Toward a Cognitive-Inspired Hashtag Recommendation for Twitter Data Analysis

Youcef Djenouri, *Member, IEEE*, Asma Belhadi, Gautam Srivastava, *Senior Member, IEEE*, and Jerry Chun-Wei Lin, *Senior Member, IEEE*

***Abstract*— This research investigates hashtag suggestions in a heterogeneous and huge social network, as well as a cognitive-based deep learning solution based on distributed knowledge graphs. Community detection is first performed to find the connected communities in a vast and heterogeneous social network. The knowledge graph is subsequently generated for each discovered community, with an emphasis on expressing the semantic relationships among the Twitter platform's user communities. Each community is trained with the embedded deep learning model. To recommend hashtags for the new user in the social network, the correlation between the tweets of such user and the knowledge graph of each community is explored to set the relevant communities of such user. The models of the relevant communities are used to infer the hashtags of the tweets of such users. We conducted extensive testing to demonstrate the usefulness of our methods on a variety of tweet collections. Experimental results show that the proposed approach is more efficient than the baseline approaches in terms of both runtime and accuracy.**

***Index Terms*— Cognitive computing, deep learning, hashtag recommendation, semantic analysis, social network.**

I. Introduction

COGNITIVE computing development has been largely increased in the last five years. In particular, Siri, Google Assistant, Cortana, and Alexa are a few of the best examples of cognitive computing technologies, which imitate human personality and address the common challenges in natural language processing [1]. Social network analysis is the process of incorporating the fundamental concepts of graph theory in studying the different properties and features of large social networks. It is a combination between nodes and edges, where nodes represent the users and edges represent the connection among the users [2]. Cognitive computing for social network analysis has recently become high research of interest where analyzing and handling the semantic features in the social network is crucial [3], [4].

Manuscript received 2 December 2021; revised 27 February 2022; accepted 20 April 2022. Date of publication 9 May 2022; date of current version 1 December 2022. *(Corresponding author: Jerry Chun-Wei Lin.)*

Youcef Djenouri is with SINTEF Digital, 7465 Oslo, Norway (e-mail: youcef.djenouri@sintef.no).

Asma Belhadi is with the School of Economics, Innovation and Technology, Kristiania University College, 0107 Oslo, Norway (e-mail: asma.belhadi@kristiania.no).

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Centre for Interneural Computing, China Medical University, Taichung 40402, Taiwan (e-mail: srivastavag@brandonu.ca).

Jerry Chun-Wei Lin is with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway (e-mail: jerrylin@ieee.org).

Digital Object Identifier 10.1109/TCSS.2022.3169838

Hashtags are one of the well-known semantic representations for various social networks, such as Twitter or Facebook [5]. The hashtags are metadata added to the main contents to easily identify the given message with a specific theme or content [6]. Hashtag recommendation has the goal to label the nontagged tweets with relevant hashtags from the tweet collection. It is used in various applications, such as business intelligence [7], multimedia analysis [8], and/or smart cities [9].

A. Motivations

Two types of solutions have been developed for hashtags recommendation, approaches which study the correlation among the tweets collection, these solutions use the pattern discovery in recommending the hashtags [10]–[12]. Other solutions attempt the use of the deep learning architectures to learn the hashtag recommendation process [13]–[15]. Solutions to pattern discovery are high time and memory consuming, whereas solutions to deep learning suffer from accuracy. This is explained by the fact that the tweets in the social network are heterogeneous, where different types of hashtags may be published. In addition, the same meaning of different hashtags may be observed, for instance, the hashtags #Play and #Game are different hashtags that represent the same meaning. To improve the accuracy of the deep learning solutions, without loss of the processing runtime, we develop in this research work, an intelligent framework, which combines the community detection to study the different correlations among the tweets, the deep learning to learn the hashtag recommendation process and the knowledge graphs to represent the semantic meaning of the different hashtags.

B. Contributions

This article presents a new intelligent framework for recommending hashtags. Our methodology consists in exploring dependencies among the tweets in the social network and uses deep learning to learn the hashtag recommendation process from highly correlated tweets, by incorporating the knowledge graph to describe the semantic meaning of the different hashtags in the social network. The main contributions of the presented work can be summarized as follows.

