Heliyon 10 (2024) e32479

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Triplétoile: Extraction of knowledge from microblogging text

Vanni Zavarella [a], Sergio Consoli [c], Diego Reforgiato Recupero [a,*], Gianni Fenu [a], Simone Angioni [a], Davide Buscaldi [b], Danilo Dessí [d], Francesco Osborne [e,f]

[a] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari, 09121, Italy
[b] Laboratoire d'Informatique de Paris Nord, Sorbonne Paris Nord University, 99 Av. Jean Baptiste Clement, 93430 Villetaneuse, Paris, France
[c] European Commission, Joint Research Centre (DG JRC), Via E. Fermi 2749, Ispra (VA), 21027, Italy
[d] Knowledge Technologies for Social Sciences Department, GESIS Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany
[e] Knowledge Media Institute, The Open University, Walton Hall, Berrill Building, Milton Keynes, 50667, UK
[f] Department of Business and Law, University of Milano Bicocca, Via Bicocca degli Arcimboldi 8, Milano, 20100, Italy

## ARTICLE INFO

Dataset link: https://data.jrc.ec.europa.eu/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139

*Keywords:*
Information extraction
Knowledge graphs
Social media analysis
Named entity recognition
Hierarchical clustering
Word embeddings

## ABSTRACT

Numerous methods and pipelines have recently emerged for the automatic extraction of knowledge graphs from documents such as scienti c publications and patents. However, adapting these methods to incorporate alternative text sources like micro-blogging posts and news has proven challenging as they struggle to model open-domain entities and relations, typically found in these sources. In this paper, we propose an enhanced information extraction pipeline tailored to the extraction of a knowledge graph comprising open-domain entities from micro-blogging posts on social media platforms. Our pipeline leverages dependency parsing and classi es entity relations in an unsupervised manner through hierarchical clustering over word embeddings. We provide a use case on extracting semantic triples from a corpus of 100 thousand tweets about digital transformation and publicly release the generated knowledge graph. On the same dataset, we conduct two experimental evaluations, showing that the system produces triples with precision over 95% and outperforms similar pipelines of around 5% in terms of precision, while generating a comparatively higher number of triples.

## 1. Introduction

Examining, connecting, and understanding content sourced from microblogging platforms holds signi cant importance in pinpointing trends, and grasping the intricacies of events and individuals' influence. However, this endeavor is particularly demanding due to the Internet's diverse array of social platforms, each marked by its own distinctiveness, and potentially featuring natural language text in varying formats, structures, and lengths.

Social analysts and various stakeholders commonly navigate this intricate realm through the utilization of social media platforms such as Hootsuite,[1] Brandwatch,[2] Talkwalker,[3] Sprout Social.[4] However, these platforms are constrained to basic queries and merely

* Corresponding author.
*E-mail address:* diego.reforgiato@unica.it (D. Reforgiato Recupero).

[1] https://www.hootsuite.com/.
[2] https://www.brandwatch.com/.
[3] https://www.talkwalker.com/.
[4] https://sproutsocial.com/.

https://doi.org/10.1016/j.heliyon.2024.e32479
Received 9 February 2024; Received in revised form 4 June 2024; Accepted 4 June 2024

Available online 10 June 2024

provide a list of pertinent documents that require manual analysis. These limitations represent a notable impediment to the flow of knowledge within the social media analysis process.

The main problem lies in the fact that existing systems do not possess an adequate depiction of the nuanced social media dynamics, thereby rendering them incapable of facilitating advanced queries regarding the entities mentioned in the posts. This limitation hinders the ability to discern potential trends, gauge the influence of events or individuals, and understand their relationships.

Consequently, the research community has put forth numerous proposals aimed at generating organized, interconnected, and machine-readable data frameworks of social analysis knowledge found within text from microblogging platforms [1–3]. Typically, this resulting representation employs Semantic Web technologies, such as ontologies and knowledge graphs. In computer science, ontologies are de ned as *explicit speci cations of a conceptualization* [4] and serve to formalize the conceptual structure of a speci c domain by delineating the categories of entities and their interrelationships. Typically, ontologies are encoded using the Web Ontology Language (OWL) and are considered the foundational pillars of the Semantic Web [5].

Knowledge graphs (KGs) are extensive networks comprising entities and relationships, imparting machine-readable and comprehensible information pertaining to a speci c domain, adhering to a formal semantic structure [6]. In recent years, KGs have become increasingly recognized for their ability to organize structured data in a semantically signi cant way, allowing them to effectively support various AI systems [7]. The relationship between two entities is typically formalized as a triple in the format of `<subject, predicate, object>`, such as `<digital transformation, revolutionize, industry>`. The structure of a KG is commonly outlined in a domain ontology. Large-scale KGs are usually generated through a semi-automated process, utilizing both structured and unstructured data. Some prominent examples include DBpedia [8],[5] Google Knowledge Graph,[6] BabelNet,[7] and YAGO.[8] Furthermore, knowledge graphs can undergo automatic re nement through link prediction techniques, which are designed to identify additional relationships among domain entities [9,10]. For instance, these approaches can facilitate the formulation of novel scienti c hypotheses by linking known entities in new ways [11].

Creating extensive and high-quality knowledge graphs from social media is a current open problem that has already been addressed by researchers [2]. Existing solutions either depend on systems that aid social media experts in structuring their knowledge or rely on information extraction pipelines [12,2,13,14]. The rst category of solutions suffer from scalability problems. Information extraction techniques have the potential for scalability but often struggle to generate outputs of sufficient quality for practical applications. Speci cally, present approaches for extracting entities and relationships from social media texts typically focus on speci c domains [2] without giving much importance to the preprocessing and linking operations of entities and relations, and their grounding. However, crafting a large-scale, coherent, and semantically sound representation of social media texts drawn from millions of posts is an entirely distinct challenge. Consequently, merely employing existing methods for entity and relationship extraction on an extensive collection of texts would yield a highly noisy outcome [15]. Therefore, several challenges should be addressed, including:

- Integrating the extracted information from various posts into a cohesive representation;
- Evaluating the validity of the resultant triples;
- De ning a flexible ontological framework to formalize a range of statements originating from social media text.

Recent advancements in natural language processing have given rise to sophisticated Large Language Models (LLMs), including Mistral [16], LLaMa 3 [17,18], Gemma [19], and GPT 4.0 [20], among others. These models exhibit the ability to generate coherent and articulate responses to user queries and perform various tasks such as text classi cation and information extraction. Despite their capabilities, concerns have emerged regarding the accuracy and reliability of the content they generate. A notable issue is their tendency to produce “hallucinations”, i.e., responses that lack grounding in factual knowledge [21]. To address this, researchers are exploring the integration of LLMs with structured knowledge representations [22,23]. This integration aims to enhance the accuracy and transparency of LLMs by linking them to reliable sources and enabling the tracking of claim origins. KGs are increasingly vital in this context and are well-suited to complement LLMs [7].

Similar challenges for KG construction have already been addressed within the scholarly domain in [15] where the authors introduced an information extraction approach merging data from various tools based on a domain ontology and allowing in this way for the creation of expansive KGs. This pioneering approach has served as a source of inspiration for subsequent research in the eld [24–28]. However, this work also encountered several limitations: i) the entity extraction modules did not capitalize on the expert knowledge gained from analyzing the resulting knowledge graphs; ii) limited capability to unify multiple instances of the same entity; iii) a shallow and manual approach for mapping verbal predicates to semantic relations; iv) a constrained methodology for evaluating triple validity, relying on a basic multilayer perceptron classi er.

Therefore, in this paper, we present *Triplétoile*, an enhanced information extraction architecture designed to overcome the aforementioned limitations. This innovative solution demonstrates the capability to extract entities from social media text and identify different instances of them. Additionally, it facilitates the extraction of various relationships among these entities by using hierarchical clustering, word embeddings, and dimensionality reduction techniques. Furthermore, we present a use case consisting of the

[5] https://www.dbpedia.org/.
[6] https://developers.google.com/knowledge-graph.
[7] https://babelnet.org/.
[8] https://yago-knowledge.org/.

application of the proposed architecture to a subset of around 100k tweets extracted from the Twitter platform[9] from 2022 and concerning the digital transformation domain.

We conducted an assessment of Triplétoile by comparing it to several alternative solutions using a benchmark dataset consisting of 500 triples. As it will be shown next, our results reveal that Triplétoile outperforms the alternatives in terms of accuracy, while at the same time generating a relatively higher number of triples.

In brief, the main research contributions of this paper encompass the following:

- We design a general, scalable, and flexible architecture for triple extraction from social media text.
- We provide a use case on Twitter where we extracted approximately 100k tweets related to digital transformation in 2022 and subsequently released a corresponding knowledge graph comprising 22,270 statements.
- On the proposed use case, we perform a formal assessment of Triplétoile in terms of precision and a comparative evaluation with respect to alternative methods.
- We publicly release the resulting triple store as a dataset within the Joint Research Centre Data Catalogue,[10] as well as within the European Data portal,[11] the official data repository of the European Commission.

