IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 July 2022, accepted 10 October 2022, date of publication 26 October 2022, date of current version 8 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217492

RESEARCH ARTICLE

Resorting to Context-Aware Background Knowledge for Unveiling Semantically Related Social Media Posts

AHMAD SAKOR[1,2], KULDEEP SINGH[3], AND MARIA-ESTHER VIDAL[1,2]
[1]TIB Leibniz Information Centre for Science and Technology, Leibniz University of Hannover, 30167 Hannover, Germany
[2]L3S Research Centre, 30167 Hannover, Germany
[3]Zerotha Research, 52058 Hannover, Germany

Corresponding author: Ahmad Sakor (ahmad.sakor@tib.eu)

This work was supported in part by the European Union (EU) H2020 Research and Innovation Action (RIA) Project CLARIFY under Agreement 875160; and in part by the Federal Ministry for Economic Affairs and Energy of Germany in the project Cognitive Economy Intelligence Plattform für die Resilienz wirtschaftlicher Ökosysteme (CoyPu), Germany, under Grant 01MK21007[A-L]. The work of Maria-Esther Vidal was supported in part by the Leibniz Association in the program "Leibniz Best Minds: Programme for Women Professors," Project TrustKG Transforming Data in Trustable Insights, under Grant P99/2020.

ABSTRACT Social media networks have become a prime source for sharing news, opinions, and research accomplishments in various domains, and hundreds of millions of posts are announced daily. Given this wealth of information in social media, finding related announcements has become a relevant task, particularly in trending news (e.g., COVID-19 or lung cancer). To facilitate the search of connected posts, social networks enable users to annotate their posts, e.g., with hashtags in tweets. Albeit effective, an annotation-based search is limited because results will only include the posts that share the same annotations. This paper focuses on retrieving context-related posts based on a specific topic, and presents PINYON, a knowledge-driven framework, that retrieves associated posts effectively. PINYON implements a two-fold pipeline. First, it encodes, in a graph, a CORPUS of posts and an input post; posts are annotated with entities for existing knowledge graphs and connected based on the similarity of their entities. In a decoding phase, the encoded graph is used to discover communities of related posts. We cast this problem into the Vertex Coloring Problem, where communities of similar posts include the posts annotated with entities colored with the same colors. Built on results reported in the graph theory, PINYON implements the decoding phase guided by a heuristic-based method that determines relatedness among posts based on contextual knowledge, and efficiently groups the most similar posts in the same communities. PINYON is empirically evaluated on various datasets and compared with state-of-the-art implementations of the decoding phase. The quality of the generated communities is also analyzed based on multiple metrics. The observed outcomes indicate that PINYON accurately identifies semantically related posts in different contexts. Moreover, the reported results put in perspective the impact of known properties about the optimality of existing heuristics for vertex graph coloring and their implications on PINYON scalability.

INDEX TERMS Social media networks, community detection, post relatedness, knowledge graph, COVID-19, knowledge retrieval.

I. INTRODUCTION

Capturing knowledge is of paramount relevance to support the new generation of data-driven digital technologies for improving quality of life [1], industrial competitiveness [2], and Web-based health data analysis [3]. Social networking channels have become an information dissemination media for personal discussions, as well as to report relevant scientific research results. In particular, it has become a common practice in the biomedical domain to announce public results

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

of clinical studies on social media channels.[1] These outcomes are relevant for the scientific community and of interest to a broader audience besides the biomedical domain. Given the wealth of knowledge encoded in these announcements, users on social media search to uncover exciting insights, information, and novel findings of trending topics such as COVID-19 and new lung cancer treatments. Nevertheless, effective search and recommendation tools are demanded to support users in hunting the most informative and meaningful social media posts. There are several approaches in the literature for recommending related posts based on the hashtags [4], sentence similarities [5], and by extracting similar concepts or entities in the post [6]. Albeit effective, existing approaches cannot utilize the plethora of knowledge available in publicly available knowledge sources, either encyclopedic or domain-specific. Especially in the biomedical domain of rare diseases such as lung cancer or a newly emerged pandemic like COVID-19, creating and curating sizable labeled training data is extremely challenging to employ deep-learning-based approaches for recommending related posts. Further, while searching, users require knowledge about specific terms, such as the name of the drug used for COVID-19 or newly tested interventions for lung cancer. However, deep-learning approaches are not suitable to search in domains, like biomedical relevant content in social media, with scarce training data (Section VI). Nonetheless, methods employing the community curated knowledge in Knowledge Graphs (KGs)–e.g., DBpedia [7], Wikidata [8], and Unified Medical Language System (UMLS) [9] may have a pivotal role for unveiling semantically related posts.

A. PROBLEM STATEMENT AND APPROACH

In this paper, we tackle the problem of identifying in a predefined dataset of social media posts the most semantically related ones to a given input post. Moreover, we address the issue of data scarcity in newly emerged scenarios such as COVID-19. We propose PINYON, a novel approach that resorts to the contextual KGs encoding domain-specific (e.g., UMLS) and encyclopedic knowledge (e.g., DBpedia) to extract semantically related posts. PINYON focuses on capturing knowledge about posts by determining relatedness among the entities mentioned in a post based on semantic similarities computed from KG embeddings–e.g., RDF2vec [10]. Moreover, PINYON exploits similarity values to compute communities of highly related posts in a given context (domain-specific or encyclopedic). Specifically, we map the problem of context-aware post recommendation into the Vertex Coloring Problem [11], [12] and propose a heuristic algorithm *PINYON-Context-Aware-Community-Detection(PINYON-CACD)* to efficiently identify highly related posts in various contexts. We rely on a natural intuition that given an input post, the recommended posts will be similar to this post and share similarities across all the recommended posts. *PINYON-CACD* extends DSATUR [11] a well-known heuristic algorithm for the vertex coloring problem. It is guided by a new heuristic for assigning a minimal number of colors. The aim here is to maximize the similarity between the vertices colored in the same color and minimize the similarity between vertices in different colors. The colored vertices are used to generate communities of posts that are semantically similar to the input post for a relevant context. In a nutshell, we legitimize (theoretically and empirically) applying vertex coloring algorithms to solve the community detection problem efficiently. More importantly, the proposed approach is interpretable and transparent, allowing error tracing and explanation at every step.

To support the PINYON approach described earlier, background knowledge (BK) from the entities' alignments is collected from existing KGs such as DBpedia, Wikidata, and UMLS. These alignments are used for computing the latent representation of the background knowledge using embeddings in the low-dimensional vector space. In total, we aligned around 100 million entities collected from the KGs. We choose the mentioned KGs because of: 1) the richness of the used KGs and how frequent they are updated. A new release for DBpedia is created every month [13]. Wikidata is updated daily with further contributions from the crowd.[2] The UMLS is updated in May and November of each year.[3] 2) The availability of entity linking tools for these KGs and the high performance of the current state-of-the-art entity linking tools for the used KGs [14]. However, PINYON is agnostic of the entity recognizer and linking tools, and it can be configured to other KGs as far as entities can be linked to those KGs. We have exhaustively evaluated PINYON in the three social media topics (COVID-19, Lung Cancer, and Soccer), containing in total 2,448,612 records (aka. posts), compared against baselines including transformer-based approaches (e.g., [5], [15], [16]). Our experiment results suggest that PINYON significantly outperforms all the existing baselines, specifically in the biomedical domain. Our results also suggest a significant finding that knowledge already encoded in domain-specific KGs such as UMLS empowers PINYON during the discovery of semantically related posts in COVID-19 and Lung cancer topics and discussions.

B. FURTHER CONTRIBUTIONS

PINYON is available as an open-source tool, and its source code is released to ensure reproducibility. It can be found in our public Github.[4]

The paper is structured as follows: we analyze related literature in Section II. Section III motivates our work by illustrating posts related to an announcement of a promising treatment for lung cancer; Section IV details our approach; Section V details the implementation of our approach;

[1] https://twitter.com/WHO/status/1317032089951358977

[2] https://dumps.wikimedia.org/wikidatawiki/entities/

[3] https://www.nlm.nih.gov/research/umls/faq_main.html

[4] https://github.com/SDM-TIB/PINYON

we present and discuss the outcomes of our empirical evaluations in Section VI. Finally, Section VII concludes our findings and outlines our future work.

II. RELATED WORK

A. RECOMMENDING RELATED POSTS IN SOCIAL MEDIA

Finding and recommending similar social media posts has attracted a considerable attention in the scientific research community [17], [18]. Work in [19] proposes a fuzzy inference system that learns the interests of the source and target users to categorize tweets. Tweet-Recommender system [20] provides tweets related to the news using topic similarities and language modeling. Several other approaches, such as [4], [18], [21], focus on hashtag-based or user interaction history for tweet recommendations. These methods primarily focused on hashtag similarities. However, they also rely on detecting communities in social network based on followers, mention, hashtag, and topic.

B. FINDING SENTENCE SIMILARITIES

Besides social media, finding similar sentences in a document or a Web article is a well-studied research domain. Kiros et al. [22] train an encoder-decoder model to predict surrounding sentences of an encoded passage in a given document. [23] introduce a transformer-based model for encoding sentences into embedding vectors for calculating semantic similarity between two sentences. Other approaches such as in [24] propose sentential embedding-based techniques for computing similarity between two sentences. Sentence-BERT [5] is the state-of-the-art, a BERT [25] based model that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. We position PINYON with Sentence-BERT and its similar extensions in biomedical domain: BERTweet [26], COVID-Twitter-BERT [16]. However, there are two fundamental differences. First, PINYON aims to discover post relatedness based on the context and meaning of the entities in a post. Second, PINYON is agnostic to the domain-specific training data and resorts to the BK captured in publicly available KGs. Hence, PINYON can easily adapt to varied domains without any training data. Experimental results in Section VI validate that PINYON significantly outperforms Sentence-BERT, BERTweet, and COVID-Twitter-BERT in fields with limited training data, such as tweets related to a new pandemic.

C. FINDING ENTITY RELATEDNESS

As our approach relies on entity relatedness, we now describe the research work aiming to predict if two entities are similar. The problem of finding related entities and patterns from an unstructured text has not been investigated widely. Some approaches learn patterns in dependency representations of sentences to find similarities between the entities and the predicates mention in different sentences [3], [27]. Other approaches compare entities based on the semantic meaning [28], or by extracting linguistic patterns [29]. Stevenson et al. [28] propose an approach that automatically learns patterns using WordNet [30] to find the similarity between entities and the predicates. For instance, in a given document, this approach aims at labeling the patterns ''president resign'' and ''executive leave job'' semantically similar. The approach starts with a small set of sample extraction patterns. It uses a similarity metric based on a version of the vector space model augmented with information from WordNet to learn similar patterns; however, it does not aim to label entities based on their similarity. Sematch [6] is a framework for the development, evaluation, and application of semantic similarity for KGs. Sematch is used to compute semantic similarity scores of the concepts, words, and entities in a KG. Sematch focuses on specific knowledge-based semantic similarity metrics; they rely on both structural knowledge in a taxonomy (e.g., depth, path length, least common subsumer) and statistical information contents (corpus-IC and graph-IC). Sematch only calculates the similarity between two entities at a time, while PINYON is free of such restriction. Some other approaches such as [20], [31], and [32] aim to find relevant tweets in domains/topics such as London Riots and news. However, due to unavailability of public code, these approaches have been omitted from our experiments. Researchers from Facebook [33] released graph embeddings trained on the Wikidata knowledge graph. These embeddings can be utilized to calculate similarities between entities. However, contextual knowledge is not considered during the computation of embeddings. Recently, researchers have employed neural networks and deep-learning to capture similar patterns in unstructured text. EquatorNLP [27] is an approach that combines deep natural language processing and advanced machine learning for the task of extracting facts related to disaster response. Another deep-learning approach [3] recognizes mentions of Adverse Drug Reactions (ADR) in social media using knowledge-infused recurrent models. This approach solves the challenge of extracting ADR entities with characteristics including long surface forms, varied, and unconventional descriptions, as compared to more formal medical symptom terminology.

D. COMMUNITY DETECTION ALGORITHMS

Existing community detection approaches have focused on the fundamental problem of grouping nodes in a network in the way that very densely connected nodes are placed in one community. In contrast, nodes in different communities are sparsely connected. As a result, detected communities provide the basis for uncovering connections that would be detected by simply traversing a network. Community detection relies on an objective function that captures the intuition that a community is a set of nodes with better internal connectivity than external connectivity. The exact optimization of this objective function is typically NP-hard. Heuristic-based e.g., METIS [34] and semEP [35] and approximation algorithms e.g., [36], [37], [38] aim at identifying sets of nodes with good values of the objective function and that can

be understood good communities [39]. METIS [34] generates a graph partition guided by a heuristic that aims at creating a coarse graph whose size is within a small factor of the size of the final partition obtained after multilevel refinement. This process is conducted in three stages: coarsening, partition of the coarsest graph, and refinement, implemented following multilevel and multi-constraint partitioning schemes to scale up to large graphs; this makes METIS a suitable method for detecting communities in large graphs like those processed in PINYON. SemEP [35] is a graph partitioning method that identifies a minimal partition of a weighted bipartite graph; it is guided for optimizing an objective function that combines the values of similarity among the nodes in a community with the density of the connections. The problem of graph partitioning implemented by SemEP is matched to the Vertex Coloring Problem. Experimental studies show that this encoding enables the detection of high-quality communities in real-world graphs of various topologies [40], [41], [42]. Built upon these results, PINYON formalizes the problem of finding related posts as a problem of community detection. PINYON resembles SemEP and resorts to well-known heuristic-based algorithms –e.g., DSATUR [11] and Welsh Powell [43]– for the Vertex Coloring Problem and semantic similarity measures to identify context-aware communities. However, since PINYON is agnostic of community detection techniques, METIS is also evaluated as a potential implementation. The results of our experimental studies indicate the empirical advantage of our approach against METIS and Welsh Powell.