1) We propose an intelligent approach to recommending the hashtags of the nontagged tweets. It explores community detection, deep learning, and knowledge graphs. Community detection is performed to find the highly

This is the author accepted version of an article published in
IEEE Transactions on Computational Social Systems (volume 9, issue 6, December 2022)
https://doi.org/10.1109/TCSS.2022.3169838

correlated communities of users in the social network. The embedded deep learning is trained on tweets of each community of highly correlated users.

2) We propose a semantic strategy to represent the different meaning of the hashtags of highly correlated users. We also use the shared knowledge base to represent the same meaning of the hashtags of the different communities. Both knowledge graphs and the shared knowledge base are used to select the relevant tweets and models in the recommendation process.
3) We perform intensive experiments on different tweets collection. The results of experiments reveal that the proposed framework outperforms the baseline algorithms in terms of both runtime and accuracy.

The rest of this article is structured as follows. Section II studies the main existing hashtag recommendation algorithms, followed by a detailed explanation of the developed framework in Section III. Section IV evaluates the results of the designed model compared with the existing approaches. Section V presents the main directions in exploring the community detection, the knowledge graphs, and the deep learning for solving the hashtag recommendation problem. Finally, Section VI shows the final conclusion of this article and provides possible extensions as the future works.

II. Related Work

Solutions to the hashtag recommendation problem [16]–[20] may be divided into two categories. Traditional-based solutions use classical data mining, machine learning approaches in recommending the relevant hashtags, and advanced-based hashtag recommendation solutions, which use the recent deep learning architectures in learning and predicting the relevant hashtags. Next, we highlight works that fall into both categories.

Zhao *et al.* [21] developed a personalized hashtag recommendation strategy based on user profiling and latent Dirichlet allocation (LDA) to first determine the frequencies of all hashtags of the top-k similar users and then recommend the most relevant hashtags to the given user. Li *et al.* [22] suggested the use of the probabilistic latent factor model and the content information to analyze the user profiles and detect the set of personalized hashtags. First, the set of user and microtopic latent factors are calculated. Then, the best microtopics are retrieved for each user by fitting the distribution of the derived models from the previous step. Gong *et al.* [14] disseminated a model that is generative. The model can integrate textual and visual information to be used for hashtag recommendation. The authors use a Gibbs model for sampling to decipher hidden topics. Xie *et al.* [23] developed a clustering-based algorithm to analyze microblogs in the context of social network analysis. To increase the accuracy of event clustering, the provenance-based community partition is used along with local modularity. Incremental clustering is also explored to filter noisy data, enable event data, and enable dynamic migration. Liu *et al*. [24] introduced a model that exploits hierarchical relations like: hashtag/hashtag, hashtag/tweet, tweet/word, and word/word. This helps for the semantic understanding of the tweets that are tagged. Next, the authors use an embedding system that is content-based to assist in the derivation of network embedding. Liu *et al.* [25] developed a deep learning-based strategy for detecting stigmatized content on online social network platforms. This method used semantic-based quantitative analysis to uncover important spatiotemporal aspects of COVID-19 stigma to provide timely warnings and risk mitigation. Almuqren and Cristea [26] took aim at the Saudi Telecommunication Company, Riyadh, Saudi Arabia, and monitored their customers' satisfaction via Twitter and questionnaire analysis in real time. They used social media mining and a quantitative method to assess customer satisfaction with a telecommunications provider. Gao *et al*. [27] developed a hybrid deep neural network system for microblog recommendation. The user interest tags with interest topics are used for deriving the candidate's recommended microblogs. In addition, a collection of heterogeneous features is extracted to show microblogs. Kumar *et al.* [12] developed a new approach that deal with the data spareness in hashtag recommendation systems. The idea is to explore the useful features of the external resources. It also used both the lexical and topical features in enriching the semantic context of the recommended hashtags. Chen *et al.* [28] developed an intelligent method based on graph convolution neural network to predict the relevant hashtags. The process is composed of three main steps. Each data representation is first handled separately, the features of images are extracted using the convolution neural network, and the features of sequences of text are extracted using the long short term memory [29]. These features are used for edge generation. The graph of features is trained to predict the hashtags of nontagged tweets. Jelodar *et al.* [30] used the LDA to study the correlation among the scholar-context documents. The probabilistic modeling is explored with the Gibbs sampling strategy to analyze one of the top eight conference publications in information retrieval and software engineering. Sert *et al.* [31] introduced the topic dynamics of Twitter content and analyze more than one million tweets with the hashtag "COVID." The study showed that the hygiene lifestyle is decreased when the official announcement of the first COVID is released. The results also revealed that a number of friends and followers with COVID hashtags are highly connected compared with others. Saini *et al.* [32] is the first work, which explores both the images and the tweet text in microblog summarizing. The new evolutionary computation algorithm is proposed with a novel multi-objective function, several dimensions for microblog summarizing are considered such tweet similarity, the maximum number of hashtags per tweet, and the redundancy factor. Naseem *et al.* [33] analyzed the COVID views by pointing out followers who share the social media on Twitter. The COVIDSENTI data are created, which is a large dataset for sentiment and Twitter analysis. The tweets are annotated within three classes, positive, negative, or neutral.

