliary knowledge is lacking, the PGKP model framework performs less well than the suggested model.

The explanation why KGPOKG which stands for Knowledge Graph Population with Out-of-KG entities also doesn't meet expectations to the proposed framework is mainly due to the fact that the KGPOKG model is out of knowledge graph entities description and representation in which identification and modelling of potential knowledge graph entities are induced into the method and the entities will be created using graph to text summarization methods so the ultimate goal is the abstract with summarization methods which is the core strategy which is used in the KGPOKG model however semantic oriented quantitative reasoning through semantic similarity measure and dense knowledge store repository comprising of the web 3.0 metadata as well as the knowledge base's entities in the knowledge store repository along with the system of semantic similarity measures like the jiang conratt index , horns index and SimRank is definitely absent in the KGPOKG framework due these reasons and due to the absence of robust deep learning model and a LDA which is topic modelling framework henceforth KGPOKG model also lags in comparison to the proposed framework.

Table 2.
Precision Percentage Vs No Recommendations

No of Recommendations	EKGPS	PGKP	KGPOKG	Proposed SGKPS
10	92.78	94.74	96.82	98.74
20	91.78	93.71	95.63	97.01
30	90.44	92.74	94.15	96.47
40	89.74	91.08	93.07	95.72
50	88.33	90.82	92.48	94.09

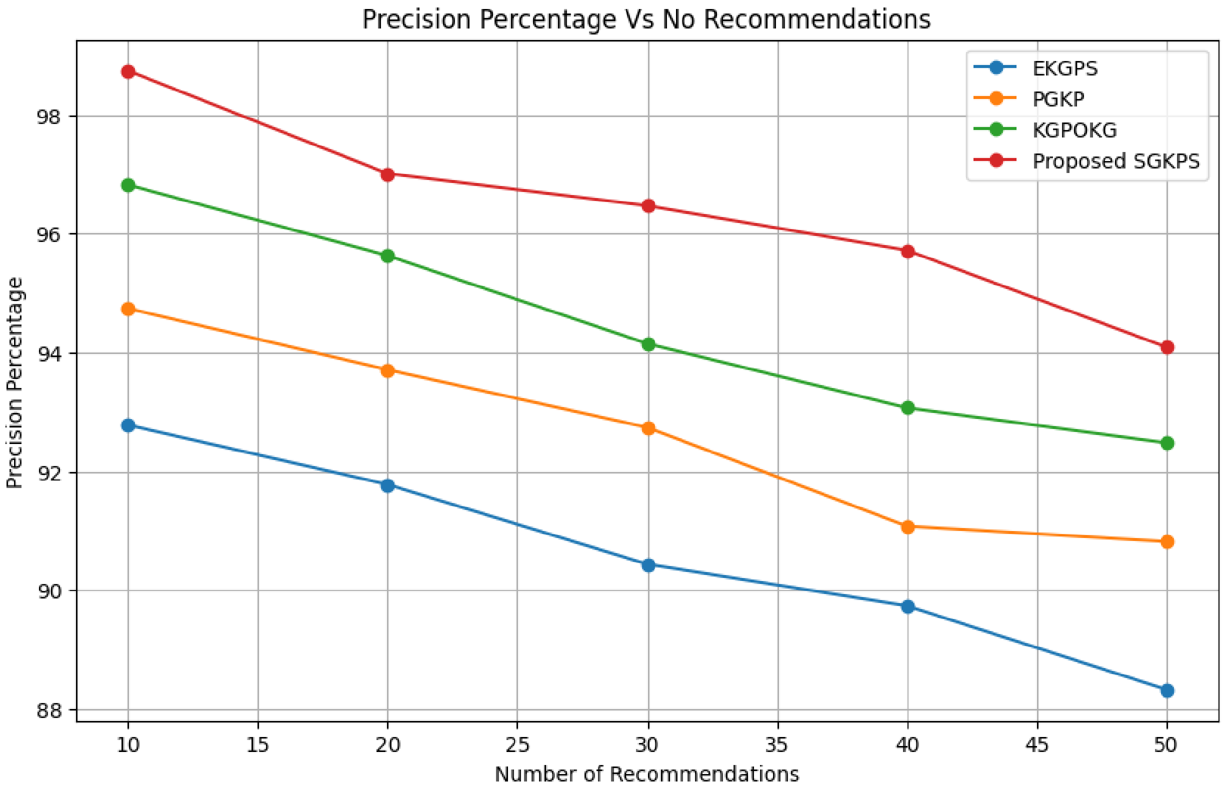

Figure 2: depicts the performance of the various model namely EKGPS, PGKP, KGPOKG and the proposed framework SGKPS