The remainder of this paper is organized as follows. The related work is illustrated in Section 2. The proposed architecture is depicted in Section 3. The use case and comprehensive analysis of the extracted triples are described in Section 4. The evaluation we have carried out, including the comparisons against state-of-the-art tools, is detailed in Section 5. Finally, conclusions and future works where we are headed are reported in Section 6.

## 2. Related work

The term “knowledge graph” was first coined in 1972, but it was not until 2012 that it gained widespread recognition after Google’s announcement[12] of the Google Knowledge Graph [29]. This event also sparked the growth of knowledge graph development and usage in the industry [30,31]. A knowledge graph is a graph of data that is designed to capture and communicate knowledge about the real world. Its nodes represent entities of interest and its edges represent the relationships between these entities [7,32].

Creating, maintaining, and refining knowledge graphs requires the use of an array of techniques for information extraction, entity selection and linking, relation extraction, and ontology engineering [33,5,34]. Numerous scholarly articles delve into the methodologies for generating knowledge graphs across different domains and under various constraints. [15,35]. Notably, Sequeda et al. [36] introduced a unique pay-as-you-go approach to overcome the challenges associated with understanding complex database schemas, providing a use case from a large company.

The extraction of knowledge graphs from web sources to answer questions related to social networks [1], such as Twitter or Facebook, has been widely discussed in literature [37,38,2]. He et al. [3] described how to build knowledge graphs for social networks by developing deep Natural Language Processing models, and holistic optimization of knowledge graphs and the social network. While authors in [39] have already acknowledged the overlap between social networks and knowledge graphs, the current research has poorly exploited this overlap so far. A number of information extraction pipelines have been proposed to create high-quality knowledge graphs within the social network analysis domain (see for example [12,2,13,14]). While information extraction techniques have the potential for scalability, they often struggle to produce outputs of sufficient quality for practical applications. Specifically, current approaches for extracting entities and relationships from social analysis texts typically focus on specific domains [2], neglecting the significance of preprocessing, linking operations, entity grounding, and the creation of a large-scale, coherent, and semantically robust representation of social network analysis texts drawn from millions of posts [15].

Haslhofer et al. [40] have emphasized the importance of connected knowledge graphs and discovery, whereas Hyvönen and Rantala [41] have highlighted the significance of new relationships extracted from the original dataset. In recent years there has been also an increasing research focus on ontologies and interoperable data [42]. In particular, Dessì et al. [15] have proposed an information extraction method that combines data from different tools using a domain ontology, enabling the creation of expansive knowledge graphs. This first approach has been a source of inspiration for further research in the field [24–28].

Implicitly, a significant amount of research has already utilized knowledge graphs. This involves combining actors, persons, and additional information such as locations using linked data [43]. While the practice of using external data linked to a network is still prevalent, it is also possible to define different types of nodes. Tamašauskaite and Groth [27] conducted a systematic review of the process of knowledge graph creation. The review methodology aimed to collect and describe the various steps involved in this process, such as data identification, construction of the knowledge graph ontology, knowledge extraction, analysis of the extracted knowledge, knowledge graph creation, and maintenance. The last step, maintenance, entails periodic updates and edits to keep the knowledge graph up to date. In their review, the authors offer suggestions, best practices, and tools that support the creation and maintenance of knowledge graphs.

In this paper, we propose a methodology specifically tailored for micro-blogging text which overcomes several limitations of existing approaches in the field. Specifically: i) our method identifies entities among the text and has a high ability to unify different

[9] https://twitter.com/.

[10] https://data.jrc.ec.europa.eu/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139.

[11] https://data.europa.eu/88u/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139.

[12] https://blog.google/products/search/introducing-knowledge-graph-things-not/.

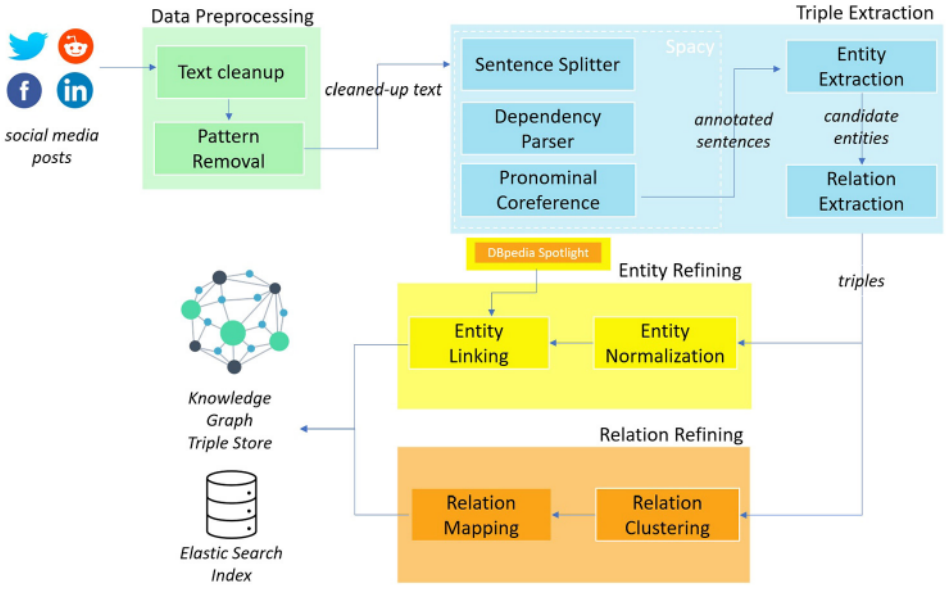



**Fig. 1.** Flowchart of the pipeline for generating a knowledge graph from micro-blogging text data.

instances of the same entity; ii) the designed entity extraction modules make use of the information acquired from analyzing the obtained knowledge graph; iii) we perform entity coreference resolution by applying pronoun anaphora resolution and a set of heuristics to normalize the identi ed entities; iv) the method recognizes relationships among the identi ed entities and comes up with an automated approach for mapping verbal predicates to semantic relations; v) nally, a robust methodology to evaluate the validity of the produced triples is adopted. To the best of our knowledge, a methodology speci cally tailored for micro-blogging text embracing all these features is the rst of its kind.

## 3. The proposed architecture

Fig. 1 shows the workflow of the pipeline that we propose in this paper.
The main blocks of the architecture include:

- *Data Preprocessing*, a step responsible for the normalization of the micro-blogging text in order to make it processable by the downstream text analysis modules;
- *Triple Extraction*, the block comprising core modules applying text processing libraries and models for the extraction of entity-relation triples;
- *Entity Re ning*, a block responsible for the cleaning and generalization of entity mentions to canonical forms, in view of subsequent entity merging;
- *Relation Clustering*, in which relation instance verbal forms are mapped to canonical forms, computed as a representative of the relation cluster they belong to.

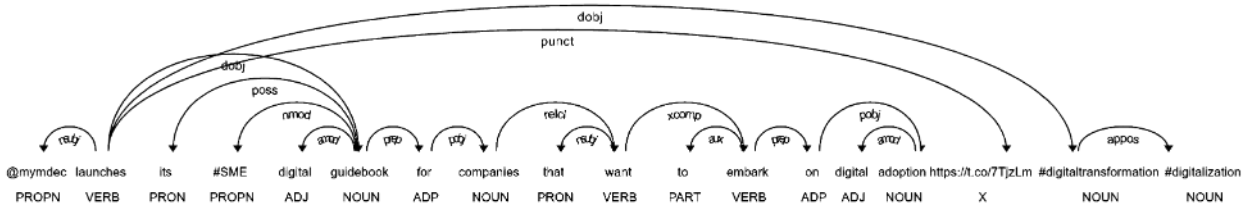
The nal output of the pipeline is a knowledge graph of generalized triples annotated with references to the micro-blogging text items they were matched in.

The following subsections describe in more detail the individual components of the pipeline across the four main blocks and how they are applied.

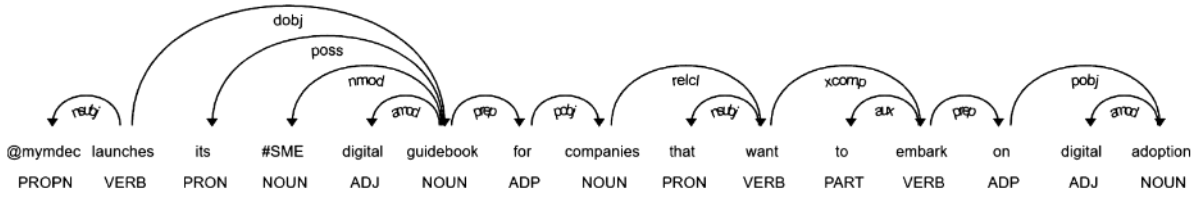
### 3.1. Data preprocessing

Twitter status updates (tweets) are short micro-blogging posts of a maximum of 280 characters: their informal (often plainly ungrammatical) genre and the abundance of platform-speci c conventions are known to be hard to process by standard NLP tools. Prior to extracting triples, we follow a two-fold approach to tweet normalization [44] which can be readily extended to normalize social content from other platforms [45,46]. On the one hand, we remove tokens and token sequences encoding platform-speci c metadata or denoting communicative conventions that (typically) do not carry any syntactic function in the tweet sentence. Namely, we remove:

- sentiment emoticons and smileys;
- reserved tokens (e.g., RT for 'retweet');




(a) Dependency parse of a tweet's original text.