When looking at the algorithms proposed to date, they tend to ignore correlations and dependencies that may exist between tweets. Overall, these actions would reduce the quality of the recommendation process for hashtags. In this work, we explore and study the major correlations among the

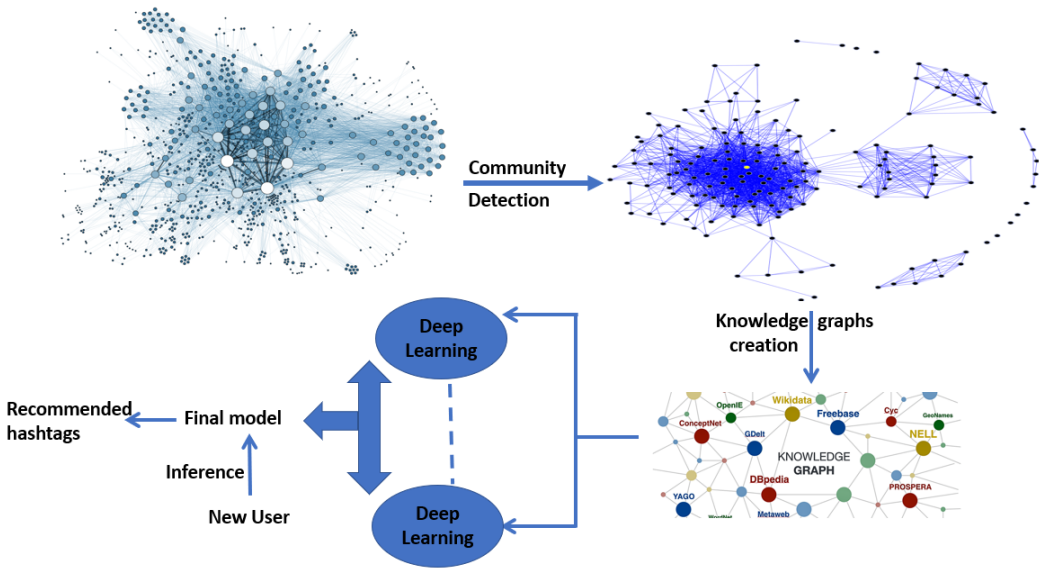

Fig. 1. CKD-HR framework.

tagged tweets. We also present a novel learning model that makes use of knowledge graphs and community detection to find the relevant tweets that should be used in the hashtag recommendation.

