# Charting the Landscape of Digital Health Towards A Knowledge Graph Approach to News Media Analysis

Vanni Zavarella
Department of Mathematics and Computer Science,
University of Cagliari
Cagliari, Italy
v.zavarella@unica.it

Sergio Consoli
European Commission, Joint Research Centre (DG JRC)
Ispra, Italy
sergio.consoli@ec.europa.eu

Diego Reforgiato Recupero
Department of Mathematics and Computer Science,
University of Cagliari
Cagliari, Italy
diego.reforgiato@unica.it

Gianni Fenu
Department of Mathematics and Computer Science,
University of Cagliari
Cagliari, Italy
fenu@unica.it

## ABSTRACT

In this paper, we present our currently on-going work on a method for analyzing digital health transformation in our society by constructing a Knowledge Graph from a large corpus of 7.8 million English news articles, dating from 1987 through 2023. We firstly sampled around 95k articles relevant to the Digital Health topic by training and deploying a Deep Learning binary classifier via fine-tuning BERT. Successively, by deploying NLP techniques, we extracted triples from the identified articles to form a Digital Health News Knowledge Graph, which consists of 431k distinct triples connecting 186k entities through 1866 relations. The constructed Knowledge Graph provides insights into the evolution of Digital Health in news media and serves as a resource for further research in the field. The analysis that we have carried out reveals significant trends in Digital Health as reflected in the news, with notable peaks coinciding with key events like the COVID-19 pandemic.

**ACM Reference Format:**

Vanni Zavarella, Sergio Consoli, Diego Reforgiato Recupero, and Gianni Fenu. 2024. Charting the Landscape of Digital Health: Towards A Knowledge Graph Approach to News Media Analysis. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3631700.3665237

## 1 INTRODUCTION

In recent years, traditional healthcare systems have witnessed a pervasive digital transformation process characterized by the integration of digital tools and innovative solutions encompassing electronic health records, telemedicine, artificial intelligence (AI), and wearable devices. By analyzing news articles, reports, and commentaries, one can glean insights into emerging trends, technological innovations, policy developments, public perceptions, and the overall evolving discourse surrounding digital health (DH) initiatives.

While a plethora of news monitoring tools already exist, offering diverse capabilities for news analysis[1], generally they feature limited capability on extracting connections and supporting advanced queries about the entities, tags and keywords they identify in the news texts or cover a limited number of target domains [2, 20]. To address this limitations, researchers have proposed diverse methodologies to construct structured, interlinked, and machine-readable data frameworks tailored for news analysis [16, 22]. A number of these frameworks utilize semantic technologies, including knowledge graphs (KGs), which are expansive networks comprising entities and relationships [9, 18] that can also undergo automatic expansion via link prediction techniques [5, 11, 15]. Information extraction pipelines have been successfully applied to the creation of extensive and high-quality KGs from text data in several domains [1, 3, 6–8, 12, 13, 21, 23, 24]. Nonetheless, constructing a large-scale, cohesive, and semantically robust representation of news content sourced from millions of news articles still poses various challenges concerning the capacity of a system to consolidate multiple instances of the same entity and grounding them onto existing repositories, as well as mapping verbal predicates to generalized semantic relations.

We present an enhanced, scalable information extraction architecture under current development, which is designed to generate a large knowledge graph of predictive triples concerning key entities in the Digital Health domain from news articles. Additionally, we report on an ongoing endeavor of building a series of data analytics visualizations on top of the generated graph and show how these can potentially support insight gathering about trends and dynamics of digital transformation in the health sector over time and in different geographical regions.