(b) Dependency parse of a tweet after preprocessing

**Fig. 2.** Example of tweet preprocessing.

- URLs.

On the other hand, we keep by default other platform-speci c tokens that can carry syntactic functions depending on the context like hashtags and @ entity mentions (e.g. *#digitaltransformation, @NASA*). Then, we identify token patterns that typically disrupt the syntactic parsing of the sentence, and remove them from the original tweet. Namely, we implemented the following preprocessing heuristics:

1. we remove sequences of $n$ entity mentions and retweet markers at the beginning of a sentence, with $n > 1$ or when the sequence is not followed by a verb. For example, we remove the leading sequence in *"@bansijpatel @RTatsat @kiranpatel1977 Thanks for updating the information with us."* but not in *"@AMDRyzen enabling #DataAnalytics in [...]"*.
2. for any sequence of size $n > 1$ hashtags/mentions/URL, we drop the sub-sequence with indexes $[1 : n]$ or drop the entire sequence if preceded by a sentence closing marker like ('!',':','?','.'). For example, in the text *"According to the @PayNews survey, 84 percent of #employees in the U.S. have instant access to #information about their pay and #bene ts #Sapper #AI #hr #support #goals[...]"* we keep only the rst element of the trailing hash tag sequence.
3. we remove a leading sequence of $n$ non-verbal tokens ($n$ empirically set to 6) ended by a column sign (e.g. *'Tech Update: Apple Watch's data 'black box' poses research problems [...]'*) as these frequent constructs (carrying a function similar to a tweet title) tend to mislead the dependency parsing.

Entity mentions and hashtags, that are typically removed from tweet preprocessing pipelines for NLP tasks such as sentiment analysis, are highly relevant for knowledge graph generation as they can be nominal subjects, objects, or modi ers of dependency parse trees and therefore be extracted as elements of candidate triples, like the tokens @mymdec and #SME in Fig. 2a. Notice, although, that the trailing sequence of purely referential elements can often lead to noisy edges, for example in the gure the parser wrongly draws a $dobj$ dependency edge from the main verb "launches" onto the hashtag #digitaltransformation.

Fig. 2b shows that the application of the preprocessing heuristics above (rule 2 in this case) can enhance the parsing on the tweet, without losing too much information.[13] The preprocessing step is carried out using the output of Spacy's English transformer pipeline *en_core_web_trf-3.6.1* after customizing the default Tokenizer in order to parse tweet metadata (e.g., mentions and hashtags).[14]

### 3.2. Triple extraction

In the triple extraction block, preprocessed tweets are split into sentences and each sentence is fed to the Spacy pipeline mentioned above. Building upon the works in [47] and [48], we de ne a set of procedures to extract candidate nominal entities and predicative triples connecting them from Spacy dependency parse trees.

#### 3.2.1. Entities

The entity extraction module detects local nominal phrases with a restricted range of syntactic modi cations (e.g., compound nouns, and adjectives). Then it connects and expands them with:

[13] Although we are currently not using them for KG generation, we are currently saving each tweet's metadata.

[14] https://github.com/explosion/spacy-models/releases/tag/en_core_web_trf-3.6.1.

CloudMile Wins 2020 Google Cloud Partner of the Year : accelerating digital transformation in Asia governments - Yahoo Finance

Lyfts $20 per month membership program gets new bike-share benefits and an annual plan ... #digitaltransformation

78 % of #healthcare organizations are currently deploying #cloud computing , with 20 % planning to deploy it in the future . More trends driving #DigitalTransformation in the industry via

Less than 15 % of the #banks considered themselves as #digitaltransformation leaders ! Lets take a look at the standard customer onboarding process for most US commercial banks.

**Fig. 3.** Visualization of candidate entities extracted from a sample of tweets.

**Table 1**
List of target and some of the discarded relation dependency paths.

| Target dependency paths |
|---|
| $[nsubj, dobj]$ |
| $[acl, relcl, dobj]$ |
| $[acl, dobj]$ |
| $[nsubjpass, agent, pobj]$ |
| $[nsubj, dobj, conj]$ |
| $[nsubj, conj]$ |
| **Sample discarded paths** |
| $[obj, pobj]$ |
| $[obl, pobj]$ |
| $[nsubj; pobj; nmod]$ |

- a non-recursive set of attached prepositional phrases;
- Spacy quantity-type entities (MONEY, PERCENT, QUANTITY, CARDINAL).

We also use pronominal anaphora links output by the Spacy pipeline component coreferee[15] and assign to it the expanded entity span of the token it points to.

Overall, the module ends up with a set $E = e_0, ..., e_n$ of non-uni ed, candidate entity phrases.

In Fig. 3 we show a sample of extracted candidate entities. For multi-token entity spans including quantifying modi ers (e.g. *'Less than 15% of the #banks'*) we maintain a structured representation separating the lexical head (*'#banks'*) from the quantifying modi cation of the noun phrase (*'Less than 15%'*), which then allows a more accurate entity normalization (see Section 3.3 below).

Notice that at this stage the hashtag *#digitaltransformation* in the second sentence and the noun phrase *digital transformation* in the rst are not mapped to the same general concept DIGITAL TRANSFORMATION23 yet, so that the triples in which they occur would still be considered as unrelated.

### 3.2.2. Relations

In the relation extraction module, for each sentence $s_i$ all the shortest paths of the dependency tree between each pair of entities $(e_m, e_n)$ containing a verb and matching any of the patterns listed in Table 1 are selected.

The target pattern set has been selected through an expert validation process. First, an open-domain text corpus was automatically annotated with entities, and all shortest paths connecting any pair of entities were collected, resulting in a total of approximately 15k path instances, ranging over 3695 path patterns. The patterns were then sorted by their frequency in the corpus. Next, the top twenty patterns, with frequencies ranging from 79 to 1098, were manually reviewed. Three independent evaluators assessed a random sample of 20 triples from each pattern. Speci cally, each evaluator was tasked to assess the correctness of all 400 relevant triples.

In order to be annotated as valid, a triple should reflect the semantics of the portion of the sentence where it was extracted. For example, the triple $< Mr.\ Lewis; give; quixotic\ guided\ tour >$ extracted from the sentence *'Mr. Lewis gives the reader a quixotic guided tour through Silicon Valley while showing how its success stories revolutionized American business.'* with path $[nsubj, dobj]$ was considered valid by the annotators. On the other hand, the triple $< air; rising; hot\ day >$ from the sentence *'Howe says it was discovered by cows drawn to cool air rising from the mouth of the cave on a hot day.'*, with path $[acl, pobj]$ was discarded as most of the annotators did not believe it accurately reflected the semantics of the corresponding text.

A majority vote was used in order to label each triple as correct/incorrect and only the subset of patterns with a prevalence of correct triples (i.e., more than 10) were considered reliable and kept in the result list.

Although this pattern expert validation process was carried out on a separate text corpus, while evaluating our pipeline on micro-blogging posts (see Section 5), we noticed that a potential source of noise was the extraction of triples via the dependency path

[15] https://github.com/richardpaulhudson/coreferee.

$[acl, dobj]$. The issue arose in instances where the noun's clausal modifier was an infinitive verb, as exemplified in the following sentence:

*'Salesforce really has the power to transform your business.'*

from where a triple $< power, transform, business >$ was wrongly extracted. Consequently, we added a constraint to the dependency path $[acl, dobj]$ in order to filter out those paths where verb nodes had a relation *aux* with an infinitive particle node. In the example above, *transform* has an *aux* relation to the particle *to* and, therefore, it is discarded. More in detail, the following expressions hold:

$\texttt{SUBJ} = power \xrightarrow{acl} \texttt{PRED} = transform \xrightarrow{dobj} \texttt{OBJ} = your\ business$

$\texttt{PRED} = transform \xrightarrow{aux} \texttt{to}.$

The entire updated process generates a set of verbal relations $V = v_0, ..., v_k$ and a set of triples $S = s_0, ..., s_k$ of the form $< e_m, v, e_n >$ where $v \in V$ and $e \in E$.

Analogously to what we pointed out for the entities, note that $v$ in $V$ are surface forms, that is individual inflected verbal forms that are unable to generalize triples over morphological or lexical variations. So for example the following triples:

$< BLEND360, acquires, EngagementFactory >$

$< BLEND360, acquired, EngagementFactory >$

$< BLEND360, bought, EngagementFactory >$

are considered distinct at this stage.

The final goal of the pipeline is to allow to generalize from the set $S = s_0, ..., s_k$ of surface form triples of type $< e_m, v, e_n >$, to the lower sized set $T = t_0, ..., t_h$ of triples of the form $< {}_m, r, {}_n >$ where each ${}_i \in E$ is an entity and $r$ is a label in a common relation vocabulary $R$.

### 3.3. Entity refining

The function of this block is to clean up and normalize the candidate entities into a normalized form that allows the merging across entity name variants.[16]

Entities are first cleaned up by removing leading/trailing punctuation marks as well as stop-words. Afterwords, we distinguish the following cases for normalization:

- For hashtags and @ mentions, we remove hashtags and @ symbols, split the "camel case" forms (e.g., *#SmartCities*) and lowercase the resulting string.
- For all other entities, we lemmatize and lowercase all component tokens whose POS tag is neither Verb nor Proper Noun, otherwise we simply lowercase.
- For nouns that have variants in American English, we finally map to the British English variant.