E. GRAPH COLORING ALGORITHMS

The field of graph coloring is relatively well studied, spanning over the last few decades [43]. It is one of the most studied NP-HARD problem in computer science [44]. Initial approaches for graph coloring such as Welsh Powell [43] introduced vertex coloring algorithm. The vertex coloring algorithm finds application in various computer science subfields such as scheduling [45], frequency assignment [46] and communication network [47]. We also point readers to the following literature related to the application of vertex coloring problem in some information extraction and link discovery tasks [35], [48], [49], [50]. Our idea in this paper is to map the studied problem of recommending semantically related posts to a graph coloring problem, which has not been done in the literature thus far. Hence, to do so, we revert too few of the fundamental approaches in vertex coloring problem such as DSATUR [11] and extend them to our use-case. We are not using the latest vertex coloring algorithms because: 1) DSATUR is one of the most fundamental algorithms in vertex coloring. Proving its extensibility to our studied problem solves purpose of application of graph coloring algorithm for unveiling semantically related posts. 2) In recent works of vertex coloring algorithms, it is a usual practice to compare proposed algorithm to DSATUR to show empirical effectiveness [51], [52]. 3) Our focus in this work is not to empirically compare hundreds of already proposed vertex coloring algorithms, instead to show if vertex coloring can be a solution for the studied problem in this paper.

III. MOTIVATING EXAMPLE

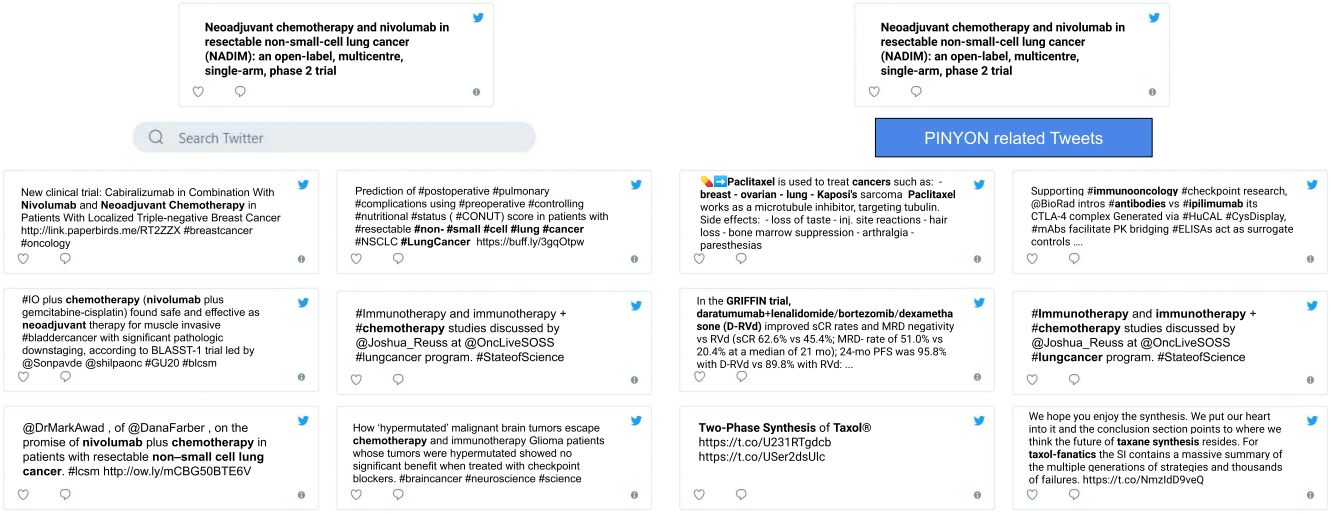
We motivate our work by presenting a post with revolutionary news about a novel treatment that can increase survival probability in patients with non-small-cell lung cancer (NSCLC) in stage IIIA. NSCLC is terminal in most patients with advanced stages of the disease. Effective interventions with the potential of increasing the median survival are celebrated by the patients, their families, and oncologists. The post refers to a scientific article by Provencio et al. [53] published in The Lancet Oncology,[5] one of the most prestigious venues in medicine.[6] The post was announced on Twitter from the account of the first author of the work[7] on September 25th, 2020. Since then, it has captured the attention of the scientific community, resulting in 43 Retweets, 10 Quote Tweets, 91 Likes. The novelty of the announced treatment relies on the promising results of assessing antitumor activity when neoadjuvant therapies are applied. They combine chemotherapy drugs (Paclitaxel and Carboplatin) plus an immunological drug (Nivolumab) before surgery. Then, this treatment is followed by adjuvant intravenous monotherapy (also with Nivolumab) for one year. The evaluation was conducted in a cohort of 46 patients who received neoadjuvant therapy. 41 (89%) of 46 patients had surgery, and at the time of the publication, these patients were alive and free of recurrence, with a median follow-up of 24.0 months. The authors claim the novelty of assessing the effectiveness of neoadjuvant Chemoimmunotherapy in NSCLC patients in stage IIIA. More importantly, the referred results support the hypothesis about the efficacy of neoadjuvant Nivolumab and platinum-based Chemotherapies, and potentially represents a paradigm shift in treating lung cancer patients with advance stages of the disease.

Figure 1 presents the original tweet; given the relevance of the announced results, it is of great interest to retrieve posts that announce similar results either in lung cancer or in other types of cancers. Social media platforms (e.g., Twitter) provide searching capabilities by indexing relevant entities in the post or using user annotations like hashtags. Figure 1a presents six tweets identified by Twitter API for two search criteria, *"Neoadjuvant chemotherapy and nivolumab"* and *"resectable non-small-cell lung cancer"*. As expected, the output includes 1) tweets discussing the prescription of chemotherapy and nivolumab in other cancers (e.g., breast, bladder, or brain tumors), or 2) outcomes of non-small-cell lung cancer treatments. Despite these results' relevance, they only include posts whose text comprises at least one of the keywords in the search criteria because the search depends on the hashtags and keywords (annotation-based) mentioned in

[5]https://www.thelancet.com/journals/lanonc/home

[6]A follow-up version of this work have been presented in ASCO2022 https://meetings.asco.org/abstracts-presentations/207335

[7]https://twitter.com/MARIANOPROVENCI/status/1309355589676-535810



(a) Twitter Search (b) PINYON related Tweets

FIGURE 1. Motivating Example. A tweet announcing the promising results of combining Neoadjuvant chemotherapy and Nivolumab in patients with Non-small-cell lung cancer in stage IIIA. a) Related tweets resulting from searching using the Twitter API the criteria *Neoadjuvant chemotherapy and Nivolumab* and *Resectable non-small-cell lung cancer*. b) Results of semantically related tweets based on PINYON. Tweets retrieved from Twitter include at least one of the keywords in the input post. Tweets identified by PINYON are semantically related to the input post based on the shared context.

the tweet. Following such search strategy affects the richness of the knowledge encoded in the retrieved posts because it is only limited within the keywords and hashtags without considering any semantics of the entities mentioned in a post. Contrary, PINYON exploits knowledge encoded in the context of the input post. This knowledge composes PINYON background knowledge. It is extracted from existing encyclopedic knowledge graphs like DBpedia or Wikidata, and domain-specific ones like Unified Medical Language System (UMLS). Furthermore, PINYON employs a graph coloring algorithm to identify semantically similar communities of posts for a relevant context. As a result, PINYON can identify the tweets that exactly match the terms in the input text (Figure 1b). More importantly, it outputs tweets that refer to treatments that combine other novel oncological treatments (e.g., GRIFFIN trials, synthesis of Taxol, or immunoncology) and are used to treat other various diseases (e.g., sarcoma and breast, ovarian, and lung cancer). Given the wealth of information present in posts announced in social media, the possibility of retrieving the posts semantically related to a given announcement, represents a fundamental change in post recommendation towards more informative and meaningful outcomes.

IV. PROBLEM STATEMENT AND SOLUTION

This section formally describes the problem statement and the approach implemented in PINYON. Please refer to Table 1 for better understanding of the used annotation.

A. PRELIMINARIES

1) THE VERTEX COLORING PROBLEM

The vertex coloring problem corresponds to the coloring of the vertices in a graph $G=(V,J)$ with the minimal number of colors such that adjacent vertices are colored with distinct colors. Formally, let $\mathcal{SC}$ be a set of colors and $\mu(.)$ is a mapping from V to $\mathcal{SC}$. The function $\mu(.)$ is a solution to the vertex coloring problem for G if $\mu(.)$ is defined as follows:

- Adjacent vertices are in distinct colors, i.e., if v_i and $v_j \in V$ are adjacent then, $\mu(v_i) \neq \mu(v_j)$.
- Number of colors in $\mu(.)$ is minimized, i.e., the optimization objective is formally defined as follows $\arg\min_{\mathcal{USC}\subseteq\mathcal{SC}} \mid \{\mu(n_i)/n_i \in V \wedge \mu(n_i) \in \mathcal{USC}\} \mid$

Lemma 4.1: Let $G=(V,J)$ be a graph. Let $\mathcal{SC}$ be a set of colors available to color $G=(V,J)$ and $\mu(.)$ be a mapping from V to $\mathcal{SC}$ that corresponds to a solution to the vertex coloring problem for $G=(V,J)$. Let v, $deg_{G(v)}$, and $\mu(v)$ be a vertex in J, the degree of v in G, and $\mu(v)$ the color of v, respectively. The node v is the only vertex in V with the color $\mu(v)$ if and only if $deg_{G(v)}$ is equal to $|J|$-1.

Proof of 4.1 is presented in Appendix I-A.

2) LINE, COMPLEMENT, AND BIPARTITE GRAPHS

a: LINE GRAPH

Given a graph $G=(V,J)$ such that $J \subseteq V \times V$, the *line graph* $LP(G)=(F,T)$ of G comprises a) a vertex f_{e_q} in F per each edge e_q in J, and b) an edge $(f_{e_q} f_{e_k})$ in T if e_q and $e_k \in J$ and share a vertex in common, i.e., the following edges belong to J: $e_q = (v_i, v_z)$ and $e_k = (v_j, v_z)$, or $e_q = (v_z, v_i)$ and $e_k = (v_z, v_j)$.

b: COMPLEMENT GRAPH

Given a graph $G=(V,J)$, the *complement graph* of G is a graph $Comp(G)=(V,K)$, where vertices of G and $Comp(G)$ are the same, and K is the complement of J, i.e., $K=((V\times V)-J)$.

TABLE 1. Summary of PINYON Notation.

Notation	Explanation
BP=($V_1 \cup V_2$,E)	Bipartite Graph, $V_1 \cap V_2=\emptyset$, $E \subseteq V_1 \times V_2$
LP(G)=(F,T)	*LP(G)* is a graph, and it is the *line graph* of *G=(V,J)*, $\|J\|=\|F\|$, $\|T\|=\|\{(e_q,e_k)/e_q,e_k \in J \wedge (e_q=(v_i,v_z) \wedge e_k=(v_z,v_j)) \vee (e_q=(v_z,v_i) \wedge e_k=(v_z,v_j))\}\|$
Comp(G)=(V,K)	*Comp(G)* is a graph. It is the complement graph of *G=(V,J)* iff *K=((V× V)-J)*
$\mathcal{SC}$	Colors available to color a graph *G=(V,J)* using a function $\mu(.)$ from V to $\mathcal{SC}$. $\mathcal{USC}$ subset of $\mathcal{SC}$ used in $\mu(.)$
$\mathcal{P}$	An input post expressed in terms of entities (words or tokens) annotated with terms from various contexts
$\mathcal{R}$	A database of posts p also described with a set p_e=$\{en_1,\ldots,en_m\}$ of entities
$\mathcal{C}$	A set of contexts c modeled as contextual knowledge graphs (KGs)
PostRelated($\mathcal{P}$,$\mathcal{R}'$,c)	Metric for the semantic relatedness of the posts in $\mathcal{R}'$ and $\mathcal{P}$ in the context c
$\mathcal{P}_c(\mathcal{R})$	Partition of $\mathcal{R}$, i.e., grouping of the elements into non-empty subsets, where every element is in exactly one subset
PostRelated($\mathcal{P}$,$\mathcal{P}_c(\mathcal{R})$,c)	Overall value of *PostRelated($\mathcal{P}$,$\mathcal{R}'$,c)* for all the $\mathcal{R}'$ in $\mathcal{P}_c(\mathcal{R})$ by using a triangular norm
$\delta(p,e_i,c)$	For entity e_i in p_e and a context c, returns the set of terms in c
$e_{i,j}=(t_i,t_j)$	t_i and t_j and terms, and $e_{i,j}$ is an edge in *BP=($V_1 \cup V_2$,E)* and a vertex in *LP(BP)=(F,T)*
$\rho_\gamma(e_{i,j},e_{z,q})$	Quantifies similarity of terms t_i,t_j,t_z, and t_q based on a similarity measure γ
$\mathcal{SP}_c(\mathcal{R})$	All the possible partitions of $\mathcal{R}$
ComS(.)	Overall relatedness among the terms in the vertices colored with the same color
ColoredSimilarity($\mu(.)$)	Aggregates *ComS(.)* for all colors in $\mu(.)$
Degree of similarity $v(e_{i,j})$	Aggregates values of the similarity of $e_{i,j}$ with the rest of other vertices
SCom	*PINYON-CACD* communities
CAC_c	A context-aware community for a context c
GComS(.)	The Global community similarity is defined using a triangular norm among the community similary of the communities in *CAC*

c: BIPARTITE GRAPH

A *bipartite graph BP=($V_1 \cup V_2$,E)* comprises vertices in $V_1 \cup V_2$ and edges are in $E \subseteq V_1 \times V_2$; the intersection of V_1 and V_2 is empty, i.e., $V_1 \cap V_2 = \emptyset$.

B. THE DSATUR ALGORITHM

The Vertex Coloring Problem is NP-hard [11], and various approximate algorithms have been proposed to provide efficient solutions to tractable instances of the problem [54].

DSATUR [11] employs an algorithm that colors each vertex of the graph once, using a heuristic to select the colors. Assuming that we have a graph $G = (V, E)$, DSATUR dynamically orders the vertices in V depending on the number of different colors appointed to the adjacent vertices of each vertex in V, i.e., vertices are picked based on the degree of saturation on the partial coloring of the graph created so far; only adjacent vertices that are already colored are taken into account. Intuitively, choosing a vertex with the highest degree of saturation enables one to color first those vertices with more restrictions and smaller sets of colors available. Ties are broken depending on the maximum vertex degree of the tied vertices, i.e., the number of neighboring nodes colored or not; DSATUR has a time complexity of $O(|V^3|)$. Furthermore, the optimality requirements of the suggested algorithms have attracted attention in past years; features of the graphs that are difficult to color, in terms of time complexity, for each algorithm [12]. Thus, DSATUR colors most k-colorable graphs optimally, i.e., k is the number of ideal colors, G is k-colorable, and $UsedColors(G) \leq k$. The propositions described in Appendix I-B list graphs for which DSATUR is optimum [55].