*UMAP Adjunct '24, July 01–04, 2024, Cagliari, Italy*

ACM ISBN 979-8-4007-0466-6/24/07
https://doi.org/10.1145/3631700.3665237

[1]Examples include: Europe Media Monitor (emm.newsbrief.eu), Brandwatch (brandwatch.com), Repustate (repustate.com), Cision Communication Cloud (cision.com), SentiOne (sentione.com) and Meltwater (meltwater.com).

| | P | R | F-Score | Acc. |
|---|---|---|---|---|
| Cross-validation | 99.40 | 97.40 | 98.39 | 98.59 |
| 200 documents Test Set | 98.9 | 93.0 | 95.8 | 96.0 |

**Table 1: Evaluation of the binary Health Tech topic classifier.**

## 2 PIPELINE

### 2.1 Dataset

Our source dataset comprises title and full text of around 7.8M English language news articles from the Dow Jones Data, News and Analytics (DNA) platform[2] ranging from September, 1987 through December, 2023. We used the curated Region codes metadata provided by DNA and merged them into a two-valued ('EU','US') attribute in order to differentiate our analysis for the European and US contexts.

### 2.2 Topic Classifier

In order to sample news articles about Digital Health technologies we trained a Deep Learning binary classifier on the concatenated title and full text of news items from DNA[3] and several RSS feeds from specialized news outlets in health tech[4], where positive training instances are digital health-related documents and negative instances are documents belonging to unrelated topics. Namely, we fine-tuned BERT[5] on a near-balanced small training set of 9097 news items (4495 positive - 4602 negative instances).

Among the positive instances, 4187 were articles from health tech news outlets, while 308 were DNA articles filtered by general Health-related Subject codes and then manually tagged as concerning Digital Health. Out of the 4602 negative instances, 3000 were DNA items sampled from 'negative' Subject tags such as *gcat* (Political/General News) or *mcat* (Commodity/Financial Market News) and 1602 were articles scraped from 'negative' topic feeds of technology news outlets[6].

Fine-tuning was performed using 10-fold stratified cross-validation with 80-20% data splits and Binary Cross Entropy as Loss function, training for 10 epochs with Early Stopping on 1 epoch of non-increasing Accuracy score. We mitigate the model over-fitting on the relatively small training set by keeping a small number of trainable parameters (4.3M) and adding a dropout regularization layer in the training phase (0.2 dropout rate).

In Table 1 we show the average cross-validation performance of the classifier on the provided dataset. Additionally, after training the classifier on the entire training set, we evaluated it on a separate dataset comprising 100 negative and 100 positive instances sampled from DNA using Subject codes. We report the classifier's performances on the test set as well in the second row in the Table.

The models achieved 98.39% on the training set and 95.8% F-Score on the hold-out test set from DNA (98.9% Precision, 93.0% Recall), which shows that, despite missing a significant number of true positives, it allows to sample a mostly relevant subset of health tech articles from the DNA multi-domain corpus[7]. From the original full DNA dataset we then filtered out around 95k health tech articles (1.2% sample) by deploying the trained classifier.

### 2.3 Triple Extraction and Refining

In the triple extraction phase, the DH news items filtered by the classifier have been passed to a spaCy NLP pipeline[8]. Then we applied several optimized procedures for identifying candidate nominal entities and forming predicative triples, namely:

- An **entity extraction module** that identifies local nominal phrases and extends them with non-recursive attached prepositional phrases and Quantity-type entities (MONEY, PERCENT, QUANTITY, CARDINAL), yielding a set $E$ $\{e_0, ..., e_n\}$ of non-unified, candidate entity phrases;
- A **relation extraction module** selecting all the shortest paths of the dependency tree between every pair of entities $(e_m, e_n)$ containing a verb and matching any of a shortlist of expert-validated patterns[9], yielding a set of verbal relations $V$ $v_0, ..., v_k$ and a set of triples $S$ $s_0, ..., s_k$ in the form $e_m, v, e_n >$ where $v \in V$ and $e \in E$.