We take advantage of such normalized versions of candidate entities in order to merge them, by using the Spacy DBpedia Spotlight Entity Linking library.[17]

The DBpedia Spotlight model is trained to perform both entity detection and linking. In order to power this module with the entity normalization performed by our pipeline, we run it on modified tweet sentences where the original subjects and objects entity spans (extracted by the Entity Extraction module of the Triple Extraction block in Fig. 1) are replaced with their normalized forms. Next, we link the normalized versions of the entities to the corresponding DBpedia entries of the Spacy native entities whose text spans are both:

- included within the subject or object text spans of the corresponding normalized version of the entities;
- overlapping with the syntactic heads of the corresponding normalized version of the entities.

In other terms, we let the Spacy DBpedia Spotlight module perform the merging of entities that were normalized to the same or similar forms, by having them linked to the same DBpedia unique entries. For example, the two candidate entities *'Gartner'* and *'@Gartner_inc'* are merged together by linking them (later formalized with a relation `owl:sameAs` to the DBpedia entry of the Gartner consulting firm http://dbpedia.org/resource/Gartner).

In case only the first condition is met, we assign a semantic 'relatedness' link between the candidate entity and the DBpedia entry, indicating that the former is not an instance of, but rather related to the latter.[18] For example, the span *'@gartner_survey'* is considered only 'related' (later mentioned with a relation `skos:related`) to the DBpedia entry for Gartner.

In Section 4 we describe how these relations are encoded in the resulting knowledge graph by inheriting from existing ontology relations.

[16] Splitting is another typical subtask of Entity Refining functions, for example by separating the individual entities in parsed coordinated noun phrases like in *'#testautomation and #datamanagement can accelerate your #digitaltransformation'* We deal with these cases earlier on at the triple extraction phase by generating a triple for each coordinated entity.

[17] https://spacy.io/universe/project/spacy-dbpedia-spotlight.

[18] We keep out the cases when only the second condition is met, as they typically arise from inaccuracies of the entity span detection.

### 3.4. Relation re ning

This block aims to nd the best predicate label $r$ for each relation verb $v$ in a triple $< e_m, v, e_n >$ and to map $v$ to $r$ in the resulting triple.

The approach we followed consisted of deriving a word embedding representation of the verb predicates from a pre-trained model, computing an optimized clustering of the relation vectors, and nally using a representative instance of each cluster to map verb predicates.

After experimenting with several (contextual and non-contextual) word embedding models and clustering algorithms, we converged to a setup using static word embeddings learned with GloVe (Global Vectors for Word Representation, [49]) and applying HDBSCAN clustering to the vectors. We tested using verb phrase contextual embeddings from Huggingface's *bert-large-uncased*[19] and Sentence-BERT.[20] However, it turned out that the optimal cluster scores, in this case, were achieved for a number of clusters too close to the number of items in the dataset.[21] In Appendix A we report the clustering scores and number of resulting clusters for some best performing con gurations using the different embedding models.

*Relation embeddings* For each single or multi-token relation predicate, we get the static, 300-dimensional word embedding vector made available for text Span objects in the Spacy *en_core_web_lg-3.6.0* pipeline.[22]

*Dimensionality reduction and clustering* We used the HDBSCAN clustering algorithm enhanced by previously applying UMAP dimension reduction technique on the word embeddings vectors.[23] HDBSCAN is a hierarchical version of the popular density-based DBSCAN algorithm, which is characterized by considering outliers and leaves unclustered the data points lying in low-density regions [50]. Consequently, high dimensional data require more observed samples to produce the suitable level of density for HDBSCAN to work properly. However, applying UMAP to perform non-linear, manifold aware dimension reduction [51] has been proven to transform the datasets down to a dimension small enough for HDBSCAN to cluster the vast majority of instances.

In order to optimize the combination of UMAP and HDBSCAN, we perform a grid search over the hyper-parameters of both algorithms and evaluate the clustering using the score indicated in Equation (1):

$$S = silhouette_X \quad clustered_X, \tag{1}$$

where the silhouette coefficient $silhouette_x$ of an instance $x \in X$ is de ned in Equation (2):

$$(b \quad a)/max(a, b), \tag{2}$$

with $a$ being the mean distance to the other instances in the same cluster and $b$ being the mean distance to the instances of the next closest cluster.

In the $S$ score formula, the $silhouette_X$ is the mean silhouette coefficient over all the instances of the dataset $X$ that were actually clustered by HDBSCAN [52] while $clustered_X$ is the fraction of instances of $X$ that were actually clustered by HDBSCAN.

In practice, we optimize for the classical measure of cluster cohesion and separation while penalizing the con gurations with low coverage of the dataset. We nally chose a subset of best-scoring hyper-parameter con gurations and plotted their $S$ score over the number of output clusters they generate, so that we are able to pick a sub-optimal con guration that balances between generalization (fewer clusters) and accuracy (cluster number closer to the dataset size).

*Relation mapping* The triples in our dataset often contain numerous distinct relations that might be seen as synonyms. For instance, relations such as "includes", "involves", "embeds" and "contains" can convey similar meanings. To minimize redundancy and support semantic retrieval of the triples in the graph, we consolidate these extracted relations into a smaller set of prede ned relations. Therefore, for each relation verb $v$ in the dataset, we replace it with the predicate label $r$ consisting of the lemma of the most frequent relation in the cluster of $v$. Otherwise, we map it to itself if $v$ was an outlier and not clustered. Thus, the three distinct triples shown in the last example of Section 3.2.2 would be merged and the resulting triple would be:
$< BLEND360, BUY, EngagementFactory >$.[24]

## 4. The use case: digital transformation monitoring

The recent surge in data science and arti cial intelligence technologies has led to signi cant insights and aided in the creation of numerous decision-making tools [53]. These instruments assist investors in decision-making and policymakers in creating policy interventions, which have the potential to boost economic growth and enhance societal well-being [54,55].

[19] https://huggingface.co/bert-large-uncased.

[20] https://sbert.net/.

[21] In other terms, these representations were not suitable for generalizing enough over relations, probably due to the context-speci c information they are encoding.

[22] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.6.0.

[23] https://umap-learn.readthedocs.io/en/latest/parameters.html.

[24] A CSV le with a sample of the most frequent normalized triples, together with the originally matched relations can be found in the project repository at https://github.com/zavavan/dtm_kg/blob/master/data-collection/twitter/sampleNormalizedTriples.tsv.

**Table 2**
A sample of statements extracted by our pipeline.

| Subject Entity | Relation | Object Entity |
|---|---|---|
| pandemic | accelerate | digital_transformation |
| arti cial_intelligence | impact | insurance_sector |
| microsoft | buy | riskiq |
| data-driven_insight | drive | decision-making |
| agile_business | demand | effective_marketing_capability |
| hootsuite | buy | ai_chatbot_ rm |
| automl | generate | data-driven_insight |
| image_classi cation | use | transfer_learning |
| new_belgium_brewing | implement | digital workflow_place_solution |
| e-rupi | back | existing indian rupee |
| 82%_of_cio | implement | new_technology |
| image_recognition_framework | use | arti cial_intelligence |
| microinsurance | close | africa_insurance_gap |
| hsbc_qatar | introduce | mobile_payment |
| ford_motor_company | explore | blockchain_technology |

In particular, the application of these cutting-edge technologies to social media and news has great potential since they provide a larger set of information than standard lower frequency socio-economic indicators [56,57].

These opportunities and challenges are inspiring the research activities at the European Commission's Competence Center on Composite Indicators and Scoreboards[25] at the Joint Research Centre (JRC)[26] aimed at the development of a tracker of societal and economic activities in European countries using unconventional data [58].

In light of this, we have deployed our prototype pipeline to develop a Digital Transformation monitoring system from alternative sources. The technological domain of Digital Transformation is widely pervasive in both scholarly and industrial publications (scienti c papers, patents) as well as in the fast-reactive news and social media, capturing the latest updates in the eld: therefore, it represents a relevant benchmark for the capacity of our prototype to link and extend existing knowledge graphs generated from conventional sources. At this aim, we have generated a knowledge graph in the domain of Digital Transformation from a topic-speci c tweet dataset.

The dataset was collected by using the Twitter public API v2 full-archive search endpoint, retrieving English language tweets from 2022 containing the hashtag #DigitalTransformation. We excluded all retweets. From the approximately 4M tweets matching the query, we sampled a dataset of around 100k[27] and run the pipeline on it.

The resulting DTSMM_KG (Digital Transformation Social Media Monitor Knowledge Graph) comprises approximately 22,270 statements. We represented all claims extracted from the tweets using the DTSMM_KG ontology we created for this purpose.[28] Table 2 shows a sample of these statements. The reader notices that we refer to statements or triples indifferently. The triples obtained after the rei cation have not been taken into account for the statistics reported in this paper.