C. PROBLEM STATEMENT

Given an input post $\mathcal{P}$, a dataset $\mathcal{R}$ of posts, and a set of contexts $\mathcal{C}$, we tackle the problem of *identifying the minimal groups of posts* in $\mathcal{R}$ that maximize the *context-aware relatedness* to $\mathcal{P}$ according to each of the contexts c in $\mathcal{C}$. Formally, assume that given a subset $\mathcal{R}'$ of $\mathcal{R}$, and a context c from $\mathcal{C}$, *PostRelated($\mathcal{P}$,$\mathcal{R}'$,c)* is a metric that quantifies the semantic relatedness of the posts in $\mathcal{R}'$ and $\mathcal{P}$ in the context c. A solution to the *context-aware post recommendation* problem (aka *CWPR*) corresponds to a partition $\mathcal{P}_c(\mathcal{R})$ of $\mathcal{R}$ such that the overall value of *PostRelated($\mathcal{P}$,$\mathcal{R}'$,c)* over all the parts $\mathcal{R}'$ of $\mathcal{P}_c(\mathcal{R})$ is maximized while the number of parts in $\mathcal{P}_c(\mathcal{R})$ is minimized. Suppose *PostRelated($\mathcal{P}$,$\mathcal{P}_c(\mathcal{R})$,c)* corresponds to the overall value of relatedness and is computed by combining the values of *PostRelated($\mathcal{P}$,$\mathcal{R}'$,c)* for all the $\mathcal{R}'$ in $\mathcal{P}_c(\mathcal{R})$ by using a triangular norm (aka t-norm) $\mathcal{T}(.,.)$ [56]. Formally, if $\mathcal{SP}_c(\mathcal{R})$ represents all the possible partitions of $\mathcal{R}$, a solution for *CWPR* is a $\mathcal{P}_c(\mathcal{R})$ with minimal cardinality

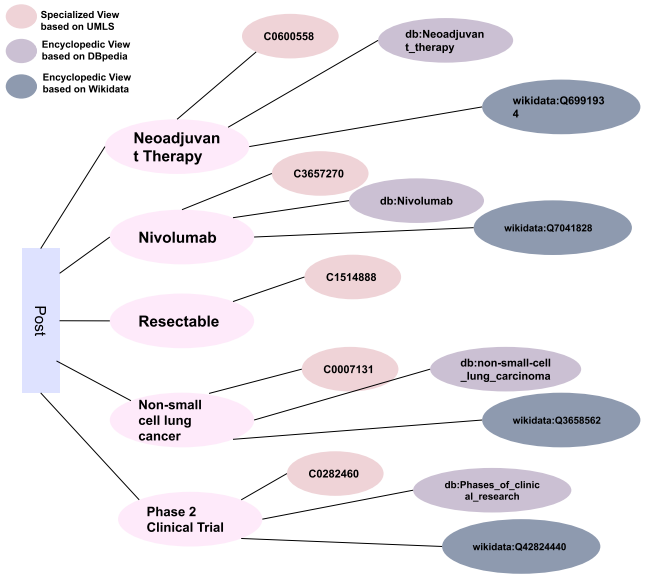


FIGURE 2. Running Example. Entities in a post are annotated with concepts in various contexts (e.g., domain-specific or encyclopedic) derived from the knowledge graphs.

and satisfying the condition:

$$\arg\max_{\mathcal{P}_c(\mathcal{R}) \subseteq \mathcal{SP}_c(\mathcal{R})} PostRelated(\mathcal{P}, \mathcal{P}_c(\mathcal{R}), c)$$

D. PROPOSED SOLUTION

The proposed solution (PINYON) resorts to existing knowledge bases (e.g., DBpedia, Wikidata, and UMLS) to represent encyclopedic or domain-specific contexts. Moreover, our proposed solution assumes that the input post $\mathcal{P}$ and the posts in $\mathcal{R}$ are annotated with the terms of these knowledge bases to encode context-aware knowledge expressed in a post. Formally, a post p is described in terms of a set p_e of words or tokens that represent entities, i.e., $p_e = \{en_1, \ldots, en_m\}$ indicates that p is expressed in terms of the m entities in p_e. Additionally, p is associated with a function $\delta(p, e_i, c)$ that, for each entity e_i in p_e and a context c, returns the set of terms in the knowledge base c that represent e_i in the context modeled by c. To illustrate, Figure 2 presents the five entities in p_e in the post in the motivating example Figure 1, and the terms from DBpedia, Wikidata, and UMLS that represent these entities. Similarly, the posts in $\mathcal{R}$ are expressed in terms of entities and terms from contextual knowledge bases. A bipartite graph $BP=(V_1 \cup V_2,E)$ embodies relatedness between the terms that describe the input post $\mathcal{P}$ and the posts in $\mathcal{R}$. BP is defined as follows:

- V_1 represents the contextual description of $\mathcal{P}$, i.e., V_1 is equal to terms in the union of $\delta(\mathcal{P}, en_i, c)$, for all the entities en_i in $\mathcal{P}_e = \{en_1, \ldots, en_m\}$ and the knowledge bases c in $\mathcal{C}$.
- V_2 corresponds to the contextual description of posts in $\mathcal{R}$, i.e., V_2 comprises all the terms in $\delta(p, en_i, c)$ for all p in $\mathcal{R}$, its entities in p_e and contexts in $\mathcal{C}$.
- E encodes context-aware relatedness among $\mathcal{P}$ and $\mathcal{R}$ and edges in E meet the following conditions:
  - Associate terms in the same context, i.e., if (t_i, t_j) belongs to E then, t_i and t_j are terms of the same knowledge base.
  - Relate terms semantically similar, i.e., if (t_i, t_j) belongs to E then, for a given similarity measure γ, $\gamma(t_i, t_j)$ is equal or greater than a given threshold ϵ.

A partition $\mathcal{P}_c(\mathcal{R})$ is computed from a partition of the edges in BP formulated as a solution of the Vertex Coloring Problem (VC). The mapping of $CWPR$ to VC is defined as follows.

1) BIPARTITE GRAPH TRANSFORMATION

The bipartite graph BP for $\mathcal{P}$ is transformed into an undirected line graph $LP(BP)=(F,T)$ as follows: 1) Edges representing context-aware relationships between terms in BP are modeled as vertices in $LP(BP)$, i.e., there exists a vertex $e_{i,j}$ in F iff there exists an edge $e_{i,j} = (t_i, t_j)$ in E. 2) Co-occurrence of a term in several posts modeled as terms that appear in several edges in BP. For each pair of different edges $e_{i,j} = (t_i, t_j)$ and $e_{z,q} = (t_z, t_q)$ in BP, such as $t_i = t_z$ or $t_j = t_q$, there is an edge between the vertices $e_{i,j}$ and $e_{z,q}$ in $LP(BP)$.

The degree of a vertex $e_{i,j}$ in $LP(BP)$ that represent the edge $e_{i,j} = (t_i, t_j)$ in BP is equal to $\binom{out(t_i)}{2} + \binom{in(t_j)}{2}$, where $out(t_i)$ and $in(t_j)$ represent the out-degree and in-degree of the vertices in BP, respectively.

2) COMPLEMENT GRAPH CREATION

The complement graph of $LP(BP)=(F,T)$ corresponds to an undirected graph $Comp(LP(BP))=(F,K)$ with the same vertices of $LP(BP)$ and with the complement edges of $LP(BP)$, i.e., $K = (F \times F) - T$. The degree of a vertex $e_{i,j}$ in $Comp(LP(BP))$ that represent the edge $e_{i,j} = (t_i, t_j)$ in BP is equal to $\binom{|V_2|-out(t_i)}{2} + \binom{|V_1|-in(t_j)}{2}$, where $|V|$ corresponds to the cardinality of a set V. Moreover, given the edges $e_{i,j} = (t_i, t_j)$ and $e_{z,q} = (t_z, t_q)$ in BP, a function $\rho_\gamma(e_{i,j}, e_{z,q})$ quantifies the similarity of the terms t_i,t_j,t_z, and t_q based on a similarity measure γ; $\rho_\gamma(e_{i,j}, e_{z,q})$ corresponds to the result of applying a t-norm $\mathcal{T}(\gamma(t_i, t_z), \gamma(t_j, t_q))$.

3) FINDING COMMUNITIES OF SEMANTICALLY RELATED POSTS

The Vertex Coloring Problem is solved over $Comp(LP(BP))$. However, the concept of number of colors used during the coloring of the vertices of $Comp(LP(BP))$ is redefined to ensure that the color assignment both minimizes the numbers of colors and the overall value of relatedness among the terms in the vertices colored with the same color (i.e., $ComS(.)$). Thus, the number of colors in $\mathcal{USC}$ for a solution $\mu(.)$ of the vertex coloring of $Comp(LP(BP))$ corresponds to 1-$ColoredSimilarity(\mu(.))$, where $ColoredSimilarity(\mu(.))$ is defined as follows:

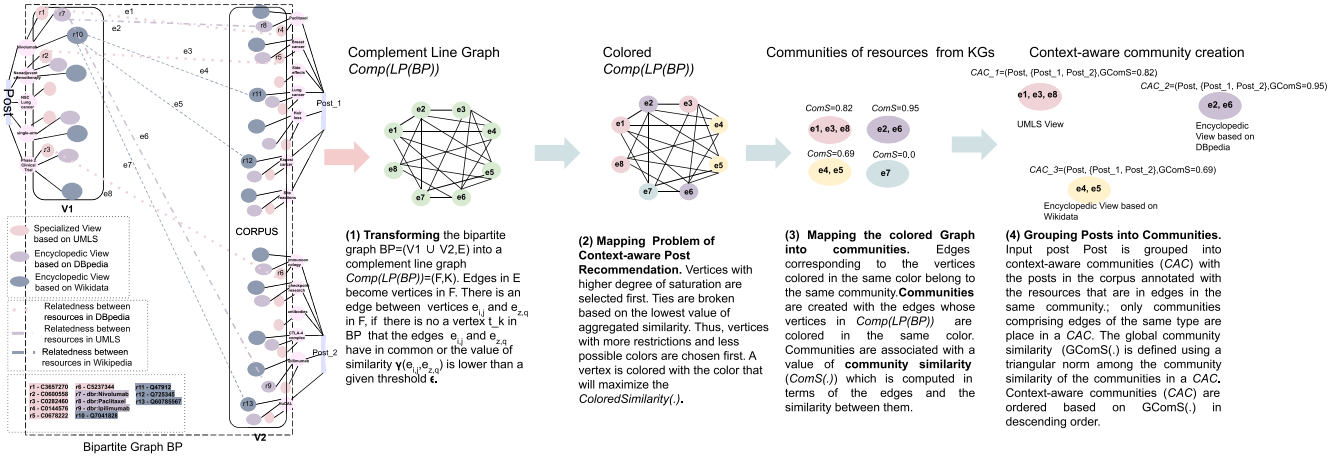


FIGURE 3. Running Example. Mapping the Problem of Context-Aware Post Recommendation into the Vertex Coloring Problem. Entities in a post are annotated with terms in different contexts (e.g., domain-specific or encyclopedic), relatedness between the terms in the same context are represented by edges of a Bipartite Graph $BP=(V_1 \cup V_2,E)$; a similarity measure quantifies relatedness of the terms in a context. (1) The complement of the line graph of BP is computed $Comp(LP(BP))=(F,K)$. (2) $Comp(LP(BP))=(F,K)$ is colored with the minimal number of colors that maximize the value of $ColoredSimilarity(.)$, (3) Colored Graph in mapped into Communities. (4) Communities are used to create Context-Aware Communities of Posts.

- Let $\mathcal{P}(F)$ be the partition of the vertices of $Comp(LP(BP))$ such that all the vertices in one part of $\mathcal{P}(F)$ are colored with the same color in $\mu(.)$, and each part is colored in a different color. We call each part pa a community, and the community similarity $ComS(pa)$ corresponds to the results of applying a t-norm $\mathcal{T}$ over all the unordered pair pairs $(e_{i,j}, e_{z,q})$ of $\rho_\gamma(e_{i,j}, e_{z,q})$. Note that a pair $(e_{i,j}, e_{z,q})$ is considered unordered because $\rho_\gamma(.,.)$ is assumed to be symmetric.
- $ColoredSimilarity(\mu(.))$ aggregates the values of the community similarity $ComS(.)$ of all communities in $\mathcal{P}(F)$. A t-norm or the average can be used to compute this aggregated value.

4) THE PINYON-CACD SOLUTION

PINYON-CACD is used to color $Comp(LP(BP))$. It follows a heuristic to color the vertices and meet the condition of minimizing the number of used colors $\mathcal{USC}$ as defined previously; it resembles the coloring heuristic proposed by Palma et al. [35] and implemented by the DSATUR algorithm. In addition to the degree of saturation, each vertex is associated with a *degree of similarity*, which represents how much similar the vertex is with respect to the rest of the vertices in the graph. The *degree of similarity* of a vertex $e_{i,j}$, aka $\upsilon(e_{i,j})$, is computed as the aggregated value of the similarity of $e_{i,j}$ with the rest of the vertices; the aggregation function can be the average (e.g., arithmetic or geometric mean) or a triangular norm.

5) PINYON-CACD IN A NUTSHELL

Intuitively, a vertex with a high degree of similarity is highly similar to many of the vertices in the graph. Vertices are ordered based on degree, and the one with the maximal degree is chosen first. In case of ties, i.e., at least two vertices have the same degree, the one with the lowest value of similarity is selected first. Thus, *PINYON-CACD* starts coloring the vertex *with more restrictions*. Then, *PINYON-CACD* iteratively traverses the list of ordered vertices and chooses the one which the highest degree of saturation and in case of ties, the vertex with the lowest *degree of similarity* is chosen. This decision also enables *PINYON-CACD* to select the vertex with fewer options to be colored. Once a vertex is chosen, a color is selected; it is among the suitable colors the one that maximizes the $ColoredSimilarity(.)$ of the assignment $\mu(.)$ of colors created so far. *PINYON-CACD* finalizes when all the vertices are colored; it creates a community per used color. Each community comprises all the vertices colored with the same color and is described in terms of the community similarity $ComS(.)$.