Figure 2 indicates that the suggested SGKPS framework is the highest within the hierarchy followed by EKGPS which is the lowest in the hierarchy , the second last position in the hierarchy is occupied by PGKP, the KGPOKG framework holds the second position within the hierarchy and the reason why the proposed SGKPS is the highest in the hierarchy is that because of following reasons, performance measures, with excellent precision, recall, accuracy, and F-measure of 96.77%, 97.18%, and 96.97%, respectively, and a low False Discovery Rate (FDR) of 0.04. The remarkable outcome can be ascribed to multiple factors: the incorporation of various external knowledge sources, including the Google Knowledge Graph Repository and web 3.0 metadata, the use of the sophisticated BLOOM LLM for superior caption generation, and the integration of a strong deep learning classifier for accurate classification of journalism knowledge.

The Bi-LSTM model, which performs complex entity classification, and efficient semantic similarity metrics like the Jiang-Conratt Index, Horn's Index, and SimRank are other advantages of the SGKPS. By combining it with current data, the model's metaheuristic optimization, namely the Invasive Weed Optimization (IWO), significantly improves knowledge graphs. This combination of deep learning, generative AI, semantic indexing, and optimization techniques creates a robust and efficient framework for knowledge graph population, setting the SGKPS apart from other models. The reason why EKGPS occupies the last position in the list is as follows, several significant limitations cause the EKGPS (Enhancing Online Knowledge Graph Population with Semantic Knowledge) framework to fall short of expectations. Although it incorporates semantic knowledge, its main technique is to combine NLP methods with semantic insights obtained from clustered news events. This method is deficient in sophisticated semantic filtering that finds exceptionally relevant terms among a larger set of terms. Furthermore, the EKGPS model lacks a dense knowledge stack with web 3.0 metadata and eBooks, which are essential for enriching knowledge entities, as well as a strong large language model (LLM) and a deep learning classifier.

Consequently, the EKGPS framework performs less well than the suggested SGKPS, which makes use of an extensive array of sophisticated semantic techniques, deep learning, and rich knowledge sources. The suggested SGKPS is more effective than the PGKP (Personal Knowledge Graph Population from User Utterances in Conversational Understanding) framework because of a number of its shortcomings. Although PGKP employs traditional techniques like slot filling, relation extraction, and assertion classification, it does not provide sophisticated semantic interpretation using semantic similarity metrics and quantitative semantics analysis.

Without a strong optimization model, the framework mainly uses traditional knowledge representation and natural language mining techniques. Furthermore, important features like topic modeling, large language models (LLM) for caption generation, and abundant auxiliary knowledge sources are absent from PGKP. As a result, although PGKP can fill knowledge graphs, it is not as effective as the suggested SGKPS framework at producing high-quality, accurately classified entities. The reason why the KGPOKG (Knowledge Graph Population with Out-of-KG Entities) framework occupies the second position in the hierarchy is that it lacks advanced semantic reasoning, quantitative similarity measures, and a dense knowledge store with web 3.0 metadata, and instead creates entities primarily through graph-to-text summarization. Furthermore, it does not make use of topic modeling frameworks like Latent Dirichlet Allocation (LDA) or a robust deep learning model, nor does it integrate tools like the Jiang-Conratt Index, Horn's Index, or SimRank. Because of these drawbacks, KGPOKG is not as effective as SGKPS. Hence the proposed framework SGKPS performs better that all the other models.

5. Conclusion

This framework has presented the inclusion of a very strong topic modelling framework along with large language models and a metaheuristic based optimization mainly the Invasive weed optimization which makes it a perfect hybrid model for knowledge graph population, the methodology of encompassing strong domain centric knowledge comprising of journalism knowledge in the knowledge stack along with eBooks , glossary ,index terms , web3.0 metadata and it's classification using a very strong deep learning Bi-LSTM based classifier to get classified instances is definitely quite strategic to the proposition. The jiang conratt index , horns index , SimRank differentiates the pipeline by using the differential threshold which is empirically decided on the situation and the context of where it is enrolled in the proposed framework yields the best in class semantics oriented reasoning to quantitative semantic similarity measures, an optimal solution set derivation using a strong Meta-Integer Experimentation Heuristics paints a pathway of the best in class entities in the knowledge graph with a precision of 96.77% , with a F-measure of 96.67% and with a recall 97.18% is being achieved with a very low value of FDR of 0.04 which makes the model best in class for knowledge graph population and specifically for noble domain of sustainable journalism which makes it one of its kind

References

[1] Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12* (pp. 542-557). Springer Berlin Heidelberg.