The pipeline goes on generalizing from the surface form triples in set $S$ to a smaller set $T$ $t_0, ..., t_h$ of triples in the form of $_m, r, _n >$ where each $_i \in E$ is the result of entity merging (see modules below) and $r$ is a label drawn from a generalized relation vocabulary $R$. In particular:

- The **entity refining module** first cleans up and normalizes candidate entities, then the normalized entities are merged by leveraging their linking to DBpedia entries via the DBpedia Spotlight service[10]. The merging process has been formalized with a relation `owl:sameAs` in the output knowledge graph.
- The **relation refining module** is used to identify the most suitable predicate label $r$ for each relation verb $v$ in a triple $e_m, v, e_n >$ and to establish the mapping from $v$ to $r$ in the resulting triple. In detail:
  - We retrieve the 300-dimensional word embeddings representation learned with GloVe [19] for each single or multi-token verb predicate;
  - We perform UMAP [10, 14] non-linear dimensionality reduction[11] and apply HDBSCAN clustering [4] on the relation vectors, optimizing via grid search across the hyperparameters of both algorithms. We use the score: $S$ $silhouette$ $clustered$, combining the mean silhouette coefficient over all clustered instances and the fraction $clustered$ of instances of $X$ actually clustered by HDBSCAN [4];

[2]https://professional.dowjones.com/developer-platform/. We discarded items with missing Title and short articles with text body character length lower than 300.
[3]We removed URLs, DNA-specific and news outlet-specific tokens (e.g., 'Reuters', 'Reuters Limited', 'techcrunch') and we truncated training instances to 1000 characters, in order to remove correlation between the topic and text length features of the article sources.
[4]E.g., healthtechdigital.com, techcrunch.com/tag/healthtech/feed/ digitalhealth.net/news/ and others.
[5]https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-2_H-128_A-2/1
[6]For example, https://techcrunch.com/tag/security/.
[7]The trained model is made publicly available at the project repository https://github.com/zavavan/dtm_kg/blob/master/resources/bert_fine_tuned_healthTech-20240516T132123Z-001.zip
[8]Including sentence splitting, POS tagging, Dependency Parsing, (https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.6.0).
[9]https://github.com/zavavan/dtm_kg/blob/master/resources/paths.txt
[10]https://www.DBpedia-spotlight.org/
[11]https://umap-learn.readthedocs.io/en/latest/parameters.html

– Finally, for each relation verb $v$ in the dataset, we replace it with a predicate label $r$ consisting of the most frequent lemma among the 'exemplars' relations returned by HDBSCAN for the cluster of $v$[12], otherwise we map it to itself if $v$ is an outlier.

The chosen hyper-parameter configuration for the combination of UMAP and HDBSCAN was sub-optimal and represents a balance between higher $S$ score and lower number of generated clusters (allowing higher generalization over relation predicates). For UMAP, we used *cosine* distance metric over *normalized* vectors, applying 0.0 *min_dist*, 2 *n_neighbors* and 2 *n_components*. As for HDBSCAN, we set *min_cluster_size* to 5, with *min_samples* and *cluster_selection_epsilon* set to 0.

## 3 KNOWLEDGE GRAPHS AND DATA ANALYTICS

Using the described pipeline we generated a Digital Health News Knowledge Graph (DHNKG), comprising roughly 431k distinct (non-reified) triples, connecting 186k unique entities via a total of 1866 generalized relations. We then reified each extracted claim of DHNKG into instances of the `dhn ont:Statement` class in the corresponding ontology (`dhn ont` namespace prefix), with each `dhn ont:Statement` representing a specific assertion derived from a collection of news items.

Out of the overall set of unique DHNKG entities, around 8% have been linked to DBpedia entries using 14975 `owl:sameAs` and 33345 `skos:related` predicates, indicating entity equality and relatedness, respectively. Overall, 23.8% of all triples had either subject or object entities linked to DBpedia. DHNKG's `owl:sameAs`-linked entities inherit DBpedia entity typization.

While the DHNKG graph is fully automatically generated, the deployed methodology has been manually evaluated[25] and proved to have reliable Precision. Therefore, we made the KG available as a public resource via data access endpoints. We set up a Virtuoso SPARQL endpoint (https://api-vast.jrc.service.ec.europa.eu/sparql/) where DHNKG can be queried and analytical information on target entities, attributes, and relations can be retrieved in some structured data formats[13].