We rei ed each claim into *dtsmm-ont:Statement* class instances, where *dtsmm-ont* is the namespace pre x of the DTSMM_KG ontology and a *dtsmm-ont:Statement* de nes a speci c claim extracted from a given number of tweets. Namely, each statement includes:

- the rei cation of the triple itself via *rdf:subject*,*rdf:predicate* and *rdf:object* predicates;
- a data property *dtsmm-ont:hasSupport* reporting the number of tweets supporting the claim;
- a number of object property instances *dtsmm-ont:comesfromTweet* ranging over ontology instances of type *dtsmm-ont:Tweet* (which was inherited from *schema:SocialMediaPosting*) supporting the claim;
- A boolean data property *dtsmm-ont:negation* flagging whether a negation of the claim's predicate was parsed from the source text.

Fig. 4 shows a shortened example of a claim rei cation having the DTSMM_KG ontology's instance *machine_learning* as *rdf:object* and support equal to six.

In Fig. 5 instead we visualize a sub-graph of DTSMM_KG showing a few sample triples having the instance *machine_learning* as the subject. Here, we report just the statements, hiding claim rei cation for the sake of readability.

The linking of the DTSMM_KG instances to the DBpedia ontology is implemented by using the *owl:sameAs* and *skos:related* predicates in order to encode entity equality and relatedness, respectively. DTSMM_KG provides 2,857 *owl:sameAs* links and 3,309 *skos:related* links to DBPedia entries. Fig. 6 shows some examples *owl:sameAs* and *skos:related* edges from a number of entities onto the DBpedia resource http://dbpedia.org/resource/Machine_learning (the node in yellow).

[25] European Commission's Competence Center on Composite Indicators and Scoreboards (COIN): https://composite-indicators.jrc.ec.europa.eu/.
[26] The Joint Research Centre (JRC) of the European Commission (EC): https://ec.europa.eu/info/departments/joint-research-centre_en.
[27] After removal of duplicates and near-duplicates, namely tweets over a 0.85 Levenshtein string similarity threshold, computed after preprocessing.
[28] The ontology de nitions are located within the same le of the triple store.

```
dtsmm-ont:statement_10100 a dtsmm-ont:Statement,
 rdf:Statement ;
dtsmm-ont:negation false ;
dtsmm-ont:comesfromTweet dtsmm:tweet_1424266328882429952 ;
...
dtsmm-ont:hasSupport 6 ;
rdf:subject dtsmm:multi_page_document_classification ;
rdf:predicate dtsmm-ont:use ;
rdf:object dtsmm:machine_learning .
```

**Fig. 4.** A shortened example of rei cation for a Statement concerning the instance *machine_learning*, grounded by 6 tweets, with the three dots referring to the hidden *dtsmm-ont:comesfromTweet* predicates.

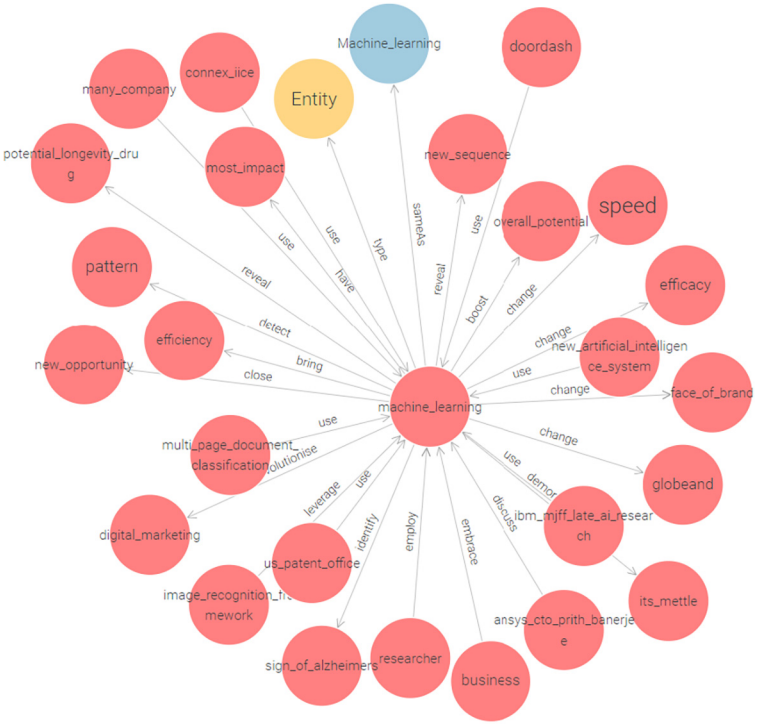



**Fig. 5.** A subgraph from DTSMM_KG showing a few sample claims for the instance *machine_learning*, prior to applying explicit statement rei cation.

The resulting data have been made publicly accessible under Creative Commons Attribution 4.0 International (CC BY 4.0) license[29] within the Joint Research Centre Data Catalogue,[30] as well as within the European Data portal,[31] the official data repository of the European Commission. The direct link to the Digital Transformation knowledge graph, available in Terse RDF Triple Language (Turtle), is https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/CC-COIN/se-tracker/DTSMM_KG.ttl. In Appendix B we also explain how the knowledge graph can be used in domain applications, along with a practical Q/A exercise via Retrieval Augmented Generation (RAG) [59].

We now report on the experimental results of generating the DTSMM_KG knowledge graph via the processing modules described in Section 3.

### *4.1. Preprocessing and triple extraction*

As for the extracted entities, around 33.9% and 6.44% included hashtags and @ entity mentions, respectively; 3.34% were complex noun phrases with prepositional attachments while around 16.6% contained quantitative modi ers of any type (currency, percent, etc.).

Out of all the generated triples, a 5.98% had either the subject or object entity made by a resolved pronominal anaphora, while 0.93% and 4% had $NEGATION$ and $INTERROGATIVE$ tags, respectively.

[29] Creative Commons Attribution 4.0 International license. https://creativecommons.org/licenses/by/4.0/

[30] https://data.jrc.ec.europa.eu/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139.

[31] https://data.europa.eu/88u/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139.


**Fig. 6.** A subgraph from DTSMM_KG showing entity linking via *owl:sameAs* and *skos:related* predicates of some KG instances to the DBpedia resource http://dbpedia.org/resource/Machine_learning (the blue node).

**Table 3**
Sample relation verb-predicate mapping.

| Relation Verb | Relation Predicate | Example |
|---|---|---|
| fuel | FUEL | *'How the UR+ Ecosystem is Fueling Cobot Market Growth'* |
| driven by | FUEL | *'Digital transformation in Ho Chi Minh is being driven by remote working'* |
| accelerated by | FUEL | *'huge social trends being accelerated by the pandemic.'* |
| identify | IDENTIFY | *'Machine learning can identify signs of Alzheimers in patients'* |
| quantify | IDENTIFY | *'Research quanti es G's potential in roaming and manufacturing'* |
| predict | IDENTIFY | *'AI-supported test can predict eye disease that leads to blindness'* |

### 4.2. Relation clustering

Starting with a set of 29,335 raw triples,[32] we derived 2,539 unique 300-dimensional word embeddings from GloVe and standardized them.

Via the grid search optimization described in Section 3, we converged to a UMAP two-dimensional representation of the vector dataset, using a $n\_neighbors = 5$ hyper-parameter, which constrains UMAP to look at rather local neighborhoods of ve data points when attempting to learn the manifold structure of the data. We then optimized UMAP dimensionality reduction and HDBSCAN clustering on this reduced dataset via the hyper-parameter grid search described in Section 3.

Table 3 shows some sample mappings from relations to their associated predicate labels, consisting of the lemma of the most frequent relation in their clusters.

## 5. Experimental evaluation

We perform a twofold evaluation of our methodology. First, we evaluate the precision, recall, and F1 by manually assessing the truthfulness of a test set of statements. Second, we evaluate our pipeline's precision against a number of alternative tools.

### 5.1. Human expert assessment

For the rst evaluation, we randomly selected 483 statements, equally distributed among high-support (support greater or equal to 5) and low-support triple groups. Each triple was assessed by three evaluators as True or False. The 'True' label was assigned only when all the following criteria were satis ed:

[32] These are surface-level candidate triples from the Triple Extractor, counted prior to entity and relation merging.

**Table 4**
Triple evaluation over a manually annotated set of 491 tweets.

| TP | FP | TN | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 464 | 19 | 20 | 0.96 | 0.95 | 0.95 |

- the subj and obj entities are linked by a relation in the tweet text;
- the assigned relation label entails the relation verb in the tweet text;
- the spans of the subject/object of extracted triples include the syntactic head of the relation's subject/object.[33]

We calculated the average pair-wise Cohen $\kappa$ inter-rater agreement, resulting in a value of 0.61. This value generally suggests a signi cant level of agreement. We also computed the Fleiss $\kappa_F$ agreement of all the 3 raters. This is ranging in [ 1, +1] and is de ned according to [60] as:

$$\kappa_F = \frac{p_o \quad p_e}{1 \quad p_e}, \tag{3}$$

where $p_o$ is the observed inter-annotator agreement and $p_e$ is the prior probability estimates of the inter-annotator agreement, that is the agreement that we would expect if the annotators were annotating randomly. The Fleiss $\kappa_F$ score reaches 0.558, which indicates a substantial agreement, although one annotator featured an outlier rating on a speci c category of cases.