6) RUNNING EXAMPLE-PINYON-CACD

Figure 3 illustrates in the steps (1), (2), and (3). **Step (1):** a bipartite graph $BP=(V_1 \cup V_2,E)$ is transformed into a complement line graph $Comp(LP(BP))$. Thresholds of similarity values are utilized to decide when two entities are similar or not; entities collected from different contexts are not similar. In settings with unrelated entities, $Comp(LP(BP))$ can include numerous edges, i.e., it may be a complete graph. **Step (2):** $Comp(LP(BP))$ is colored. The color assignment optimality depends on the topology of the $Comp(LP(BP))$. Lemma 4.1 and propositions in subsection IV-B state the graph topologies where the coloring is optimal. Thus, the resulting partition of the edges in $Comp(LP(BP))$ maximizes the value of $ColoredSimilarity(.)$. **Step (3):** the colored $Comp(LP(BP))$ is utilized to generate communities; a community only comprises edges connecting highly similar terms from the same context. Each community pa is associated with $ComS(pa)$; Figure 3 illustrates four partitions resulting from

the execution of steps (1), (2), and (3). **Step (4):** posts are grouped into context-aware communities *(CACs)* composed of the input post $\mathcal{P}$ and the posts in $\mathcal{R}$ that are highly related to $\mathcal{P}$ in a given context c. Note that *PostRelated*($\mathcal{P}$,$\mathcal{P}_c$($\mathcal{R}$),c) corresponds to the result of applying an aggregation function over all the values of *ComS(pa)*, where *pa* is a community of edges associating terms of the context c.

7) CONTEXT-AWARE COMMUNITIES

A context-aware community for a context c, CAC_c, is defined inductively as follows:

8) BASE CASE SIMPLE COMMUNITY $CAC_c(\{cc\})$

For $cc = \{e_1, \ldots, e_m\}$ in *SCom*, $CAC_c(\{cc\}) = (\mathcal{P}, \mathcal{SP}, \text{ComS(cc)})$, where $\mathcal{SP}$ corresponds to the posts in $\mathcal{R}$ annotated with at least one $t_{i,k}$ from an edge $e_i = (t_{i,j}, t_{i,k})$ in cc.

9) INDUCTIVE CASE COMPOSED COMMUNITY $CAC_c(C_{r,s})$

$CAC_c(C_{r,s})$=($\mathcal{P}$,$\mathcal{SP}$,GComS($C_{r,s}$)) is created from
$CAC_c(C_r)$=($\mathcal{P}$,$\mathcal{SP}$,GComS(C_r)) and
$CAC_c(C_s)$=($\mathcal{P}$,$\mathcal{SP}$,GComS(C_s)),
where $C_{r,s}$=$C_r \cup C_s$ and $GComS(C_{r,s})$ is the aggregated value of $GComS(C_r)$ and $GComS(C_s)$.

10) RUNNING EXAMPLE-PINYON-CACD (Cont.)

Figure 3 illustrates, in step (4), the context-aware communities created from the running example. Three communities are created, one per context, i.e., CAC_1 for UMLS, CAC_2 for DBpedia, and CAC_3 for Wikidata. Note that describing each community CAC_c, where $c \in$ {UMLS, DBPedia, Wikidata}, based on the communities created by *PINYON-CACD* enables the traceability of the whole process of post recommendation. This is a unique feature of *PINYON* that cannot be achieved with any of the baselines for post recommendation included in the empirical study.

V. THE PINYON APPROACH

This section describes the techniques that implement the proposed solution reported in Section IV-D. Figure 4 depicts the components of the pipeline for retrieving semantically related posts. The pipeline comprises, first, the phase of encoding where the CORPUS and the input post are annotated and represented as a complement line graph *Comp(LP(BP))=(F,K)*. Second, communities of posts are detected to retrieve the related posts in the decoding phase.

A. CONTEXT-AWARE POST ENCODING

The step of Context-Aware Post Encoding comprises the components of Context-Aware Corpus Annotation, Context-Aware Post Annotation, and Bipartite Graph Creation. Given an input post $\mathcal{P}$, a dataset $\mathcal{R}$ of posts, and a set of contexts $\mathcal{C}$, annotations for the input post and the corpus are created. The created annotations are utilized to build the Bipartite Graph, which is transformed to a complement line graph.

1) CONTEXT-AWARE CORPUS ANNOTATION

The dataset of posts is annotated by identifying first the entities for each post in $\mathcal{R}$. Any Named Entity Recognition & Linking tool could be used in this step (e.g., TagMe [57], Falcon [58], and DBpedia Spotlight [59]). However, the current PINYON implementation resorts to two versions of Falcon [58] for performing this task. This decision is supported on the experimental results reported by Sakor et al. [58], [60], which show that Falcon outperforms these state-of-the-art engines in the existing benchmarks. Falcon 2.0 [60] identifies entities in a short text and links the recognized entities to DBpedia and Wikidata knowledge graphs. BioFalcon[8] recognizes and links entities in a short text to UMLS. Once the annotation is completed, a dictionary of the recognized entities is created where the posts' ids for each entity's mention are stored; the dictionary is used during the last step of the PINYON approach to retrieve posts with a specific entity mention. The Context-Aware Corpus component is computed once-for-all, and there is no need to perform it again for new input posts. All the remaining components of the approach pipeline have to be computed again for each new input post.

2) CONTEXT-AWARE POST ANNOTATION

As in the previous step, the input post $\mathcal{P}$ is annotated by recognizing the entities in the input post then linking the recognized entities to each corresponding KG in $\mathcal{C}$. The same Named Entity Recognition & Linking tools are utilized.

3) BIPARTITE GRAPH CREATION

During this step, a bipartite graph $BP=(V_1 \cup V_2,E)$ is created. The BP graph embodies relatedness between the terms that describe the input post $\mathcal{P}$ and the posts in $\mathcal{R}$. The BP graph is formed by creating edges between all the entities in $\mathcal{P}$ and each post's entities in $\mathcal{R}$ (Figure 3). The created bipartite graph is transformed into a Complement Line Graph *Comp(LP(BP))*=(F,K). Edges in the BP graph become vertices in the complement line graph and there is an edge between vertices $e_{i,j}$ and $e_{z,q}$ if there is no a vertex t_k in BP that the edges $e_{i,j}$ and $e_{z,q}$ have in common or the value of similarity $y(e_{i,j}, e_{z,q})$ is lower than a given threshold ϵ. The similarity between the edges is the average value of similarity between all the pairs of entities that form the edges. This process is similarity metric-agnostic. However, in the current PINYON implementation, the similarity between two entities is computed based on the cosine similarity between the vectors (embedding) representing the entities. Two embedding techniques are considered, RDF2Vec [10] and CUI2Vec [61]. RDF2Vec is utilized to retrieve embedding for entities in DBpedia and Wikidata KGs, while CUI2Vec is used to create entity embeddings for entities in UMLS.

B. CONTEXT-AWARE POST DECODING

This step comprises the components of Context-Aware Community Creation, Context-Aware Aggregated Relatedness,

[8]https://labs.tib.eu/sdm/biofalcon/

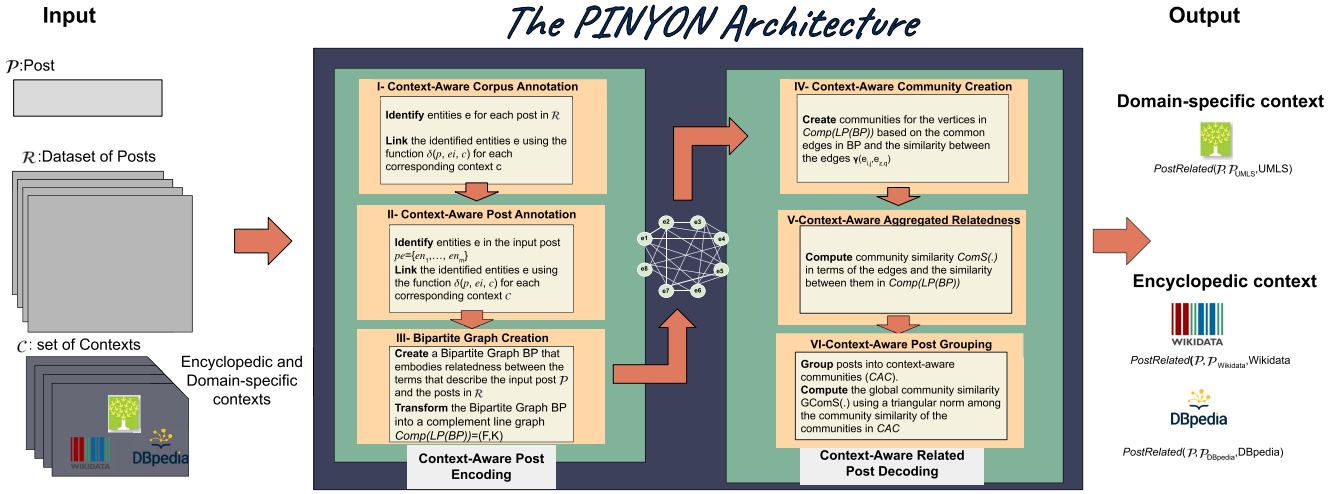


FIGURE 4. **The PINYON approach architecture. The pipeline receives an input post $\mathcal{P}$, a dataset $\mathcal{R}$ of posts, and a set of contexts $\mathcal{C}$ (e.g., DBpedia, Wikidata, UMLS) and outputs related posts in the corresponding contexts *PostRelated*($\mathcal{P}$,$\mathcal{P}_{\mathcal{C}}(\mathcal{R})$,$c$). During the encoding phase, the input post $\mathcal{P}$ and the Corpus $\mathcal{R}$ are annotated with terms from the provided contexts $\mathcal{C}$. The annotations are used to build the Bipartite Graph BP; then the Bipartite Graph is transformed to Complement Line Graph *Comp(LP(BP))=(F,K)*. The complement line graph is utilized to create context-aware communities during the decoding phase. The posts are grouped into the created communities, and community similarity is calculated to determine the related posts.**

Algorithm 1 PINYON-CACD

1: **Input:** graph G=(F,K)
2: **Output:** vertices color c(v): $v \in$ C
3: **Begin**
4: C:=∅; U:=F; Compute $deg_{G(u)}$
5: **For each** $v \in$ F
6: **if** $deg_{(v)}$ == $|F|$-1
7: **then** $c(v) = c'$; where c' is unassigned color
8: select one uncolored vertex v randomly with $\max_{v \in U}\{deg_G U(v)\}$; ties are broken based on the lowest value of aggregated similarity
9: $c(v) := 1$; $C := C \cup \{v\}$; $U := U \setminus \{v\}$
10: update $CACD_{G(C}$ and $deg_{G(U)}$
11: **repeat**
12: find an uncolored vertex v with $max_{v \in U}\{CACD_{G(C)}(v)\}$
13: **if** a subset U' of multiple vertices with the same max degree of saturation is found
14: **then** select one uncolored vertex v randomly with $max_{v \in U'}\{deg_{G(U)}(v)\}$; ties are broken based on the lowest value of aggregated similarity
15: find the least possible color k that can color the selected vertex v
16: $c(v) := k$; $C := C \cup \{v\}$; $U := U \setminus \{v\}$; update $CACD_{G(C)}$ and $deg_{G(U)}$
17: **until** U=∅
18: **End**

and Context-Aware Post Grouping. Given a complement line graph *Comp(LP(BP))*=(F,K), communities of edges are created using different community detection techniques. The posts are grouped into the created communities, and a community similarity is calculated to determine the related posts.

1) CONTEXT-AWARE COMMUNITY CREATION

During this step, communities of edges are created. This component could be implemented by any community detection approach, e.g., METIS [34] or SemEP [35]. However, PINYON maps the problem of context-aware community detection to the graph vertex coloring problem. Algorithm1 sketches the details of *PINYON-CACD*.

a: CREATING A REDUCED COMPLEMENT LINE GRAPH

PINYON-CACD receives *Comp(LP(BP))*=(F,K) and follows 4.1 to reduce the size of the graph to be colored. Thus, *PINYON-CACD* first identifies all the nodes that meet this condition, and assigns to each one a new color; no other node will be colored with these assigned colors. We have observed that numerous vertices meet this condition. As a result, a smaller portion of the original complement line graph is colored following the DSATUR heuristic. Thus, as shown in Algorithm1 Lines 5-7, for each vertex v in F, such as $deg_{G(v)}$ is equal to $|F|-1$, a new color $c(v)$ is assigned.

Furthermore, if this *reduced* complement line graph has one of the topologies presented in the propositions in Section IV-B, *PINYON-CACD* generates an optimal coloring. Thus, *PINYON-CACD* generates a mapping $\mu(.)$ that corresponds to a solution to the vertex coloring problem for *reduced* complement line graph. Vertices in the *reduced* complement line graph with a higher degree of saturation, but that have not been colored yet, are selected first.

b: COLORING A REDUCED COMPLEMENT LINE GRAPH

PINYON-CACD orders the vertices in F, which have not been colored so far, dynamically based on the number of

different colors assigned to the adjacent vertices of each vertex in F, i.e., the vertices are chosen based on the degree of saturation on the partial coloring of the graph built so far, and only colored adjacent vertices are considered. The degree of saturation represents the number of different colors used in the neighbor vertices of a vertex. Intuitively, selecting a vertex with the *maximum degree of saturation* allows for coloring first the vertices with more restrictions and for which there are a smaller number of available colors. Ties are broken based on the lowest value of aggregated similarity (Lines 8-9). Thus, vertices with more restrictions and less possible colors are chosen first. A vertex is colored with the color that maximizes the Colored Similarity. This process is repeated until all the vertices are colored (Lines 11-17).

c: CREATING COMMUNITIES

Once the graph coloring is done, the colored graph is mapped into communities. Edges corresponding to the vertices colored in the same color belong to the same community. Communities are created with the edges whose vertices in *Comp(LP(BP))* have the same color.

d: PINYON-CACD TIME COMPLEXITY

The worse-case time complexity of *PINYON-CACD* is $O(|F'|^2)$, where F' is the subset of F without the nodes v with $deg_{G(v)}$ equal to $|F|-1$.

e: PINYON AGNOSTIC-SOLVER IMPLEMENTATION

PINYON approach is also agnostic of the technique used for coloring the graph, i.e., any graph coloring algorithms could be utilized. For example, Welsh and Powell graph coloring algorithm [43]. In our ablation study (Section VI), we replace the *PINYON-CACD* technique with Welsh&Powel. Additionally, we use METIS to study the effect of the PINYON technique for computing Context-Aware Communities.