As an example, one may start up by exploring the KG (the internal name 'DHNEWS_KG') via SPARQL queries like the one in Figure 1, which uses regex constraints to retrieve the graph's information about entities matching a target name of interest, like 'biogen', the American multinational biotechnology company known as Biogen Inc.

```
SELECT ?s ?p ?o
FROM <DHNEWS_KG>
WHERE { ?s ?p ?o .FILTER regex(str(?s),  biogen ) }.
```

**Figure 1: Explorative search query on DHNKG.**

The result set will include a triple in which an internal graph resource (specifically http://dhnewskg.org/dhnewskg/resource/biogen) is linked to the DBpedia entry for Biogen Inc., using the `owl:sameAs` predicate. At this point, a more fine-grained query like the one in Figure 2 can be run to return all the 480 statements in 'DHNEWS_KG' having the target entity `dhnewskg:biogen` as object. Finally, for each result statement, a link to its corresponding URL in a Virtuoso Faceted Browser endpoint is returned, allowing further navigation.

```
PREFIX dhnewskg: <http://dhnewskg.org/dhnewskg/resource/>
PREFIX dhnewskg-ont: <http://dhnewskg.org/dhnewskg/ontology#>
SELECT ?statement
FROM <DHNEWS_KG>
WHERE { ?statement a rdf:Statement .
?statement rdf:subject dhnewskg:biogen . }
```

**Figure 2: Targeted query returning all DHNKG statements with the graph entity `dhnewskg biogen` as `rdf subject`.**

By deploying the presented pipeline, we also generated an aggregated analysis of the news reflection on the Digital Health dynamics over the years[14].

In Fig. 3 below we measure the presence of the Digital Health discourse in the news over time by plotting the month-sampled time series of the ratio of news articles about Digital Health over the total of English language DNA news mentioning EU and US, respectively. While representing an underestimation in absolute terms (due to the classifier's low recall), the plot describes significant patterns in the raising and consolidation of the impact of DH technologies, with some remarkable differences between the EU and US contexts.

The DH topic is minimally represented in the news (around 1%) until a first breakthrough recorded on 2005 in both regions, with the rise in exposure likely due to advancements in technology, particularly to the implementation and progressive adoption of Health Information Exchange (HIE) networks to facilitate the electronic sharing of patient data among healthcare providers. A second peak took place in 2017 in the US, while EU news showed a rather steady growth of the topic's momentum from 2015 on. Finally, a third major spike occurred worldwide in 2020, arguably related to the global effort towards applying innovative technologies in the fight against COVID-19.

## 4 CONCLUSIONS

The paper describes the ongoing development of a Digital Health News Knowledge Graph (DHNKG) using a dataset of 7.8 million English news articles from 1987 to 2023. The work involves training a deep learning classifier to identify relevant articles and employing NLP techniques to extract knowledge triples. The resulting DHNKG offers insights into digital health trends, particularly highlighting the impact of events like the COVID-19 pandemic. The analysis also compares digital health coverage in the US and EU. We are currently working on expanding the visualization and analytics

[12]https://hdbscan.readthedocs.io/en/latest/api.html

[13]Currently the access is password protected, with credentials available upon request to authors.

[14]This is a preliminary result of an ongoing endeavour to build data analytics visualizations on top of the information in DHNKG, in order to support DH trend analysis and insight gathering.