We evaluated the 483 statements extracted by our pipeline using the majority vote of the three annotators, yielding a precision of 0.96. To compute the recall, the three annotators were assigned an additional task: extracting triples that they deemed correct from the same tweets containing the 483 selected statements. The total count of these tweets was 491, which exceeded the number of triples due to some being extracted from multiple tweets. The total number of triples manually extracted from the annotators was 484 (we considered the union of all the triples extracted by each annotator). Consequently, we were able to calculate the number of true positives (TP), false positives (FP), and true negatives (TN). Table 4 displays the TP, FP, and TN values for the 484 manually extracted triples. The table shows that the achieved recall was 0.95 and the F1 score was 0.95.

Individual rater estimates ranged from 0.90 to 0.96. Overall, these results indicate that the pipeline can extract triples with good precision from noisy text like tweets, while at the same time missing only a few triples.

Upon analyzing the results, we identi ed the primary error sources in the following descending order: i) failure in the syntactic parsing of the sentence (5 cases), ii) inaccuracy of relation clustering/mapping (4 cases), and iii) error in pronominal anaphora resolution (4 cases).

### 5.2. Comparative evaluation

In the second evaluation, we randomly sampled 500 tweets from the 100k-sized original dataset and used our pipeline to extract candidate entities. We then merge this set of entities with the one generated by the DyGIEpp Extractor [61].

DyGIEpp is an IE framework that is able to jointly extract a set of 6 pre-de ned types of entities (Method, Task, Material, Metric, Other-Scienti c-Term, and Generic). To detect the entities DyGIEpp uses a feed-forward neural network on textual span representations and computes a score for each entity type; an entity is detected considering the highest score for an entity type if a minimum threshold is met.

We then employed four alternative methods to identify relationships between these entities and thus extract the statements from the 500 tweets. Speci cally, we compared:

- **OpenIE Extractor**, the IE tool of the Stanford Core NLP suite [62], which is used to extract open-domain relationships composed by only one verb among candidate entities from our pipeline;
- **PoST Extractor**, a module built on top of the Stanford Core NLP suite that uses PoS tags to nd all verbs that exist between two candidate entities in a sentence to extract verb relations, using a window of max token distance 15 between the entities;
- **Dependency-based Extractor**, a module that extracts dependency trees using the dependency parser of the Stanford Core NLP suite, maps entities previously extracted using DyGIEpp into the sentence tokens, and exploits 12 hand-crafted paths[34] to nd verbs that connect entities.
- **Entity and Relationship Re ner**, a module that applies *Entity and Relationship Handlers* as described in [48] to the set that includes *OpenIE Extractor*, *PoST Extractor*, and *Dependency-based Extractor* triples. Its resulting triples have normalized entities that underwent several preprocessing steps, and the relationships are mapped to a controlled vocabulary which ensures that extracted verbs with a similar meaning are mapped to the same relationship.

[33] For example, a triple $< 78\%\_of\_\#healthcare, USE, Digital\_Transformation >$ would be marked as False if extracted from the text *'78% of #healthcare organisations deploy #DigitalTransformation'* as the syntactic head is organisations.

[34] https://github.com/danilo-dessi/SKG-pipeline/blob/main/resources/path.txt.

**Table 5**
Precision (P) of the triples extracted from a set of alternative methods from a set of 500 tweets, using a combination of Triplétoile and DyGIEpp candidate entities.

| Extraction Method | Precision |
|---|---|
| OpenIE Extractor | 0.52 |
| PoST Extractor | 0.17 |
| Dependency-based Extractor | 0.77 |
| Entity and Relationship Re ner | 0.31 |
| Triplétoile | **0.82** |

The number of extracted triples from the dataset ranged from 339 for the Dependency-based Extractor to a maximum of 1,015 for the PoST method. The latter is a quite predictable outcome as the PoST Extractor combines type-restricted and open-domain entities and at the same time extracts as candidate relations all the verbs occurring between any pair of entities in text, without ltering on the dependency relations connecting them.

After PoST Extractor, our Triplétoile pipeline is the one generating the largest number of triples (663) among the methods using the extended range of candidate entities, with OpenIE Extractor producing 588 triples Entity and Relationship Re ner reaching approximately 348 triples. In order to use these numbers as an indirect assessment of the relative recall levels of the different pipelines, we manually assessed also the precision on a limited random sample of 150 triples generated by each method.[35]

In order to evaluate the precision of these tools against Triplétoile, we manually assessed such 150 triples produced by every technique.[36]

Similarly to the previous evaluation, three evaluators reviewed each triple as 'True' or 'False'. The annotators' agreement reached a $\kappa_F$ of 0.86, indicating a strong agreement. Finally, we calculated the precision of the ve methodologies using the majority vote. We report the results in Table 5.

While not as high as in the previous test set, the precision of our pipeline on this smaller sample largely outperforms all the alternative methods. Interestingly, it also yielded a signi cant advantage over the Dependency-based Extractor method, which deploys very similar syntactic information from the sentence. This may be due to the application of the processing step upstream of the triple extraction process.

## 6. Conclusions and future works

In this paper, we presented Triplétoile, an information extraction architecture optimized to generate open-domain knowledge graphs from micro-blogging text. The method is mostly unsupervised and does not integrate information from a target domain during the extraction process. Nonetheless, in a topic-speci c test collection of tweets related to the domain of Digital Transformation, the pipeline proved to outperform some of the state-of-the-art methods, generating mostly valid triples. Moreover, we showed that around 12% of entity occurrences are linked to DBpedia entries, suggesting that the method is potentially useful for tracking relevant entities in the target social media text collection.

We are currently experimenting with the transferability of the pipeline across different target domains and preliminary results are promising. As an example, we deployed it for the extraction of a signi cantly larger graph of 431k triples in the domain of Digital Health and found out that a 8% of the 86k extracted entities could be linked to DBpedia entries of domain relevant types (e.g., *dbpedia:Disease*, *dbpedia:Company*, *dbpedia:Drug*). Moreover, the pipeline runtime proved to scale linearly with the size of the document set, which in this case consisted of a larger corpus of 95k news items (23.8M words against 2.9M words of the current tweet collection).

A current limitation of our method is that it does not rely on the ontology speci cation of a target domain in order to customize the entity and relation extraction process. As a consequence, extracted entities are currently un-typed, which does not support the execution of more structured queries. Moreover, a domain-speci c classi cation schema for relations would allow setting up a supervised learning of the relation mapping. We expect this to bene t from ne-tuning contextual word embedding representations using Transformer architectures.

Therefore, we plan to work on an enhanced version of the pipeline that builds upon the entity and relation spans generated with the current approach and further classi es them into a granular representation tailored to the speci c domain.

As a longer-term goal, this will also help analyzing more thoroughly the usage of social media and other dynamic information sources for tracking and expanding existing knowledge graphs generated for scienti c and technological domains.

Last but not least, in light of the recent emergence of numerous scalable large language models, we intend to explore their potential to improve triple extraction methods [63]. Simultaneously, we aim to capitalize on the resultant knowledge graph to develop knowledge plugins [64], thus augmenting the pro ciency of these language models across various natural language processing tasks.

[35] Notice that these test sets are not generated from the same tweet subset for each pipeline. Notice also that the random sampling was done without using any information on the triple support, which was not available for the alternative pipelines.

[36] Note that this test set is not generated from the same tweet subset, for each pipeline; moreover, the random sampling was done without using any information on the triple support, which was not available for the alternative pipelines.

**Table A.6**
The table presents clustering score values and the number of output clusters for the top three performing UMAP-HDBSCAN con gurations across three tested embedding models. It's worth noting that the dataset comprises a total of 29,335 relation instances for contextualized BERT and Sentence-BERT embeddings. In contrast, for static GloVe embeddings, we consolidated single occurrences of each relation form, resulting in a nal set of 2,539 relations due to their context-independent vector representations.

| Embedding Model | Silhouette | Clustered Ratio | Num Clusters |
|---|---|---|---|
| BERT | 0.9387 | | 1107 |
| BERT | 0.9287 | | 918 |
| BERT | 0.9171 | | 1063 |
| Sentence-BERT | 0.6852 | | 869 |
| Sentence-BERT | 0.6794 | | 978 |
| Sentence-BERT | 0.6767 | | 1050 |
| GloVe | 0.6505 | | 327 |
| GloVe | 0.6362 | | **332** |
| GloVe | 0.6345 | | 511 |

## CRediT authorship contribution statement

**Vanni Zavarella:** Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Data curation. **Sergio Consoli:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis. **Diego Reforgiato Recupero:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Gianni Fenu:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Simone Angioni:** Software, Conceptualization. **Davide Buscaldi:** Resources, Investigation, Formal analysis. **Danilo Dessí:** Writing – review & editing, Software, Methodology, Data curation. **Francesco Osborne:** Writing – review & editing, Methodology, Formal analysis, Data curation.

## Declaration of competing interest

The authors declare the following nancial interests/personal relationships which may be considered as potential competing interests: Diego Reforgiato Recupero is an associate Editor of the Information Science section at the Heliyon journal.