2) CONTEXT-AWARE AGGREGATED RELATEDNESS

In this step, a community similarity (*ComS(.)*) value is computed for each community from the previous step. The *ComS(.)* is computed in terms of the edges and their similarity. As in the Bipartite Graph Creation step, the entities' embedding is utilized to compute the similarity between two entities in order to compute the similarity between two edges. However, in this step, the community similarity is computed among all the pairs of edges inside the community.

3) CONTEXT-AWARE POST GROUPING

During this step, the input post is grouped into context-aware communities (CAC) with the posts in the corpus annotated with the resources that form edges in the same community. The dictionary created in the Context-Aware Corpus Annotation step is utilized to retrieve posts annotated with the entities that comprise the edges. The global community similarity (*GComS(.)*) is defined using a triangular norm among the community similarity of the communities in a CAC. CAC are ordered based on *GComS(.)* in descending order. The posts annotated with the highest number of entities that exist in a single community(considering ordering the communities based on the global community similarity) are selected first as related posts.

C. AGNOSTIC OF THE PINYON ARCHITECTURE

The PINYON approach is agnostic of the studied social media platform. Twitter is just a use-case for our experiments. The PINYON approach is agnostic of the community detection technique. Any community detection technique can be used for our approach. The PINYON approach is agnostic of the post annotation tools. Any tool that ensures identifying entities in a short text and supports the input context can be used. The PINYON approach is also agnostic of the entities similarity metric. Any entities' embedding that supports the input contexts can be used.

VI. EXPERIMENTAL STUDY

We study the following research questions: **RQ1**) How does knowledge represented in public KGs empower PINYON to outcome semantically related posts? **RQ2**) What is the effect of the contextual description on the accuracy of retrieving related posts? **RQ3**) How does the specificity of the entities in the input texts influence the accuracy of retrieving related posts?.

A. EXPERIMENTAL CONFIGURATION

1) BASELINES

Our study includes various types of baselines; they depend on the particular to be analyzed. In all experiments, we pass the same input fed to our approach to all the baselines.

a: BASELINE FOR RECOMMENDATING POSTS

We include various BERT models in the study; they are used to discover the posts in a CORPUS related to an input post. Our rationale to include language-model based baselines is: these language models are trained on large corpus and consists of domain-specific contextual knowledge when fine-tuned later on specific data. Therefore, they become natural choice to compare efficacy of our approach in recommending semantically similar posts. The various baseline models include: Sentence-BERT [5], BERTweet [26], COVID-Twitter-BERT [16], BioBERT-NLI [15], and CovidBERT-NLI.[9] These models are used as follows. Posts in the CORPUS and the input post, are short sentences and BERT models, especially Sentence-BERT, determine semantic textual similarity across these sentences [5]. Sentence-BERT is a modification of the pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity [5].

The other baselines are BERT-based methods trained over tweets, tweets related to COVID-19, or literature in the biomedical domain. Since the training data is similar to our

[9]https://huggingface.co/gsarti/covidbert-nli

corpus, we also compare with BERTweet, COVID-Twitter-BERT, BioBERT-NLI, and CovidBERT-NLI.

BERTweet is a large-scale language model pre-trained for English Tweets. COVID-Twitter-BERT is a transformer-based model pre-trained on a large corpus of Twitter messages on the topic of COVID-19. BioBERT-NLI is a BioBERT model fine-tuned on the Stanford natural language inference corpus [62] and the MultiNLI [63] datasets using the sentence-transformers library [5] to produce universal sentence embeddings. CovidBERT-NLI is the model CovidBERT trained by DeepSet on AllenAI's CORD19 Dataset[10] of scientific articles about COVID-19. CovidBERT-NLI uses the original BERT vocabulary and was subsequently fine-tuned on the SNLI and the MultiNLI datasets using the sentence-transformers library.

In addition to the BERT models, Twitter Search API [64] is studied. The comparison with the results of the Twitter Search API allow us to assess the accuracy of post retrieval whenever exact keywords or hashtags of the input query.

b: BASELINE FOR COMMUNITY DETECTION

The communities detected by *PINYON-CACD* are compared with the results generated by METIS, ''one of the state-of-the-art community detection approaches for large networks [65]''.

c: BASELINE FOR THE VERTEX COLORING PROBLEM

The quality and efficiency of the graph coloring generated by *PINYON-CACD* are compared to high-performance method Welsh and Powell [43]. In the experiments, the Welsh and Powell graph coloring algorithm implements the CACD graph coloring of the Context-Aware Related Post Decoding step.

2) BENCHMARKS

PINYON and the baselines are studied over two benchmarks of datasets comprising texts (sentences) of tweets.

a: POSTS ABOUT COVID-19

A dataset of tweets (TweetsCOV19) [66] is employed in experiment 1. It contains tweets in the English language for May 2020 [66], which is the month when the initial scientific studies related to COVID-19 analysis started coming out [67]. The total number of tweets utilized is 1,922,405 tweets. The tweets contain both scientific and general tweets, while both are related to COVID-19. For each tweet, we extract, link, and retrieve the corresponding embedding vectors of KG entities using RDF2Vec [10]. However, when UMLS is used as the underlying KG, CUI2Vec [61] is used to create entity embeddings because UMLS is not an RDF triplestore. The total number of extracted entities is 252,245. A considerable amount of tweets do not include entities that can be linked to existing KGs. Thus, the number of entities is lower than the number of tweets.

[10]https://pages.semanticscholar.org/coronavirus-research

b: POSTS ABOUT THE WORLD CUP FINAL

The second experiment focuses on generalizability. The second experiment employs a dataset of tweets [68]; it contains a random collection of 521,802 tweets starting from the 16th round until the World Cup Final that took place on July 15th, 2018. Each record in the dataset contains a tweet about the 2018 World Cup Final, including football players and their teams' names. The football players' names are considered as entities. We use the state-of-the-art entity recognition and linking tools – e.g., DBpedia Spotlight [59], TagMe [57], and Falcon 2.0 [60] – to recognize entities from the tweets and link them to DBpedia and Wikidata KGs. The primary task in this experiment is to retrieve tweets related to *La Liga* football players' from the tweets in the FIFA dataset by providing a tweet related to *La Liga* football players as an input. We choose *La Liga* topic because during the time frame of the tweets in the FIFA dataset there was much interest about *La Liga* topic.[11] In order to build the gold standard for the second experiment, we collect all the *La Liga* football players' names from different sources,[12] then link the players' names to their corresponding URIs in DBpedia and Wikidata using the previously mentioned entity recognition and linking tools. In our experiments, we consider as the gold standard of relevant tweets, the FIFA dataset of tweets which contains mentions of *La Liga* football players.

3) METRICS

We measure the performance of PINYON performance in terms of the accuracy of recommended related posts, the quality of the communities, and the solver execution time.

a: MEASURING RECOMMENDATION PERFORMANCE

We report the performance using the standard metrics of **Precision (P)**, **Recall (R)**, and **F-Score (F)** when the gold standard is available. On the other hand, **Accuracy** is utilized when we lack information about which tweets are relevant for an input tweet.

- **Precision (P)** is the fraction of *relevant posts* among the *retrieved posts*.
- **Recall (R)** is the fraction of *relevant posts* that have been retrieved over the *total amount of relevant posts*.
- **F-Score** is the *harmonic mean* of **P** and **R**.
- **Accuracy** is the fraction of *related posts* among all the top k *studied posts*.

b: QUALITY OF THE DETECTED COMMUNITIES

For studying the quality of the computed communities, we utilize the metrics defined by the research community to measure the quality of a network community [69], [70], [71]. The metrics are **Conductance**, **Coverage**, **Modularity**, **Performance**, and **Total Cut**. Let $Q = \{C_1, \ldots C_n\}$ be the set of communities obtained by PINYON:

[11]shorturl.at/qFPW1
[12]https://en.wikipedia.org/wiki/Category:La_Liga_players

- **Conductance**: measures relatedness of entities in a community and how different they are to entities outside the community [72]. It is computed as the ratio between the number of edges inside a community and the number of edges leaving the community. Thus, conductance is considered the simplest notion of a community quality, based on the intuition that a good network community comprises nodes that have better internal than external connectivity. Conductance is a lower is better metric, and we report the inverse of the conductance $1 - Conductance(S)$.
- **Coverage**: compares the fraction of intra-community similarities among entities to the sum of all similarities among entities [72].
- **Modularity**: is the value of the intra-community similarities among the entities divided by the sum of all the similarities among the entities, minus the sum of the similarities among the entities in different communities, in the case they were randomly distributed in the communities [73]. It measures the strength of a community partition according to the degree distribution. The value of the modularity lies in the range $[-0.5, 1]$, which can be scaled to $[0, 1]$ by computing $\frac{Modularity(Q)+0.5}{1.5}$.
- **Performance**: sums the number of intra-community relationships, plus the number of non-existent relationships between communities [72]. A large value of modularity indicates a good community structure.
- **Total Cut**: sums all similarities among entities in different communities [74]. Values of total cut are normalized by dividing by the sum of the similarities among the entities; inverse values are reported, i.e., $1 - NormTotalCut(Q)$.

c: PERFORMANCE OF THE SOLVERS

Execution time is defined as the elapsed time required to perform the Context-Aware Related Post Decoding step. It is measured as the absolute wall-clock system time, as reported by the `time` command of the Linux operating system.

4) IMPLEMENTATION

PINYON is implemented in Python 3.6. We run experiments on a server equipped with 96 cores and 900GB RAM running Ubuntu 18.04. All the resources used in the reported experimental study are publicly available.[13]

B. EXPERIMENT 1- MEDICAL DOMAIN

This experiment studies PINYON performance on a large dataset of tweets [66] related to the medical domain.

1) EXPERIMENT SETUP

PINYON is executed against three different configurations of the background knowledge (i.e., DBpedia, Wikidata, and UMLS). Moreover, an ablation study is conducted by utilizing METIS and Welsh Powell solvers. We fix a threshold ($\epsilon = 0.50$ from Section IV) to build the communities. The threshold is determined by running the 10-fold cross-validation to choose the best setting.

a: INPUT POSTS

Eight different tweets related to two topics (Lung Cancer and COVID-19) are used as testbeds for the experiment. Building testbeds for tweet relatedness is not our contribution, and we inherit the testbed settings from [20], [31], and [32]. The eight tweets are described in Table 2. These tweets are selected based on the following parameters: three Medical doctors were asked to provide 24 (each doctor with eight tweets) tweets related to COVID-19 and Lung Cancer. A fourth medical doctor specializing in Lung Cancer recommended four tweets related to the topic. A fifth doctor specializing in general Medicine recommended four tweets related to COVID-19 to be part of the testbed.

b: GOLD STANDARD

Due to the lack of test datasets for tweets in the medical domain, we asked six medical doctors (with at least the degree of general Medicine) to evaluate the experiment's performance. The medical doctors were asked to determine if the retrieved tweets are relevant with respect to the input tweet; medical doctors received tweets after anonymizing the approach name. These Medical doctors are not the same as the ones who selected the tweets in the testbed. *A tweet should be marked as relevant by Medical doctors if it satisfies the following criteria*: (a) fits the topic of the input tweet, (b) confirms the information in the input tweet or, (c) contains entities that are relevant to the entities in the input tweet, d) if a drug is mentioned in the input tweet, then all the occurrences of similar drugs appeared in the related tweets should be used for the same medical prescription. It is important to note that although the number of tweets in a testbed is eight, the corresponding tweets in the dataset are in millions (one-to-many mapping). Hence, in the ideal case, the solution implemented in PINYON for the vertex coloring problem should remove a lot of noise (irrelevant graph nodes representing the tweets in the communities) before finalizing the related communities of posts.

2) PERFORMANCE OF THE SOLVERS FOR DISCOVERING COMMUNITIES

a: SIZE OF THE COMPLEMENT LINE GRAPH PER INPUT TWEET

The eight input tweets induce large and complex complement line graphs, which impose challenges for the whole process of related post recommendation. Table 3 reports the size of the complement line graph *Comp(LP(BP))* generated during the Context-Aware Post Encoding for each of the tweets presented in Table 2. As we can see in Table 3 the numbers of the generated nodes and edges are huge; an average of 178,160 nodes and 20,117,886,682 edges. However, built upon Lemma 4.1, *PINYON-CACD* executes

[13] https://github.com/SDM-TIB/PINYON

TABLE 2. Overview of the tweets used for Experiment 1.

#	Tweet	Topic
T1	Neoadjuvant chemotherapy and nivolumab in resectable non-small-cell lung cancer (NADIM): an open-label, multicentre, single-arm, phase 2 trial	Lung Cancer
T2	#oncoalert Exciting data from NADIM for neoadjuvant chemo-IO in resectable NSCLC. All stage IIIA (74% N2) with 57% pathCR and 77.1% 2yr PFS in mod-ITT pop. Highly anticipate ph3 trial results and utility of pathCR (not just MPR!!) as predictive surrogate biomarker for survival!	Lung Cancer
T3	Interesting neoadjuvant trial of chemo with #immunotherapy that enrolled 46 resectable stage IIIA NSCLC patients. At 24 months, PFS was 77·1% (95% CI 59·9–87·7). Approach was not associated with surgery delays	Lung Cancer
T4	This is quite remarkable, look forward to seeing randomized neoadjuvant chemo-IO results. Definitely the most promising area for resectable NSCLC right now!	Lung Cancer
T5	Remdesivir, the only antiviral drug authorized for treatment of COVID-19 in the United States, fails to prevent deaths among patients, according to a study of more than 11,000 people in 30 countries sponsored by the World Health Organization.	COVID-19
T6	We now have evidence that an inexpensive drug, #Fluvoxamine, may be effective in preventing patients with mild #COVID19 cases from developing severe complications. Learn more and help us continue the research:	COVID-19
T7	Greater access to Remdesivir through the @NHFJamaica for the treatment of Covid-19 infection. However, please note that clinical trials for this drug are still ongoing in other countries.	COVID-19
T8	A new study from the World Health Organization has found that remdesivir — one of the anti-viral drugs that has been touted as a potential treatment for COVID-19 since — has "little or no effect" on COVID patients' chances of survival	COVID-19

TABLE 3. Size of the graphs *(*Comp(LP(BP))) generated during the Context-Aware Post Encoding for each of the Tweets in Experiment 1.