III. CKD-HR Framework

A. Principle

This section presents our framework for community detection with a knowledge graph and deep learning for hashtag recommendation (CKD-HR). As illustrated in Fig. 1, the purpose is to accurately select the relevant tweets for recommending the relevant hashtags of the nontagged tweets. It explores community detection, knowledge graph, and deep learning to determine the set of relevant hashtags to the nontagged tweets. The main idea is to apply the divide-and-conquer model to get the subsolutions and then combine all the subsolutions into a global solution. The process starts by applying the community detection to the social network and creating communities inside the network. Each community contains highly correlated users, each user provides tagged tweets. The knowledge graph is then created for each community in the network, which results in a graph inside each community of users describing its hashtags in the tweets. For instance, the hashtag #Play has high similarity with #Game, less similarity with #Enjoy, and nonsimilarity with #Food. The knowledge graph of each community is then matched to each other in order to create a set of shared knowledge bases containing similar concepts and properties of each pair of communities in the network. In addition, the embedded deep learning model is trained on each tweet of communities. To tag the tweets of the new user, the most relevant communities to such tweets are selected, and the models of such communities are used in the inference step for recommending the hashtags of such tweets. The selection of the relevant communities is based on both the knowledge graphs and the shared knowledge base among communities. Note that the results of the CKD-HR solution are the best model trained with the set of recommended hashtags for the new tweets in live. The best model can also be used to recommend hashtags for future tweets. In the remainder of this section, we show how to use all these concepts in the CKD-HR framework.

B. Community Detection

We investigate community detection to identify the correlation between the set of tweets in the large social network. In this stage, we consider the social network, in which each user is a node, and an edge is formed between two users if they are connected. After constructing the user graph, we utilize the Louvain algorithm [34] to create highly connected user communities. It optimizes the correlation among communities, which implies comparing the densities of users within communities to users from other communities. It combines several communities as one single node and creates condensed subgraphs by incrementally using this hierarchical clustering model. The procedure begins by putting each user in a separate community, and then progressively merges communities until the gain in correlation is stable. All existing communities are scanned for each iteration, and for each community, it examines neighbors' communities and merges the most comparable community to the processed one. This work will result in a collection of communities. Each community, as defined by the proposed model, is a strongly correlated subgraph. In this situation, a community is a group of users who are closely linked and correlated.

C. Deep Learning

In this article, we present a deep learning architecture based on entity embedding to deal with textual data. In the first step, the structure is represented as a vector of features by generating embeddings. Before being passed to the output

layer, the created feature vectors are coupled to two fully connected layers. To compress the structured data into feature vectors, the bag of words approach was used. First, a series of visual words are formed from the data. Data and words from each row are combined to create a matrix called DW with d rows and w columns (the number of samples and words, respectively). Each element $DW(i, j)$ indicates whether or not the ith data are present in the jth word. The input of the network is the matrix of visual words of highly correlated tweets, and the output will be a set of hashtags of such tweets. The aim is to learn the hashtag recommendation process for each community detected above. As a result of this step, k different models are created, each of which is trained from the tweets of the users of the given community. Note that k is the number of communities detected in the previous step.

D. Knowledge Graphs

In order to represent the semantic meaning of the hashtags, the knowledge graphs are incorporated into the hashtag recommendation process. The knowledge graph is created for describing the hashtags of each community of users. The hashtags of each community are parsed, where a link between two hashtags is built if these two hashtags are similar to a certain degree. In order to find similar hashtags among communities, knowledge graph matching is used. For each two pair of communities, the knowledge graph matching process [11] is used in order to find the similar hashtags among the knowledge graphs of these communities. As a result of this step, two main components are created.

1) A set of the knowledge graph, each knowledge graph represents the relations of the hashtags in the given community of users.
2) Shared knowledge base which represents the similar hashtags among each two pair of communities if it exists.

E. Selection of Relevant Models/Tweets

To recommend hashtags for the tweets of the new user, we need to efficiently explore the knowledge graph, and the knowledge base builds in the previous step. The words are generated from each tweet, and the similarity between each word and the hashtag of each knowledge graph is determined. The knowledge graph with high similarity with the words of each tweet is returned and added to the set of relevant knowledge graphs (RKG). The shared knowledge base is then explored where the knowledge graphs that have shared hashtags with the elements of RKG are recursively added to RKG. The models of the communities where their knowledge graphs are in RKG are finally explored to infer/recommend the hashtags of the tweets of the new user by concatenating the outputs of all selected models.