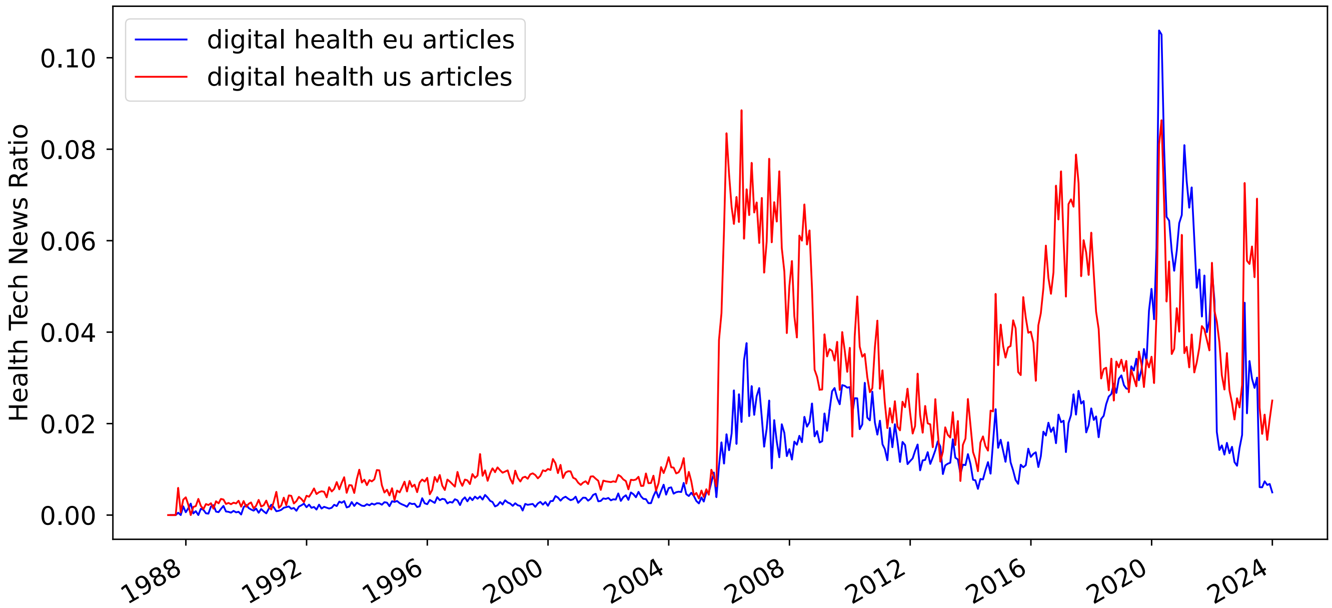


**Figure 3: Monthly time series of the ratio of news articles about Digital Health over the total of DNA news about Europe and the US respectively.**

tools to further exploit the DHNKG and extract useful hidden insights. A current limitation of the KG extraction method is that the entity and relation extraction processes are not backed by underlying ontology specifications. We are currently working on an enhanced version of the pipeline that builds upon the current entity and relation spans and further classifies them into domain-specific categories, leveraging fine-tuning of contextual word embedding representations from Large Language Models [17].

## ACKNOWLEDGMENTS

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

## REFERENCES

[1] Harith Alani, Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. *Semantic Web* 8, 6 (2017), 873–893. https://doi.org/10.3233/SW-160240

[2] Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017. On the Creation of a Security-Related Event Corpus. In *Proceedings of the Events and Stories in the News Workshop*, Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard Hovy, Teruko Mitamura, and David Caswell (Eds.). Association for Computational Linguistics, Vancouver, Canada, 59–65. https://doi.org/10.18653/v1/W17-2709

[3] Carlos Barbosa, Lucas Félix, Vinícius Vieira, and Carolina Xavier. 2019. Sara - A Semi-Automatic Framework for Social Network Analysis. In *Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web* (Florianópolis). SBC, Porto Alegre, RS, Brasil, 59–62. https://doi.org/10.5753/webmedia_estendido.2019.8137

[4] Fatima Batool and Christian Hennig. 2021. Clustering with the Average Silhouette Width. *Computational Statistics and Data Analysis* 158 (2021). https://doi.org/10.1016/j.csda.2021.107190

[5] Agustín Borrego, Danilo Dessi, Inma Hernández, Francesco Osborne, Diego Reforgiato Recupero, David Ruiz, Davide Buscaldi, and Enrico Motta. 2022. Completing Scientific Facts in Knowledge Graphs of Research Concepts. *IEEE Access* 10 (2022), 125867–125880.