Tweets	Nodes	Edges	Nodes Pruned by PINYON-CACD	Pruning rate
T1	197895	20,162,431,025	5,645,480,687	72%
T2	205934	26,408,812,356	6,602,203,089	75%
T3	162389	15,370,187,321	3,074,037,464	80%
T4	140392	11,709,913,664	3,630,073,235	69%
T5	173628	21,146,682,384	5,498,137,419	74%
T6	184937	24,201,693,969	5,808,406,552	76%
T7	172865	18,882,308,225	5,475,869,385	71%
T8	187246	23,061,064,516	7,379,540,645	68%

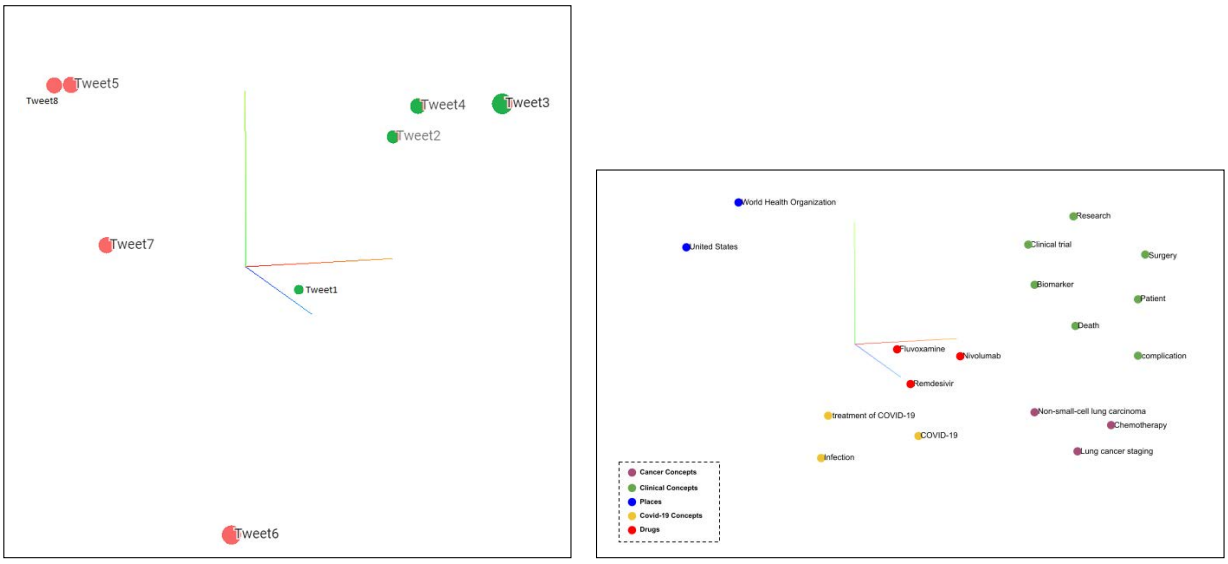
the Context-Aware Related Post Decoding step efficiently; it can identify the colors assigned to exclusively one vertex. Thus, *PINYON-CACD* prunes the complement line graph before starting coloring, and the number of nodes and edges are reduced considerably (68-80%). As a result, *PINYON-CACD* minimizes the execution time needed for the Context-Aware Related Post Decoding compared to the other solvers (i.e., METIS, Welsh Powell) Figure 8. The pruning rate is related to how much noise is presented in the tweet's text. For example, Tweet2 and Tweet3 look very similar utilizing the mentioned entities. However, the entities *months* and *surgery* in Tweet3 add more noise than Tweet2 making the pruning rate of Tweet3 higher than Tweet2.

b: VECTOR REPRESENTATION OF INPUT POSTS

To have a more precise understanding of the properties of our approach (mapping the problem of Context-Aware Post Recommendation into the Vertex Coloring Problem) and the BERT models baselines, we visualize the embedding for each tweet using Sentence-BERT (Figure 5a) and the embedding for each entity mentioned in a tweet using CUI2Vec (Figure 5b). As we can observe in Figure 5a, Tweet6 and Tweet7 are far in the embedding space from Tweet5 and Tweet8 even though they share the same entities and are related to the same topic (COVID-19); the same observation is applied between Tweet1 and the remaining Lung Cancer related tweets. On the other hand, we can observe in Figure 5b that the entities mentioned in the tweets are grouped together in the embedding space powering the *PINYON-CACD* approach finding semantically related posts based on the similarity between entities mentioned in a tweet without being affected by the noise presented in the tweet's text as the *PINYON-CACD* remove such noise while pruning the complement line graph.

c: TIME PERFORMANCE OF THE CONTEXT-AWARE RELATED POST DECODING STEP

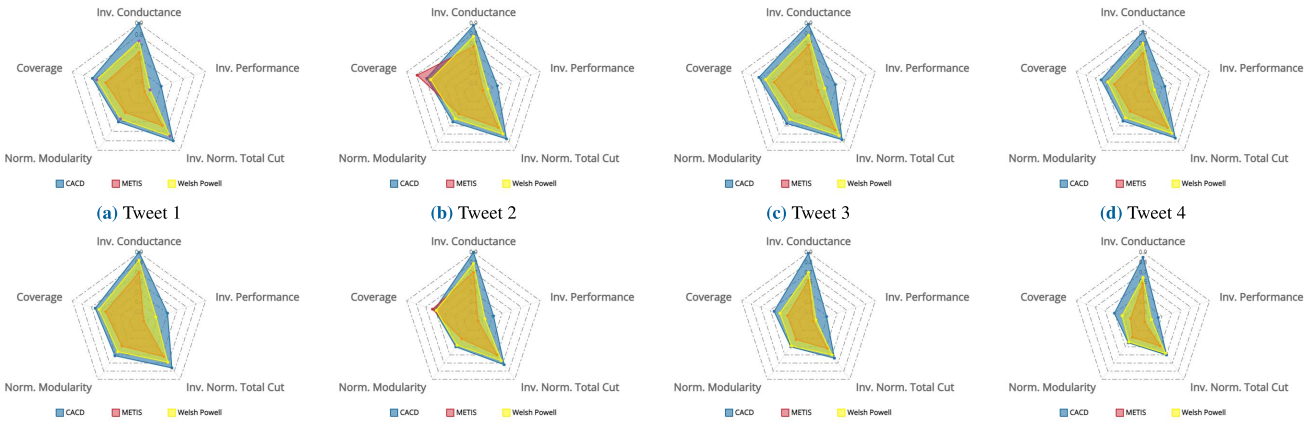
We also studied the time required for the Context-Aware Related Post Decoding by executing Experiment1 but with a different number of tweets in the studied corpus; we exclude the time required for The Context-Aware Post Encoding since



(a) Tweets Embedding Visualization

(b) Entities Embedding Visualization

FIGURE 5. Comparison of Text- and Entity-based Embeddings of Biomedical Tweets.



(a) Tweet 1 (b) Tweet 2 (c) Tweet 3 (d) Tweet 4

(e) Tweet 5 (f) Tweet 6 (g) Tweet 7 (h) Tweet 8

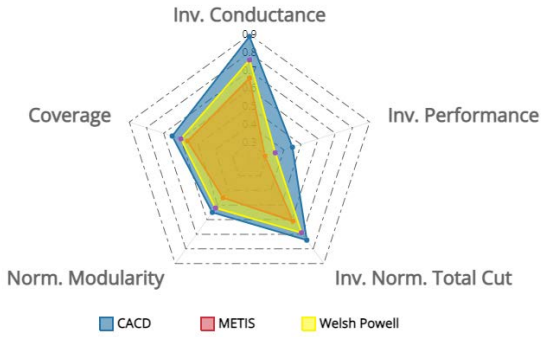
FIGURE 6. Quality of the computed communities. Communities evaluated in terms of five metrics (higher values are better); Communities for the eight tweets in VI-B are reported. *PINYON-CACD* exhibits the best performance.

it is similar for all the solvers. As we can observe in Figure 8, the execution time for PINYON-METIS and PINYON-Welsh-Powell increases considerably w.r.t the size of the studied corpus (840-86k seconds). In contrast, *PINYON-CACD* reports a much lower execution time (7-14k seconds) over the same graph because of the graph pruning step. PINYON-DSATUR (the *PINYON-CACD* approach without the graph pruning step) reports the highest execution time since the size of the graph to be colored is much higher than the pruned graph in the case of *PINYON-CACD* (68-80% difference).

3) QUALITY OF THE DISCOVERED COMMUNITIES

a: COMMUNITIES' QUALITY

Figure 6 shows the quality of the generated communities for the tweets in Table 2. Communities for each tweet are computed using three solvers; CACD, METIS, and Welsh Powell. The communities generated by PINYON include closely related posts in all tweets. However, *PINYON-CACD* exhibits higher quality in terms of the five community-based metrics. These results corroborate our hypothesis that *PINYON-CACD* is able to group together entities into communities that are highly related and provide an explanation for the results reported in Table 4. Further, we can observe that the communities for the tweets related to Lung Cancer (Tweet 1-4) are of higher quality than those related to COVID-19 (Tweet 5-8). The reason behind such higher communities' quality is the richness of the BK for terms related to Lung Cancer compared to COVID-19. Tweet5 and Tweet8 have in common several entities that have different types. But, the mention of `United States` in Tweet5 enables to recognize the resource `COVID-19 pandemic in the United States,Q83873577` which provides contextual knowledge to facilitate the discovery of highly related tweets in

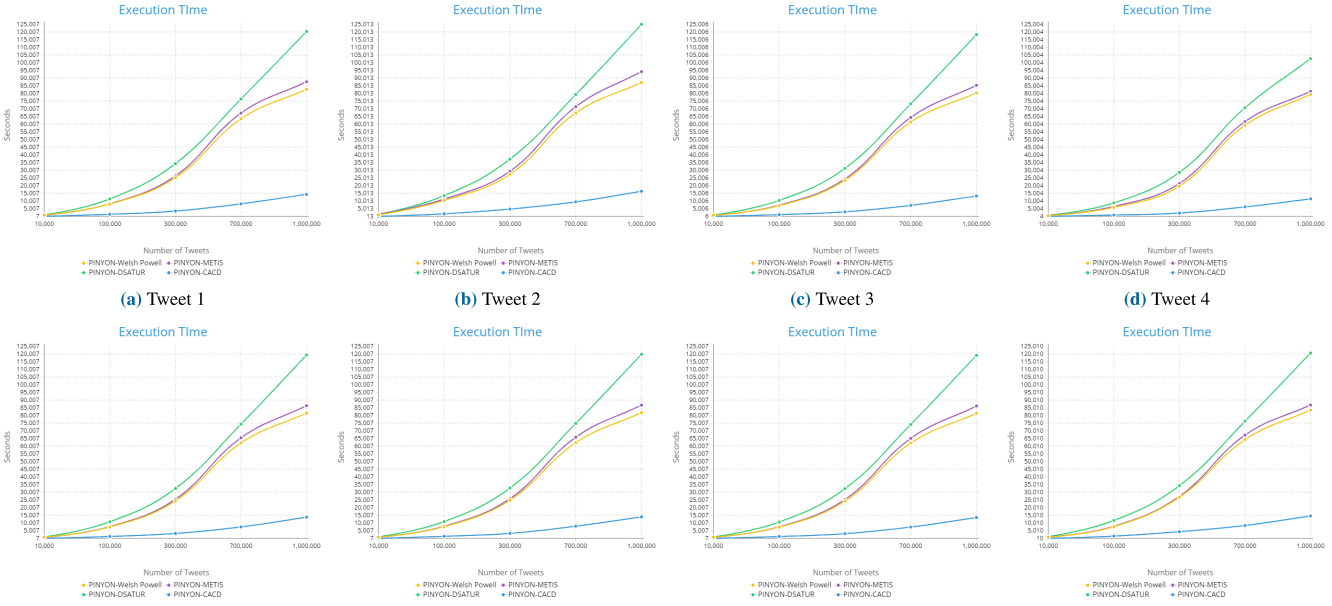


(a) Aggregated values for the communities quality metrics (average values for each metric) across all the tweets.

	Solver					
	DSATUR		METIS		Welsh Powell	
Metric	Avg	SD	Avg	SD	Avg	SD
Inv. Conductance	0.89	0.02	0.66	0.04	0.76	0.06
Inv. Performance	0.45	0.06	0.29	0.05	0.35	0.05
Inv. Norm. Total Cut	0.74	0.08	0.61	0.07	0.69	0.07
Norm. Modularity	0.55	0.06	0.45	0.05	0.52	0.05
Coverage	0.65	0.09	0.56	0.14	0.60	0.09

(b) Aggregated values for the community quality metrics. The average (Avg) and standard deviation (SD) are computed for each metric across all the tweets for the three solvers

FIGURE 7. Report of the Aggregated values for the community quality metrics.



(a) Tweet 1 (b) Tweet 2 (c) Tweet 3 (d) Tweet 4

(e) Tweet 5 (f) Tweet 6 (g) Tweet 7 (h) Tweet 8

FIGURE 8. Context-Aware Related Post Decoding execution time for each tweet in Experiment 1.

the corpus. As a result, the values of performance, modularity, and coverage in Tweet5 are higher.

b: AVERAGE COMMUNITIES' QUALITY

For understanding better the differences between the solvers, we computed the average and standard deviation (SD) for each metric used to measure the quality of the communities across all the tweets. From Figure 7b and Figure 7a We can observe that the aggregated values for the *PINYON-CACD* and Welsh Powell solvers are higher than METIS solver, which confirms the results reported in Table 4.