F. Improvement With Hyperparameter Optimization

In this research work, different deep learning models have been deployed each trained a given community of tweets. In order to improve the learning process of the hashtag recommendation, we propose to optimize the hyperparameters of the models used.

Definition 1: We called HP to be the set of hyperparameters of the deep learning models used in the training.

For instance, the number of epochs, the number of batches, the number of layers in the network, and the activation functions are the hyperparameters of the models.

Definition 2: We define the domain value of each hyperparameter hp, noted $V(hp)$, which contains all possible values of hp.

For instance, the domain value of the activation function is the set {Binary step, linear, sigmoid, tanh, leaky, relu, parameterized relu, exponential linear unit, swish, softmax}.

Definition 3: The configuration space $\mathcal{CS}$ is defined by the set of all possible configurations. Each configuration in $\mathcal{CS}$ is a vector of possible values in $V(hp)$.

For instance, the configuration ("Softmax", 16, 100, 0.05) represents the following settings.

1) Activation function is set to "Softmax."
2) The number of batches is set to 16.
3) The number of epochs is set to 100.
4) The loss rate is set to 0.05.

Algorithm 1 CKD-HR Algorithm

Input:
$T = \{T_1, T_2, \ldots, T_n\}$: the set of n tweets for training;
$T' = \{T'_1, T'_2, \ldots, T'_m\}$: the set of m tweets used in the inference phase;
Output:
$best\_model$: the best model generated in the training phase;
H: the set of recommended hashtags for each tweet in T';
**********Before Training **********
$C \leftarrow Louvain(T)$;
$KG \leftarrow \emptyset$;
for c ∈ C **do**
  $KG \leftarrow KG \cup KnowledgeGraphConstruction(T, C)$;
end for
**********Training**********
$error\_loss \leftarrow \infty$;
$best\_model \leftarrow \emptyset$;
for generation in maximum_generations **do**
  $S \leftarrow generate\_solutions(generation)$;
  $current\_model \leftarrow DeepLearning(T, KG, S)$;
  if error(current_model) ≤ error_loss **then**
    $best\_model \leftarrow current\_model$;
    $error\_loss \leftarrow error(current\_model)$;
  end if
end for
**********Inference**********
$H_c \leftarrow \emptyset$;
for $T'_i \in T'$ **do**
  $H\_i \leftarrow best\_model(T'_i)$;
  $H \leftarrow H \cup H_i$;
end for
return $(best\_model, H)$.

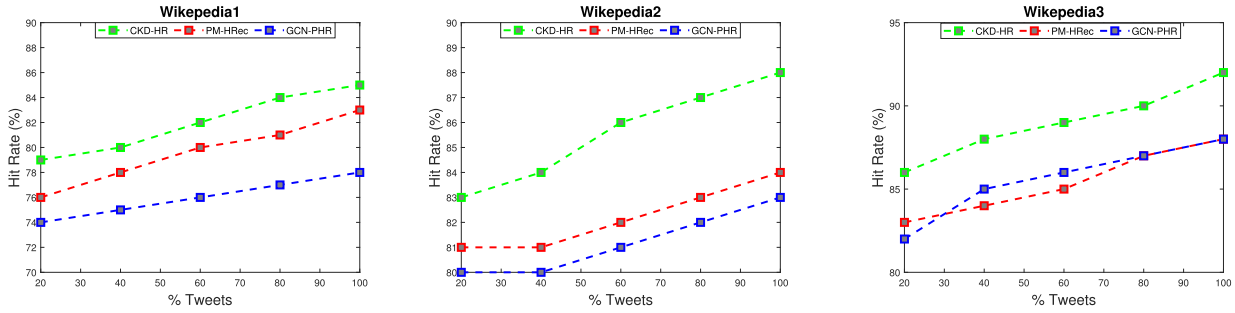

Fig. 2. CKD-HR versus state-of-the-art solutions: accuracy on small data.