## Data availability

The data about the produced Knowledge Graph are publicly accessible under Creative Commons Attribution 4.0 International (CC BY 4.0) license at https://data.jrc.ec.europa.eu/dataset/f7be47f7-49a2-44e8-9dc8-043735af4139.

## Acknowledgements

We acknowledge nancial 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).

## Appendix A. Relation clustering performance for different embeddings

See Table A.6.

## Appendix B. Example of use in domain applications via Q A and RAG

The knowledge graph we have developed is directly applicable to various domain applications, particularly within the realm of Digital Transformation monitoring. Our approach bridges the gap between the wealth of information available in real-time data streams, like social media, and more static, conventional sources. This fusion yields a dynamic and comprehensive view of the Digital Transformation landscape, aiding in real-time monitoring, informed decision-making, and predictive analytics. For instance, the European Commission's Competence Center on Composite Indicators and Scoreboards at the Joint Research Centre (JRC) is at the forefront of exploring unconventional data to track societal and economic activities across European countries. This activity aligns with our efforts and showcases a practical application where our knowledge graph can signi cantly contribute. We have utilized our prototype pipeline to create a monitoring system speci cally tailored to the domain of Digital Transformation. This technological area is not only prevalent in academic and industrial literature, such as scienti c papers and patents but is also actively discussed

in dynamic platforms such as news outlets and social media, which often provide the most current insights. The pervasive nature of Digital Transformation makes it an excellent domain for demonstrating the utility of our knowledge graph.

Moreover, the knowledge graph might also serve as a critical resource for enriching Retrieval Augmented Generation (RAG) models [59]. In detail, RAG models combine the power of language models with a retrieval component, and our knowledge graph can be used as a novel RAG approach to fetch relevant information during the generation process. By querying our knowledge graph, a RAG model can dynamically pull in contextual data related to Digital Transformation, thus enhancing the quality and relevance of its outputs.

In the following, we provide a Q/A exercise showing how the knowledge graph can be used in domain applications via RAG. Suppose for example that you wish to know whether the multinational Microsoft is making signi cant investments in Computer Security. One might supply the following question to a RAG system:

*Is Microsoft dedicating substantial resources to computer security technologies?*

Using a Named-Entity Recognition model,[37] the system is able to recognize the entities *Microsoft* and *Computer Security* from the text.

Our knowledge graph, referred to as *DTSMM_KG*, can be queried via SPARQL to detect whether *Microsoft* entities are declared into its ontology:

```
SELECT DISTINCT *
FROM <DTSMM_KG>
WHERE { <http://dtsmmkg.org/dtsmmkg/resource/microsoft> ?p ?o . }
```

which would produce the following resulting triples (in RDF Turtle format):

```
@prefix dtsmm: <http://dtsmmkg.org/dtsmmkg/resource/> .
@prefix dtsmm-ont: <http://dtsmmkg.org/dtsmmkg/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

dtsmm:microsoft a dtsmm-ont:Entity ;
 owl:sameAs <http://dbpedia.org/resource/Microsoft>,
 <http://dbpedia.org/resource/Xbox_Live> .
```

From this we learn that the *Microsoft* resource is de ned and exists into our KG, and also that it is the well-known DBpedia entity http://dbpedia.org/resource/Microsoft, which would allow us to infer additional information, external to our knowledge-base, via the DBpedia SPARQL endpoint[38] with query:

```
SELECT DISTINCT *
FROM <DTSMM_KG>
WHERE { <http://dbpedia.org/resource/Microsoft> ?p ?o . }
```

which would produce 960 knowledge triples about Microsoft (to see all the different triples, you can browse directly DBpedia to http://dbpedia.org/resource/Microsoft).

This existing knowledge can be enriched with the relations extracted from our Digital Transformation knowledge graph, via the SPARQL query:

```
prefix dtsmm: <http://dtsmmkg.org/dtsmmkg/resource/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT *
FROM <DTSMM_KG>
WHERE {
 ?statement rdf:subject dtsmm:microsoft .
 ?statement rdf:predicate ?relation .
 ?statement rdf:object ?object .
}
```

[37] See e.g. https://huggingface.co/search/full-text?q=named-entity+recognition&type=model

[38] Available at https://dbpedia.org/sparql.

which would produce exactly 48 statements representing new knowledge deriving from our KG. For example, looking at the acquire predicate type (i.e. http://dtsmmkg.org/dtsmmkg/ontology#acquire), we would know that Microsoft has acquired companies like *Cloudknox_Security*, *CyberX* and *RiskIQ*. In SPARQL we might then ask for information about this last:

```
SELECT DISTINCT *
FROM <DTSMM_KG>
WHERE { <http://dtsmmkg.org/dtsmmkg/resource/riskiq> ?p ?o . }
```

with Turtle result as follows:

```
@prefix dtsmm: <http://dtsmmkg.org/dtsmmkg/resource/> .
@prefix dtsmm-ont: <http://dtsmmkg.org/dtsmmkg/ontology#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

dtsmm:cybersecurity_firm_riskiq a dtsmm-ont:Entity ;
 owl:sameAs <http://dbpedia.org/resource/RiskIQ> ;
 skos:related <http://dbpedia.org/resource/Computer_security> .
```

The results would tell us that this is a Computer Security company. If we now would supply via RAG the existing 960 DBpedia knowledge triples on Microsoft plus the extracted relation triples deriving from our KG in-context to a LLM (in this example we used OpenAI GPT-4 Turbo[39]), and then ask the question, specifying to be brief:

*Is Microsoft dedicating substantial resources to computer security technologies?*

we would get the following answer from the system:

*Yes, Microsoft is dedicating substantial resources to computer security technologies, as evidenced by its acquisitions of companies like RiskIQ, a leader in global threat intelligence and attack surface management, and CyberX, which specializes in securing IoT devices.*

where the latter information derives exactly from our KG, showing the power of the supplied new knowledge.

In summary, when generating textual content, the RAG model can then reference the most recent updates and developments in Digital Transformation encapsulated within our knowledge graph. This ensures that the generated content is not only linguistically coherent but also factually accurate and up-to-date, reflecting the latest trends and information. Such enrichment is particularly valuable in applications where staying current with industry changes is critical, such as health-care, market analysis, or creating summaries for decision-makers. Our knowledge graph acts as a pool of novel knowledge taken from social media that RAG models can tap into, by supplying them with a repository of timely and relevant information.

## References


[1] P.S. Raji, S. Surendran, RDF approach on social network analysis, in: 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), 2016, pp. 1–4.

[2] J. Dörpinghaus, S. Klante, M. Christian, C. Meigen, C. Düing, From social networks to knowledge graphs: a plea for interdisciplinary approaches, Soc. Sci. Humanit. Open 6 (2022) 100337, https://doi.org/10.1016/j.ssaho.2022.100337.

[3] Q. He, J. Yang, B. Shi, Constructing knowledge graph for social networks in a deep and holistic way, in: Companion Proceedings of the Web Conference 2020, WWW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 307–308.

[4] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyaschev, Ontology-based data access: a survey, in: IJCAI International Joint Conference on Arti cial Intelligence, vol. 2018-July, 2018, pp. 5511–5519.

[5] A. Hogan, The semantic web: two decades on, Semant. Web 11 (2020) 169–185, https://doi.org/10.3233/SW-190387.

[6] L. Ehrlinger, W. Wöß, Towards a de nition of knowledge graphs, in: M. Martin, M. Cuquet, E. Folmer (Eds.), 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, in: CEUR Workshop Proceedings, vol. 1695, CEUR-WS.org, 2016, pp. 1–4, https://ceur-ws.org/Vol-1695/paper4.pdf.

[7] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: opportunities and challenges, Artif. Intell. Rev. (2023) 1–32.

[8] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, Semant. Web 6 (2015) 167–195, https://doi.org/10.3233/SW-140134.

[9] A. Kumar, S.S. Singh, K. Singh, B. Biswas, Link prediction techniques, applications, and performance: a survey, Phys. A, Stat. Mech. Appl. 553 (2020) 124289.

[10] M. Nayyeri, G.M. Cil, S. Vahdati, F. Osborne, M. Rahman, S. Angioni, A. Salatino, D.R. Recupero, N. Vassilyeva, E. Motta, et al., Trans4e: link prediction on scholarly knowledge graphs, Neurocomputing 461 (2021) 530–542, https://doi.org/10.1016/j.neucom.2021.02.100.

[11] A. Borrego, D. Dessi, I. Hernández, F. Osborne, D.R. Recupero, D. Ruiz, D. Buscaldi, E. Motta, Completing scienti c facts in knowledge graphs of research concepts, IEEE Access 10 (2022) 125867–125880.