4) PERFORMANCE OF PINYON

The Accuracy metric is used to measure the performance of the task of retrieving semantically related tweets(top 40 tweets). The results of this experiment are reported in Table 4. A 2-fold cross-validation method was implemented while evaluating the results, and a majority voting to solve any disagreements. The results in Table 4 propose that PINYON-CACD-UMLS employing only the knowledge in UMLS KG as background knowledge for the approach outperforms all the baselines and the other configuration of PINYON. The reason behind the good performance is the richness of UMLS KG in the medical domain (14,608,809 different entities in the medical domain), which empowers the PINYON approach to retrieve related tweets in the studied topics (Lung Cancer and COVID-19). Additionally, *PINYON-CACD* have been customized to select the communities following heuristics that ensure that very related entities are put together in the same community (answering **RQ2** & **RQ3**). Twitter Search API reports poor accuracy and performs only exact string matching. Therefore, we passed the entities mentioned

in the input tweet to the search API instead of the entire tweet content to get better results. Nonetheless, the reported accuracy of the API is not improved because of the strategy of matching hashtags and keywords without considering any semantics [64]. Note that METIS and Welsh Powell exhibit a competitive behavior, but as off-the-shelf components their algorithms are not customized to ensure that the overall relatedness among all the entities included in the same community is maximized. The reported accuracy for the baseline Sentence-BERT suggests that the specific domains as the medical domain represents a challenge even for pre-trained BERT models. BioBERT-NLI and CovidBERT-NLI report higher accuracy since they are fine-tuned over medical corpus. BERTweet is trained over 850M English Tweets (5M Tweets related to COVID-19). Still, we can observe a drop in the accuracy for the COVID-19 tweets because of its inability to capture the underlying semantic meaning of entities, which PINYON successfully does. COVID-Twitter-BERT is trained over 97M tweets related to the topic COVID-19. Thus, COVID-Twitter-BERT reports the second-highest accuracy for COVID-19 tweets. A noticeable drop in accuracy for the baselines is observed for the tweets associated with the COVID-19 topic. This drop is the lack of labeled training data for the studied topic (COVID-19), which is always the case in new pandemics. Contrary, the richness of the PINYON background knowledge in the medical domain using drugs and diseases mentions prevents the drop in the accuracy of our approach for both topics (answering **RQ1** successfully).

C. EXPERIMENT 2-SPORT DOMAIN

This experiment aims to evaluate PINYON generalization. The experiment aims to accommodate understanding of PINYON's behavior in a domain that is different from the medical domain (Sport domain).

a: TWEETS RELATED TO LA LIGA CHAMPIONSHIP-GOLD STANDARD

This experiment aims to study the performance of PINYON on a tweets dataset from the sports domain to validate general domain performance. The dataset of tweets from the 2018 World Cup Final (Section VI-A2) is utilized. Considering the domain of this experiment is general, we choose Sentence-BERT and BERTweet as the baselines. Additionally, we study the effect of the different features of the utilized knowledge graphs in the background knowledge. PINYON is executed against two different configurations of the background knowledge (i.e., DBpedia, Wikidata). Furthermore, two different thresholds for building the communities are used (i.e., 0.50 and 0.65); the same configuration from the previous experiment for the ablation study is applied.

PINYON Performance: Average results of five runs are reported in Table 5 (different random input post from the gold standard); the best results are in bold; the average number of nodes, edges, and pruning rate of the complement line graph are 135,984, 12,294,158 and 68% accordingly; the average execution time for *PINYON-CACD* is 45,367 seconds. The accuracy of the retrieved relevant tweets is measured by precision (P), recall (R), and F-Score (F). Table 5 reports the results. We observe an increase in the performance of the baselines compared to experiment 1. The explanation for this increase is because of the generalizability of the studied benchmark, which matches the data used to train these models. PINYON-CACD-Wikidata benefits from the knowledge collected in a community-maintained knowledge graph, where predicates are proposed and then get accepted only if these additions meet specific criteria. Contrary, DBpedia data is automatically extracted from publicly available Wikipedia dumps, which may affect data quality as, in consequence, negatively impact on the studied solvers. Furthermore, the results reveal the importance of the value threshold that represents when two entities can be considered similar. The thresholds of 0.5 and 0.65 correspond to the percentiles 45 and 88 in the distribution of the values of similarities between the entities in the La Liga Tweets. As a result, when the threshold $\epsilon = 0.65$ is evaluated, 88% of the combinations of entities are considered not similar, enabling all the approaches to increase the precision. Particularly, in the case of *PINYON-CACD*, the value of $\epsilon = 0.65$ prevents that all this pair of entities are never placed together in the same community. Since the recall is not improved as the precision, a more precise method for tuning this threshold is required. The results enable us to answer **RQ1** and **RQ2** in the sports domain. We conclude that the richness of the contextual description encoded in the background knowledge and the features of the entities' mentions in the tweets affect the quality of the solutions.

D. DISCUSSION

PINYON's configurations outperform the baselines across all the experimental settings. These results are grounded on the strategy followed by PINYON, where the information encoded in the contextual description is enough to determine the semantically related tweets. Moreover, integrating various KGs into the background knowledge empowers PINYON to capture knowledge from different domains and facilitate the accurate computation of entity and post-relatedness. It is also important to note that PINYON exhibits better performance (Experiment 1) because of the detailed description of the biomedical concepts present in UMLS. Moreover, it is worthy of mentioning that Wikidata is a community-maintained knowledge graph where predicates are proposed and then get accepted only if these additions meet specific criteria. Contrary, DBpedia data is automatically extracted from publicly available Wikipedia dumps, which may affect data quality. Albeit empowering PINYON with the knowledge encoded on Wikipedia, the number of entities compared to Wikidata affects the performance of PINYON-DBpedia as reported in Table 4 and Table 5. Our proposed approach and evaluation provide key insights: 1) Our approach is explainable and interpretable. Mapping the problem of Context-Aware Post Recommendation into the Vertex Coloring Problem permits us to explain the results generated by PINYON. Starting from

TABLE 4. PINYON Performance compared to the baselines for Experiment 1. TweetKB COV19 is the input dataset. The Accuracy metric is reported. Best results are in bold. The columns in Orange are Tweets related to Lung Cancer. The columns in Yellow are tweets related to COVID-19. Three different solvers are utilized for the ablation study; METIS for graph partitioning and community detection, CACD and Welsh Powell for graph coloring. Three different BKs are considered for this experiment.

Baselines		T1	T2	T3	T4	T5	T6	T7	T8
Sentence-BERT		0.58	0.62	0.50	0.41	0.42	0.45	0.39	0.47
BERTweet		0.75	0.69	0.73	0.77	0.62	0.66	0.60	0.63
COVID-Twitter-BERT		0.59	0.57	0.58	0.59	0.73	0.70	0.75	0.71
BioBERT-NLI		0.65	0.68	0.60	0.61	0.50	0.56	0.52	0.56
CovidBERT-NLI		0.54	0.52	0.49	0.55	0.63	0.65	0.60	0.61
Twitter Search API		0.18	0.21	0.15	0.23	0.20	0.13	0.23	0.16
PINYON									
BK	**Engine**								
UMLS	CACD	**0.85**	**0.83**	**0.84**	**0.86**	**0.85**	**0.82**	**0.81**	**0.79**
	METIS	0.82	0.79	0.80	0.82	0.81	0.81	0.76	0.74
	Welsh Powell	0.83	0.81	**0.84**	0.84	0.83	0.80	**0.81**	0.78
DBpedia	CACD	0.79	0.81	0.78	0.82	0.80	0.77	0.79	0.74
	METIS	0.72	0.75	0.70	0.76	0.71	0.69	0.72	0.68
	Welsh Powell	0.75	0.78	0.73	0.80	0.75	0.72	0.74	0.70
Wikidata	CACD	0.82	**0.83**	0.80	0.83	0.82	0.78	**0.81**	**0.79**
	METIS	0.75	0.73	0.71	0.69	0.74	0.67	0.70	0.69
	Welsh Powell	0.79	0.75	0.76	0.78	0.80	0.72	0.79	0.73

TABLE 5. PINYON Performance compared to the Baselines for Experiment 2. FIFA dataset is the input dataset. The reported metrics are Precision, Recall, and F-Score. Best results are in bold. Three different solvers are utilized for the ablation study; METIS for graph partitioning and community detection, CACD and Welsh Powell for graph coloring. Two different BKs are considered.

		Threshold=0.50			Threshold=0.65		
Baselines		**P**	**R**	**F**	**P**	**R**	**F**
Sentence-BERT		0.60	0.75	0.67	0.71	0.68	0.69
BERTweet		0.76	0.80	0.78	0.80	0.74	0.77
PINYON							
BK	**Solver**						
Wikidata	CACD	**0.85**	**0.88**	**0.86**	**0.89**	**0.81**	**0.85**
	METIS	0.79	0.82	0.80	0.85	0.77	0.81
	Welsh Powell	0.83	0.87	0.85	0.87	0.79	0.83
DBpedia	CACD	0.79	0.83	0.81	0.82	0.76	0.79
	METIS	0.72	0.76	0.74	0.78	0.72	0.75
	Welsh Powell	0.77	0.81	0.79	0.80	0.75	0.77

the last step in PINYON (i.e., context-aware community creation), one can trace back in which of the computed communities a particular post appears, (Figure 3). 2) Moreover, PINYON can also find out, based on context description, why a particular post belongs in a community considering the community creation depends on the entities present in the posts. As a result, PINYON corresponds to a white box framework, which allows error tracing while recommending a particular post for a given community.

1) LIMITATIONS

PINYON suffers from the following limitations. First is the constraint to recognize dark entities. Dark entities are entities that do not exist in any knowledge graph [60]. Since PINYON resorts to various knowledge graphs to build its background knowledge, PINYON is not able to extract any knowledge (i.e., contextual description) about these dark entities. Further, the specificity of the entities in the corresponding KGs affects the accuracy of PINYON. Thus, if an entity is miss-linked to a KG class, the embeddings' quality will also be affected. It's important to mention that this limitation is a limitation of the studied corpus and the entity linking tools, and it's not a limitation in our algorithm design. New entities are added regularly to the used community-maintained KGs, enabling the entity linking tools to recognize the newly added entities. So adding more knowledge to the used KGs can overcome these types of limitations. Moreover, the solution

presented in this paper resorts to external components like Falcon2.0, FALCON, TagMe, and DBpedia Spotlight for entity linking. Hence, inheriting the pitfalls of such a tool. For instance, if a tool fails to extract and link a set of entities to the corresponding knowledge graph, PINYON will also be negatively affected. We also ignore edge features (relationship between entities within a tweet), and one possible extension is to study the effect of edge features on our approach.

2) SUCCESS CASES

PINYON overcomes the common problem of lacking training data by depending on the knowledge encoded in its background knowledge to report a high accuracy. The variety of the knowledge graphs used to enrich the approach's background knowledge enables PINYON to perform semantically related posts retrieving with high accuracy, as observed in Section VI. PINYON is agnostic of similarity metrics (calculated using RDF2Vec or CUI2Vec). The lack of training data is observed in the current pandemic of COVID-19. The researchers cannot train the state-of-the-art models for tasks related to health data on the Web. PINYON is a good starting point in such cases, and can also be utilized to build datasets in a specific domain.

VII. CONCLUSION AND FUTURE WORK

Our approach PINYON adopts two novel concepts. First, various KGs are employed as background knowledge; second, a vertex coloring algorithm leverages the extended background knowledge for creating the communities of related posts. Our empirical evaluations provide evidence that the approach outperforms the baselines on several benchmarks in two domains. More importantly, this work has highlighted the importance of capturing background knowledge encoded in existing KGs and the impact that this knowledge has on the tackled problem. Thus, our work broadens the repertoire of knowledge-driven tools for supporting the new generation of data-driven digital technologies. PINYON is a first step towards a larger research agenda, and the social impact (clinical evaluation of results, misinformation on social networking platforms, etc.) of the proposed approach has not been considered in the scope of the paper. In the future, we plan to extend PINYON in the following directions: 1) extending applicability incorporating the credibility of information source and data quality, 2) employing biomedical entity linking tools [75] in PINYON architecture, 3) include domain-specific KGs in RDF in order to compare KBs processed with the same technique, i.e. RDF2vec, and 4) incorporating specific knowledge sources such as NCI Thesaurus [76] to expand the coverage of background knowledge.

APPENDIX I. PROOFS AND PROPOSITIONS

A. LEMMA 4.1

1) PROOF ⇒

By contradiction, suppose v is the only vertex in V with the color $\mu(v)$ and $deg_{G(v)}$ is different to $|J|$-1. Consider the vertex u in V, such that u is different to v, and v and u are not adjacent vertices, i.e., (v, u) is not in J. However, this leads to a contraction because $\mu(v)$ is minimal, and v and u should be colored with the same color.

⇐

By Contradiction, suppose that $deg_{G(v)}$ is equal to $|J| - 1$ and there is a vertex u in V, such that u is different to v and $\mu(v) = \mu(u)$. This leads to a contraction, since v and u are adjacent vertices and cannot be colored with the same color. ■

B. DSATUR PROPOSITIONS

1) PROPOSITION 1-DSATUR OPTIMALITY [12]

Let $G = (V, E)$ be a graph, a core of CG named $CG' = (V', E')$, is a sub-graph of G, i.e., $V' \subseteq V$ and $E' \subseteq E$, and there is no vertex v in V' such that, $degree(v)$ is 1. DSATUR optimally colors G if the core of CG is as follows: a single vertex; a bipartite graph, i.e., a graph that can be partitioned into two set of vertices such that each vertex in each set is connected to a vertex in the other set; A wheel, i.e., a graph formed by connecting one vertex with all the vertices of a cycle; A complete multipartite graph, i.e., a graph in which vertices are adjacent if and only if they belong to different partitions of the graph; a cactus, i.e., a graph in which any pair of cycles has a vertex in common; and a necklace, i.e., a graph made up of r beads where each bead comprises one cycle of length k which is incident with a path of length l.

2) PROPOSITION 2-DSATUR OPTIMALITY [12]

Let $G = (V, E)$ be a graph that corresponds to a polygon tree, i.e., (i) G is a cycle (Base Case), or (ii) G comprises two polygon trees G' and G'' that share exactly one edge. DSATUR optimally colors G.

REFERENCES

[1] S. Munevar, "Unlocking big data for better health," *Nature Biotechnol.*, vol. 35, no. 7, pp. 684–686, Jul. 2017.
[2] R. Y. Zhong, S. T. Newman, G. Q. Huang, and S. Lan, "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives," *Comput. Ind. Eng.*, vol. 101, pp. 572–591, Nov. 2016.
[3] G. Stanovsky, D. Gruhl, and P. Mendes, "Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 142–151.
[4] A. Alsini, A. Datta, and D. Q. Huynh, "On utilizing communities detected from social networks in hashtag recommendation," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 971–982, Aug. 2020.
[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3973–3983.
[6] G. Zhu and C. A. Iglesias, "Sematch: Semantic entity search from knowledge graph," in *Proc. 1st Int. Workshop Summarizing Presenting Entities Ontologies*, vol. 1556, 2015, pp. 1–12.
[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "DBpedia: A nucleus for a web of open data," in *The Semantic Web*. Berlin, Germany: Springer, 2007, pp. 722–735.
[8] D. Vrandecic, "Wikidata: A new platform for collaborative data collection," in *Proc. 21st Int. Conf. Companion World Wide Web*, Lyon, France, 2012, pp. 1063–1064.
[9] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, pp. 267–270, Jan. 2004.