[6] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, and L. Secchi. 2023. Data-Driven Methodology for Knowledge Graph Generation Within the Tourism Domain. *IEEE Access* 11 (2023), 67567–67599. https://doi.org/10.1109/ACCESS.2023.3292153

[7] Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2021. Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. *Future Generation Computer Systems* 116 (2021), 253–264. https://doi.org/10.1016/j.future.2020.10.026

[8] Jens Dörpinghaus, Sonja Klante, Martin Christian, Christof Meigen, and Carsten Düing. 2022. From social networks to knowledge graphs: A plea for interdisciplinary approaches. *Social Sciences Humanities Open* 6, 1 (2022), 100337. https://doi.org/10.1016/j.ssaho.2022.100337

[9] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016 (CEUR Workshop Proceedings, Vol. 1695)*, Michael Martin, Martí Cuquet, and Erwin Folmer (Eds.). CEUR-WS.org, 1–4. https://ceur-ws.org/Vol-1695/paper4.pdf

[10] David Freire-Obregón, Kevin Rosales-Santana, Pedro A. Marín-Reyes, Adrian Penate-Sanchez, Javier Lorenzo-Navarro, and Modesto Castrillón-Santana. 2021. Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment. *Pattern Recognition Letters* 149 (2021), 179–184. https://doi.org/10.1016/j.patrec.2021.06.014

[11] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications* 553 (2020), 124289.

[12] T. Man, A. Vodyaho, D.I. Ignatov, I. Kulikov, and N. Zhukova. 2023. Synthesis of multilevel knowledge graphs: Methods and technologies for dynamic networks. *Engineering Applications of Artificial Intelligence* 123 (2023). https://doi.org/10.1016/j.engappai.2023.106244

[13] Jose L. Martinez-Rodriguez, Ivan Lopez-Arevalo, and Ana B. Rios-Alvarado. 2018. OpenIE-based approach for Knowledge Graph construction from text. *Expert Systems with Applications* 113 (2018), 339–355. https://doi.org/10.1016/j.eswa.2018.07.017

[14] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML]

[15] Mojtaba Nayyeri, Gokce Muge Cil, Sahar Vahdati, Francesco Osborne, Mahfuzur Rahman, Simone Angioni, Angelo Salatino, Diego Reforgiato Recupero, Nadezhda Vassilyeva, Enrico Motta, et al. 2021. Trans4E: Link prediction on scholarly knowledge graphs. *Neurocomputing* 461 (2021), 530–542. https://doi.org/10.1016/j.neucom.2021.02.100

[16] Andreas L. Opdahl, Tareq Al-Moslmi, Duc-Tien Dang-Nguyen, Marc Gallofré Ocaña, Bjørnar Tessem, and Csaba Veres. 2022. Semantic Knowledge Graphs for the News: A Review. *ACM Comput. Surv.* 55, 7, Article 140 (dec 2022), 38 pages. https://doi.org/10.1145/3543508

[17] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. arXiv:2306.08302 [cs.CL]

[18] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. 2023. Knowledge graphs: opportunities and challenges. *Artificial Intelligence Review* (2023), 1–32.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[20] Ralf Steinberger, Bruno Pouliquen, and Erik van der Goot. 2013. An introduction to the Europe Media Monitor family of applications. arXiv:1309.5290 [cs.CL]

[21] G. Tamašauskaite and P. Groth. 2023. Defining a Knowledge Graph Development Process Through a Systematic Review. *ACM Transactions on Software Engineering and Methodology* 32, 1 (2023). https://doi.org/10.1145/3522586

[22] Fiona Anting Tan, Debdeep Paul, Sahim Yamaura, Miura Koji, and See-Kiong Ng. 2023. Constructing and Interpreting Causal Knowledge Graphs from News. arXiv:2305.09359 [cs.CL]

[23] Y. Xiao, C. Li, and M. Thürer. 2023. A patent recommendation method based on KG representation learning. *Engineering Applications of Artificial Intelligence* 126 (2023). https://doi.org/10.1016/j.engappai.2023.106722

[24] S. Yu, C. Peng, C. Xu, C. Zhang, and F. Xia. 2023. Web of Conferences: A Conference Knowledge Graph. In *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. 1172–1175. https://doi.org/10.1145/3539597.3573024

[25] Vanni Zavarella, Diego Reforgiato Recupero, Sergio Consoli, Gianni Fenu, Simone Angioni, Davide Buscaldi, Danilo Dessí, and Francesco Osborne. inpress. Knowledge Graphs for Digital Transformation Monitoring in Social Media. In *3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG), co-located with ESWC 2024*, CEUR-WS (Ed.). Hersonissos (Crete), Greece.