[39] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

[12] C. Barbosa, L. Félix, V. Vieira, C. Xavier, Sara - a semi-automatic framework for social network analysis, in: Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web, SBC, Porto Alegre, RS, Brasil, 2019, pp. 59–62.
[13] H. Alani, A. Gangemi, V. Presutti, D. Reforgiato Recupero, A.G. Nuzzolese, F. Draicchio, M. Mongiovì, Semantic web machine reading with FRED, Semant. Web 8 (2017) 873–893, https://doi.org/10.3233/SW-160240.
[14] J.L. Martinez-Rodriguez, I. Lopez-Arevalo, A.B. Rios-Alvarado, OpenIE-based approach for knowledge graph construction from text, Expert Syst. Appl. 113 (2018) 339–355, https://doi.org/10.1016/j.eswa.2018.07.017.
[15] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain, Future Gener. Comput. Syst. 116 (2021) 253–264, https://doi.org/10.1016/j.future.2020.10.026.
[16] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint, arXiv:2310.06825, 2023.
[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: open and efficient foundation language models, arXiv preprint, arXiv:2302.13971, 2023.
[18] W. Huang, X. Ma, H. Qin, X. Zheng, C. Lv, H. Chen, J. Luo, X. Qi, X. Liu, M. Magno, How good are low-bit quantized llama3 models? An empirical study, arXiv preprint, arXiv:2404.14047, 2024.
[19] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M.S. Kale, J. Love, et al., Gemma: open models based on Gemini research and technology, arXiv preprint, arXiv:2403.08295, 2024.
[20] OpenAI, Gpt-4 technical report, 2023, arXiv:2303.08774.
[21] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: an innate limitation of large language models, arXiv preprint, arXiv:2401.11817, 2024.
[22] L. Siddharth, L.T.M. Blessing, K.L. Wood, J. Luo, Engineering knowledge graph from patent database, J. Comput. Inf. Sci. Eng. 22 (2021) 021008, https://doi.org/10.1115/1.4052293.
[23] L. Siddharth, J. Luo, Retrieval augmented generation using engineering design knowledge, arXiv:2307.06985, 2024.
[24] Y. Xiao, C. Li, M. Thürer, A patent recommendation method based on kg representation learning, Eng. Appl. Artif. Intell. 126 (2023), https://doi.org/10.1016/j.engappai.2023.106722.
[25] T. Man, A. Vodyaho, D. Ignatov, I. Kulikov, N. Zhukova, Synthesis of multilevel knowledge graphs: methods and technologies for dynamic networks, Eng. Appl. Artif. Intell. 123 (2023), https://doi.org/10.1016/j.engappai.2023.106244.
[26] S. Yu, C. Peng, C. Xu, C. Zhang, F. Xia, Web of conferences: a conference knowledge graph, in: WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining, 2023, pp. 1172–1175.
[27] G. Tamašauskaite, P. Groth, De ning a knowledge graph development process through a systematic review, ACM Trans. Softw. Eng. Methodol. 32 (2023), https://doi.org/10.1145/3522586.
[28] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, L. Secchi, 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.
[29] P. Hitzler, A review of the semantic web eld, Commun. ACM 64 (2021) 76–83, https://doi.org/10.1145/3397512.
[30] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A survey on knowledge graphs: representation, acquisition, and applications, IEEE Trans. Neural Netw. Learn. Syst. 33 (2022) 494–514, https://doi.org/10.1109/TNNLS.2021.3070843.
[31] N. Noy, Y. Gao, A. Jain, A. Patterson, A. Narayanan, J. Taylor, Industry-scale knowledge graphs lessons and challenges, Queue 17 (2019), https://doi.org/10.1145/3329781.3332266.
[32] A. Hogan, E. Blomqvist, M. Cochez, C. D'Amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Comput. Surv. 54 (2021), https://doi.org/10.1145/3447772.
[33] P. Ristoski, H. Paulheim, Semantic web in data mining and knowledge discovery: a comprehensive survey, J. Web Semant. 36 (2016) 1–22, https://doi.org/10.1016/j.websem.2016.01.001.
[34] T. Tudorache, Ontology engineering: current state, challenges, and future directions, Semant. Web 11 (2020) 125–138, https://doi.org/10.3233/SW-190382.
[35] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, H. Sack, AI-KG: An Automatically Generated Knowledge Graph of Arti cial Intelligence, Lecture Notes in Computer Science, vol. 12507, 2020, pp. 127–143.
[36] J.F. Sequeda, W.J. Briggs, D.P. Miranker, W.P. Heideman, A Pay-as-You-Go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases, Lecture Notes in Computer Science, vol. 11779, 2019, pp. 526–545.
[37] D. Collarana, M. Galkin, C. Lange, S. Scerri, S. Auer, M.-E. Vidal, Synthesizing Knowledge Graphs from Web Sources with the Minte+ Framework, Lecture Notes in Computer Science, vol. 11137, 2018, pp. 359–375.
[38] E. Gabrilovich, N. Usunier, Constructing and mining web-scale knowledge graphs, in: SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016, pp. 1195–1197.
[39] N. Choudhary, N. Rao, S. Katariya, K. Subbian, C.K. Reddy, Self-supervised hyperboloid representations from logical queries over knowledge graphs, in: The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021, 2021, pp. 1373–1384.
[40] B. Haslhofer, A. Isaac, R. Simon, Knowledge Graphs in the Libraries and Digital Humanities Domain, Springer International Publishing, Cham, 2018, pp. 1–8.
[41] E. Hyvönen, H. Rantala, Knowledge-based relation discovery in cultural heritage knowledge graphs, in: CEUR Workshop Proceedings, vol. 2364, 2019, pp. 230–239.
[42] S. Cristofaro, E.M. San lippo, P. Sichera, D. Spampinato, Towards the representation of claims in ontologies for the digital humanities, in: CEUR Workshop Proceedings, vol. 2949, 2021, pp. 1–12, https://ceur-ws.org/Vol-2949/paper6.pdf.
[43] M. Mountantonakis, Y. Tzitzikas, Large-scale semantic integration of linked data: a survey, ACM Comput. Surv. 52 (2019), https://doi.org/10.1145/3345551.
[44] L. Siddharth, L. Blessing, J. Luo, Natural language processing in-and-for design research, Des. Sci. 8 (2022) e21, https://doi.org/10.1017/dsj.2022.16.
[45] S. Tuarob, C.S. Tucker, Quantifying product favorability and extracting notable product features using large scale social media data, J. Comput. Inf. Sci. Eng. 15 (2015) 031003, https://doi.org/10.1115/1.4029562.
[46] F. Chiarello, A. Bonaccorsi, G. Fantoni, Technical sentiment analysis. Measuring advantages and drawbacks of new products using social media, Comput. Ind. 123 (2020) 103299, https://doi.org/10.1016/j.compind.2020.103299, https://www.sciencedirect.com/science/article/pii/S0166361520305339.
[47] D. Dessì, F. Osborne, D.R. Recupero, D. Buscaldi, E. Motta, CS-KG: a large-scale knowledge graph of research entities and claims in computer science, in: U. Sattler, A. Hogan, C.M. Keet, V. Presutti, J.P.A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings, in: Lecture Notes in Computer Science, vol. 13489, Springer, 2022, pp. 678–696.
[48] D. Dessì, F. Osborne, D.R. Recupero, D. Buscaldi, E. Motta, SCICERO: a deep learning and NLP approach for generating scienti c knowledge graphs in the computer science domain, Knowl.-Based Syst. 258 (2022) 109945, https://doi.org/10.1016/j.knosys.2022.109945.
[49] J. Pennington, R. Socher, C.D. Manning, Glove: global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543, http://www.aclweb.org/anthology/D14-1162.
[50] C. Malzer, M. Baum, A hybrid approach to hierarchical density-based cluster selection, in: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, vol. 2020-September, 2020, pp. 223–228.
[51] L. McInnes, J. Healy, J. Melville, Umap: uniform manifold approximation and projection for dimension reduction, arXiv:1802.03426, 2020.
[52] F. Batool, C. Hennig, Clustering with the average silhouette width, Comput. Stat. Data Anal. 158 (2021), https://doi.org/10.1016/j.csda.2021.107190.
[53] S. Consoli, D. Reforgiato Recupero, M. Saisana, Data Science for Economics and Finance: Methodologies and Applications, Springer Nature, Switzerland AG, 2021.

[54] M. Taddy, Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions, McGraw-Hill, United States, 2019.
[55] T. Marwala, Economic Modeling Using Arti cial Intelligence Methods, Springer, Switzerland, 2013.
[56] L. Barbaglia, S. Consoli, S. Manzan, Forecasting with economic news, J. Bus. Econ. Stat. 41 (2022) 708–719, https://doi.org/10.1080/07350015.2022.2060988.
[57] S. Consoli, S. Barbaglia, S. Manzan, Fine-grained, aspect-based sentiment analysis on economic and nancial lexicon, Knowl.-Based Syst. 247 (2022) 108781, https://doi.org/10.1016/j.knosys.2022.108781.
[58] M. Colagrossi, S. Consoli, F. Panella, L. Barbaglia, Tracking socio-economic activities in European countries with unconventional data, in: ACM International Conference Proceeding Series, 2022, pp. 323–330.
[59] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red, Hook, NY, USA, 2020, pp. 9459–9474.
[60] R. Falotico, P. Quatto, Fleiss' kappa statistic without paradoxes, Qual. Quant. 49 (2015) 463–470, https://doi.org/10.1007/s11135-014-0003-1.
[61] D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, relation, and event extraction with contextualized span representations, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5784–5789.
[62] G. Angeli, M.J. Johnson Premkumar, C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 344–354.
[63] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: a roadmap, arXiv:2306.08302, 2023.
[64] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D.R. Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, IEEE Access 11 (2023) 22468–22489.