[10] G. Vandewiele, B. Steenwinckel, T. Agozzino, M. Weyns, P. Bonte, F. Ongenae, and F. D. Turck, "pyRDF2Vec: Python implementation and extension of RDF2Vec," IDLab, Tech. Rep., 2020.
[11] D. Brélaz, "New methods to color the vertices of a graph," *Commun. ACM*, vol. 22, no. 4, pp. 251–256, 1979.
[12] R. Janczewski, M. Kubale, K. Manuszewski, and K. Piwakowski, "The smallest hard-to-color graph for algorithm DSATUR," *Discrete Math.*, vol. 236, nos. 1–3, pp. 151–165, Jun. 2001.
[13] M. Hofer, S. Hellmann, M. Dojchinovski, and J. Frey, "The new DBpedia release cycle: Increasing agility and efficiency in knowledge extraction workflows," in *Semantic Systems in the Era of Knowledge Graphs*, E. Blomqvist, P. Groth, V. de Boer, T. Pellegrini, M. Alam, T. Käfer, P. Kieseberg, S. Kirrane, A. Meroño-Peñuela, and H. J. Pandit, Eds. Cham, Switzerland: Springer, 2020, pp. 1–18.
[14] K. Singh, C. Lange, M. E. Vidal, J. Lehmann, S. Auer, A. S. Radhakrishna, A. Both, S. Shekarpour, I. Lytra, R. Usbeck, A. Vyas, A. Khikmatullaev, and D. Punjani, "Why reinvent the wheel: Let's build question answering systems together," in *Proc. World Wide Web Conf.*, 2018, pp. 1247–1256.
[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[16] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," 2020, *arXiv:2005.07503*.
[17] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Twitter-based user modeling for news recommendations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2962–2966.
[18] A. Belhadi, Y. Djenouri, J. C. Lin, and A. Cano, "A data-driven approach for Twitter hashtag recommendation," *IEEE Access*, vol. 8, pp. 79182–79191, 2020.
[19] D. K. Jain, A. Kumar, and V. Sharma, "Tweet recommender model using adaptive neuro-fuzzy inference system," *Future Gener. Comput. Syst.*, vol. 112, pp. 996–1009, Nov. 2020.
[20] R. Krestel, T. Werkmeister, T. P. Wiradarma, and G. Kasneci, "Tweet-recommender: Finding relevant tweets for news articles," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 53–54.
[21] A. Alshammari, S. Kapetanakis, N. Polatidis, R. Evans, and G. Alshammari, "Twitter user modeling based on indirect explicit relationships for personalized recommendations," in *Proc. Int. Conf. Comput. Collective Intell.* Cham, Switzerland: Springer, 2019, pp. 93–105.
[22] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3294–3302.
[23] D. Cer, Y. Yang, S. Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, and B. Strope, "Universal sentence encoder for English," in *Proc. Conf. Empirical Methods Natural Language Process., Syst. Demonstrations*, 2018, pp. 169–174.
[24] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*.
[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[26] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 9–14.
[27] L. Döhling and U. Leser, "Equatornlp: Pattern-based information extraction for disaster response," in *Proc. CEUR Workshop*, vol. 798, Jan. 2011, pp. 127–138.
[28] M. Stevenson and M. A. Greenwood, "Learning information extraction patterns using wordnet," in *Proc. 5th Intl. Conf. Language Resour. Eval. (LREC)*, May 2006, pp. 95–102.
[29] A. Zouaq, D. Gasevic, and M. Hatala, "Linguistic patterns for information extraction in OntoCmaps," in *Proc. WOP*, vol. 929, 2012, pp. 1–12.
[30] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[31] J. P. Carvalho, H. Rosa, and F. Batista, "Detecting relevant tweets in very large tweet collections: The London riots case study," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Naples, Italy, Jul. 2017, pp. 1–6.
[32] T. Noro and T. Tokuda, "Searching for relevant tweets based on topic-related user activities," *J. Web Eng.*, vol. 15, pp. 249–276, Jul. 2016.
[33] A. Lerer, L. Wu, J. Shen, T. Lacroix, L. Wehrstedt, A. Bose, and A. Peysakhovich, "PyTorch-BigGraph: A large-scale graph embedding system," 2019, *arXiv:1903.12287*.
[34] G. Karypis and V. Kumar, "METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices," Tech. Rep., 1997.
[35] G. Palma, M. Vidal, and L. Raschid, "Drug-target interaction prediction using semantic similarity and edge partitioning," in *Proc. Int. Semantic Web Conf.*, 2014, pp. 131–146.
[36] M. T. Hajiaghayi and T. Leighton, "On the max-flow min-cut ratio for directed multicommodity flows," *Theor. Comput. Sci.*, vol. 352, nos. 1–3, pp. 318–321, Mar. 2006.
[37] D. A. Spielman and S.-H. Teng, "Spectral partitioning works: Planar graphs and finite element meshes," in *Proc. 37th Conf. Found. Comput. Sci.*, Burlington, NJ, USA, 1996, pp. 96–105.
[38] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci.*, Berkeley, CA, USA, Oct. 2006, pp. 475–486.
[39] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proc. 19th Int. Conf. World Wide Web*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, Eds. Raleigh, North Carolina, 2010, pp. 631–640.
[40] I. T. Ribón, G. Palma, A. Flores, and M. Vidal, "Considering semantics on the discovery of relations in knowledge graphs," in *Proc. Eur. Knowl. Acquisition Workshop*, Bologna, Italy, 2016, pp. 666–680.
[41] A. Rivas, I. Grangel-González, D. Collarana, J. Lehmann, and M. Vidal, "Unveiling relations in the industry 4.0 standards landscape based on knowledge graph embeddings," in *Proc. Int. Conf. Database Expert Syst. Appl.*, Bratislava, Slovakia, 2020, pp. 179–194.
[42] S. Vahdati, G. Palma, R. J. Nath, C. Lange, S. Auer, and M. Vidal, "Unveiling scholarly communities over knowledge graphs," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, Porto, Portugal, 2018, pp. 103–115.
[43] D. J. A. Welsh and M. B. Powell, "An upper bound for the chromatic number of a graph and its application to timetabling problems," *Comput. J.*, vol. 10, no. 1, pp. 85–86, 1967.
[44] P. Formanowicz and K. Tanas, "A survey of graph coloring-its types, methods and applications," *Found. Comput. Decis. Sci.*, vol. 37, no. 3, p. 223, 2012.
[45] F. T. Leighton, "A graph coloring algorithm for large scheduling problems," *J. Res. Nat. Bureau Standards*, vol. 84, pp. 489–506, Nov. 1979.
[46] A. Gamst, "Some lower bounds for a class of frequency assignment problems," *IEEE Trans. Veh. Technol.*, vol. VC-35, no. 1, pp. 8–14, Feb. 1986.
[47] T.-K. Woo, S. Y. W. Su, and R. Newman-Wolfe, "Resource allocation in a dynamically partitionable bus network using a graph coloring algorithm," *IEEE Trans. Commun.*, vol. 39, no. 12, pp. 1794–1801, Dec. 1991.
[48] A. Anagnostopoulos, L. Becchetti, A. Fazzone, C. Menghini, and C. Schwiegelshohn, "Spectral relaxations and fair densest subgraphs," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 35–44.
[49] A. Adiga, V. Cedeno-Mieles, C. J. Kuhlman, M. V. Marathe, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns, "Inferring probabilistic contagion models over networks using active queries," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2018, pp. 377–386.
[50] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe, "A framework for community identification in dynamic social networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 717–726.
[51] E. Malaguti, M. Monaci, and P. Toth, "An exact approach for the vertex coloring problem," *Discrete Optim.*, vol. 8, no. 2, pp. 174–190, 2011.
[52] R. Lewis, *A Guide to Graph Colouring*, vol. 7. Berlin, Germany: Springer, 2015.
[53] M. Provencio, E. Nadal, A. Insa, M. R. García-Campelo, J. Casal-Rubio, M. Dómine, M. Majem, D. Rodríguez-Abreu, A. Martínez-Martí, J. D. C. Carpeño, and M. Cobo, "Neoadjuvant chemotherapy and nivolumab in resectable non-small-cell lung cancer (NADIM): An open-label, multicentre, single-arm, phase 2 trial," *Lancet Oncol.*, vol. 21, no. 11, pp. 1413–1422, 2020.
[54] P. S. Segundo, "A new DSATUR-based algorithm for exact vertex coloring," *Comput. Oper. Res.*, vol. 39, no. 7, pp. 1724–1733, Jul. 2012.
[55] M.-E. Vidal, S. Castillo, M. Acosta, G. Montoya, and G. Palma, *On the Selection of SPARQL Endpoints to Efficiently Execute Federated SPARQL Queries*. Berlin, Germany: Springer, 2016, pp. 109–149.

[56] E. P. Klement, R. Mesiar, and E. Pap, "Triangular norms: Basic notions and properties," in *Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*. Amsterdam, The Netherlands: Elsevier, 2005, pp. 17–60.
[57] P. Ferragina and U. Scaiella, "TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities)," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, 2010, pp. 1625–1628.
[58] A. Sakor, I. O. Mulang, K. Singh, S. Shekarpour, M. E. Vidal, J. Lehmann, and S. Auer, "Old is gold: Linguistic driven approach for entity and relation linking of short text," in *Proc. Conf. North*, 2019, pp. 2336–2346.
[59] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: Shedding light on the web of documents," in *Proc. 7th Int. Conf. Semantic Syst.*, Graz, Austria, 2011, pp. 1–8.
[60] A. Sakor, K. Singh, A. Patel, and M.-E. Vidal, "Falcon 2.0: An entity and relation linking tool over Wikidata," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 3141–3148.
[61] A. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. Kohane, "Clinical concept embeddings learned from massive sources of multimodal medical data," in *Proc. Pacific Symp. Biocomput.*, Jan. 2020, pp. 295–306.
[62] *Snli*, 2021, Accessed: Jul. 2021.
[63] *Multisnli*, 2021, accessed: Jul. 2021.
[64] *Twitter Search Api*, TwitterInc, San Francisco, CA, USA, 2022, Accessed: Apr. 2022.
[65] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proc. 22nd Int. Conf. World Wide*. New York, NY, USA: Association for Computing Machinery, 2013, pp. 1089–1098.
[66] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, and S. Dietze, "TweetsCOV19—A knowledge base of semantically annotated tweets about the COVID-19 pandemic," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.* New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 2991–2998.
[67] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman, "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)," *Science*, vol. 368, no. 6490, pp. 489–493, May 2020.
[68] Rituparna. (2018). *Kaggle Fifa 2018 Tweets*. Accessed: Oct. 19, 2021. [Online]. Available: https://www.kaggle.com/rgupta09/world-cup-2018-tweets
[69] H. S. Pattanayak, H. K. Verma, and A. L. Sangal, "Community detection metrics and algorithms in social networks," in *Proc. 1st Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, Dec. 2018, pp. 483–489.
[70] M. Chen, T. Nguyen, and B. K. Szymanski, "On measuring the quality of a network community structure," in *Proc. Int. Conf. Social Comput.*, Sep. 2013, pp. 122–127.
[71] S. Saha and S. Ghrera, "Network community detection on metric space," *Algorithms*, vol. 8, no. 3, pp. 680–696, Aug. 2015.
[72] M. Gaertler, *Clustering*. Berlin, Germany: Springer, 2005, pp. 178–215.
[73] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, Feb. 2006.
[74] A. Buluc, H. Meyerhenke, I. Safro, P. Sanders, and C. Schulz, *Recent Advances in Graph Partitioning*. Cham, Switzerland: Springer, 2016, pp. 117–158.
[75] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing," in *Proc. 18th BioNLP Workshop Shared Task*, 2019, pp. 319–327.
[76] S. De Coronado, M. W. Haber, N. Sioutos, M. S. Tuttle, and L. W. Wright, "NCI thesaurus: Using science-based terminology to integrate cancer research results," in *Medinfo*. 2004, pp. 33–37.

AHMAD SAKOR received the B.Sc. degree in informatics engineering specializing in artificial intelligence from Arab International University, Syria, and the M.Sc. degree in computer science from the University of Bonn, Germany. He is currently pursuing the Ph.D. degree in computer science from Leibniz Universität Hannover. He is a Research Assistant at the Joint Laboratory of L3S Research Center and TIB–Leibniz Information Centre for Science and Technology. His research experiences are in question answering systems, and its related components (entity linking and relation linking). His research interests include natural language processing, and knowledge extraction, combining heuristics and machine learning approaches.

KULDEEP SINGH received double master's degree in computer science from the Technical University of Berlin, Germany, and Aalto University, Finland, and the Ph.D. degree from the University of Bonn, Germany, focusing on answering questions over knowledge graphs. He is currently working as the Principle Product Manager. He received the prestigious Marie Curie Fellowship from the European Union for his Ph.D. studies. He regularly publishes in top conferences, such as The Web Conference, CIKM, ECML, EACL, ESWC, SIGIR, and ISWC.

MARIA-ESTHER VIDAL is a Full Professor with the Leibniz University of Hannover and leads the Scientific Data Management (SDM) Group at the TIB-Leibniz Information Centre for Science and Technology. She is also a member of the L3S Research Centre, and a Full Professor (retired) at Universidad Simón Bolívar (USB), Venezuela. She researches on data management, semantic data integration, and machine learning over knowledge graphs. She is a coauthor of more than 230 peer-reviewed articles in semantic web, databases, and artificial intelligence. She has been awarded the Science Award on Responsible Research by Stifterverband with the recommendation of the Leibniz Association and with the program "Leibniz Best Minds: Programme for Women Professors" supported by the Leibniz Association, Germany. She is also actively shaping her research communities. Under her direction, her team has developed technologies of predominant relevance in the whole process of knowledge graph creation from heterogeneous data and query processing. She serves as an expert in several advisory boards, summer schools, and doctoral consortiums. She has advised more than 25 doctoral students, and more than 120 master's and bachelor's students in computer science. She has been a Doctoral and Habilitation Committee Member in France, Italy, the Netherlands, Germany, Ireland, Argentina, Uruguay, and Venezuela. She is an Editorial Board Member of renowned journals, such as *JWS* and *JDIQ* and the General Chair, the Co-Chair, a Senior Reviewer of major scientific events, such as, ESWC, WWW, ISWC, and AAAI.