[2] Shi, B., & Weninger, T. (2018, April). Open-world knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[3] Shi, B., & Weninger, T. (2017, February). Proje: Embedding projection for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

[4] Ji, G., Liu, K., He, S., & Zhao, J. (2016, February). Knowledge graph completion with adaptive sparse transfer matrix. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

[5] Jiang, T., Liu, T., Ge, T., Sha, L., Chang, B., Li, S., & Sui, Z. (2016, December). Towards time-aware knowledge graph completion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1715-1724).

[6] Trouillon, T., Dance, C. R., Gaussier, É., Welbl, J., Riedel, S., & Bouchard, G. (2017). Knowledge graph completion via complex tensor factorization. *Journal of Machine Learning Research*, *18*(130), 1-38.

[7] Gallofré Ocaña, M., & Opdahl, A. L. (2022). Supporting newsrooms with journalistic knowledge graph platforms: Current state and future directions. *Technologies*, *10*(3), 68.

[8] Berven, A., Christensen, O. A., Moldeklev, S., Opdahl, A. L., & Villanger, K. J. (2018). News Hunter: building and mining knowledge graphs for newsroom systems. *NOKOBIT*, *26*, 1-11.

[9] Tchechmedjiev, A., Fafalios, P., Boland, K., Gasquet, M., Zloch, M., Zapilko, B., ... & Todorov, K. (2019). ClaimsKG: A knowledge graph of fact-checked claims. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18* (pp. 309-324). Springer International Publishing.

[10] Ward, S. J. (2019). Journalism ethics. In *The handbook of journalism studies* (pp. 307-323). Routledge.

[11] Fidalgo, J. (2013, June). Journalism is changing–and what about journalism ethics. In *IAMCR 2013 Conference. In: Ethics of Society and Ethics of Communication Working Group. Dublin, Ireland* (pp. 25-29).

[12] Eveland Jr, W. P., & Scheufele, D. A. (2000). Connecting news media use with gaps in knowledge and participation. *Political communication*, *17*(3), 215-237.

[13] Statista Research Department. (2024). *Presence of sustainability in media plans in Europe 2023, by stakeholdertype*[Dataset].Statista.https://www.teosuisse.net/?_=%2Fstudy%2F167869%2Fsustainability-in-advertising-and-marketing-in-europe%2F%23KJWqMdlUlBntOaMGRBz0jpjhec9jFFI%3D

[14] Wessler, Hartmut (2017). Sustainable media events? Production and discursive effects of staged global political media events in the area of climate change. *GESIS Data Archive, Cologne. ZA6768 Data file Version 1.0.0, https://doi.org/10.4232/1.12740.*

[15] Yuriy Syerov. (2023). Social Media Big Dataset for Research, Analytics, Prediction, and Understanding the Global Climate Change Trends. IEEE Dataport. https://dx.doi.org/10.21227/71ms-8v86

[16] C. A, R., M. S, N., & Sai, C. M. (2022). Role of Social Media in Fostering a Sustainable Society through Crowdfunding. In S. K. Arumugam, K. Desai, & B. Toms (Eds.), Sustainable Development - A Diverse Deep-Dive (pp. 82-95).The Native Tribe. Retrieved from https://www.thenativetribe.org/book/cu/sd

[17] Slaček Brlek, S., & Kaluža, J. (2024). *Journalism and new technologies, 2020* [Data set]. Arhiv družboslovnih podatkov = Social Science Data Archives. https://doi.org/10.17898/ADP_NOVTEH20_V1

[18] (2024). Journalism-Methodology [Dashboard]. Retrieved from https://www.workwithdata.com/topic/journalism-methodology

[19] Turkel, E., Saha, A., Owen, C. R., Martin, J. G., & Vasserman, S. (2021). *A method for measuring investigative journalism in local newspapers* (V2) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/HSZ2QL

[20] Fernàndez-Cañellas, D., Marco Rimmek, J., Espadaler, J., Garolera, B., Barja, A., Codina, M., ... & Bou-Balust, E. (2020). Enhancing online knowledge graph population with semantic knowledge. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19* (pp. 183-200). Springer International Publishing.

[21] Li, X., Tur, G., Hakkani-Tür, D., & Li, Q. (2014, December). Personal knowledge graph population from user utterances in conversational understanding. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 224-229). IEEE.

[22] Möller, C. (2022, May). Knowledge Graph Population with Out-of-KG Entities. In *European Semantic Web Conference* (pp. 199-214). Cham: Springer International Publishing.