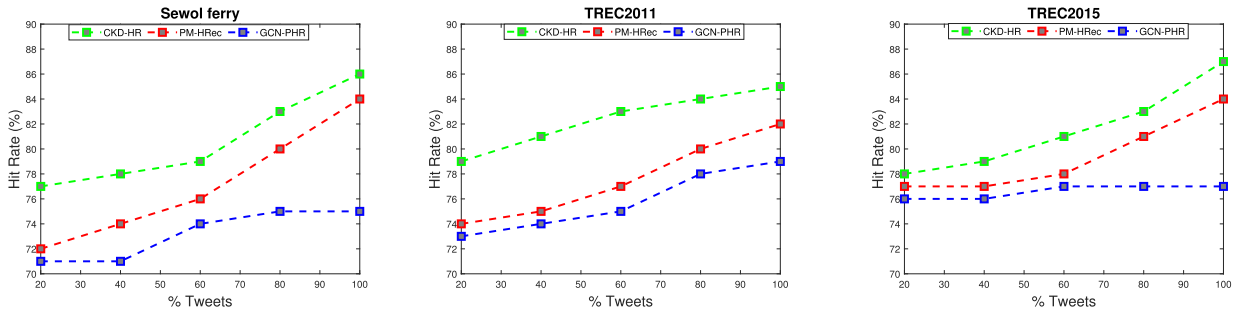

Fig. 3. CKD-HR versus state-of-the-art solutions: accuracy on medium data.

Finding the best values for all hyperparameters of the models of CKD-HR needs to explore all the configuration space in $\mathcal{CS}$. This is high time consuming where the number of possible configuration is the combination of all possible values of the hyperparameters in hp times the number of communities, which is determined by

$$|\mathcal{CS}| = \prod_{i=1}^{|HP|} |V(hp_i)| \times |C|. \quad (1)$$

The configuration space is huge. For instance, if we only consider 500 possible values for epoch parameter, 200 possible value for error rate, and 100 communities, the size of the configuration space is 10 million configurations. Therefore, traditional search-based strategies, such as branch and bound [35], and A* [36] became inefficient for such a high number of configurations. Therefore, the evolutionary computation approaches are applied to accurately find a configuration that is as close to the optimal one as possible. We used the genetic algorithm, the well-known algorithm for evolutionary optimization. First, the initial population is randomly generated, where each individual is created from the configuration space. The crossover and mutation operators are then applied to explore the configuration space. To maintain consistent population size, every individual is evaluated by making use of the hashtag recommendation accuracy, while keeping the best individuals and removing the rest. This process is repeated in multiple iterations until the max number of iterations is reached. Although in this study, we consider the ratio of correct recommended hashtags as a fitness function for evaluating the generated individuals, hyperparameter optimization can be considered as a multi-objective problem, taking into account the cost of training the model, memory usage, and other criteria, such as precision, recall, and F1_score, in determining the best individuals of the evolutionary algorithm.

G. Pseudocode

Algorithm 1 shows the pseudocode of the CKD-HR algorithm. The input data is the set of n tweets used for training. These data are accompanied by the ground truth represented by the set of hashtags. We also provide a set of m tweets to test the trained model (line 1). The output is the best-trained model and the set of recommended hashtags for each tweet in the test data (line 2). The process begins by detecting communities in the training data using the Louvain algorithm (line 4). The knowledge graph is then created for each detected community (lines 5–8). The tweets with their associated knowledge graphs are trained using deep learning by optimizing the hyperparameters (lines 10–19). As a result of the training phase, the weights of the best-trained model (best_model) are returned. In the inference phase, the weights of the best_model for each test tweet are passed to recommend the hashtags of that tweet (lines 21–25). The algorithm returns both the best model and the recommended hashtags of the test tweets (line 26). We note that the training step, which is performed only once regardless of the number of tweets used for inference, is a time-consuming process that requires multiple generations of the evolutionary process to find the optimal model. The inference step, on the other hand, involves only one loop and requires only a simple propagation of the learned model during the training phase.

IV. Performance Evaluation

A. Experimental Environment

Extensive tests were conducted to assess the performance of the developed method using benchmark hashtag

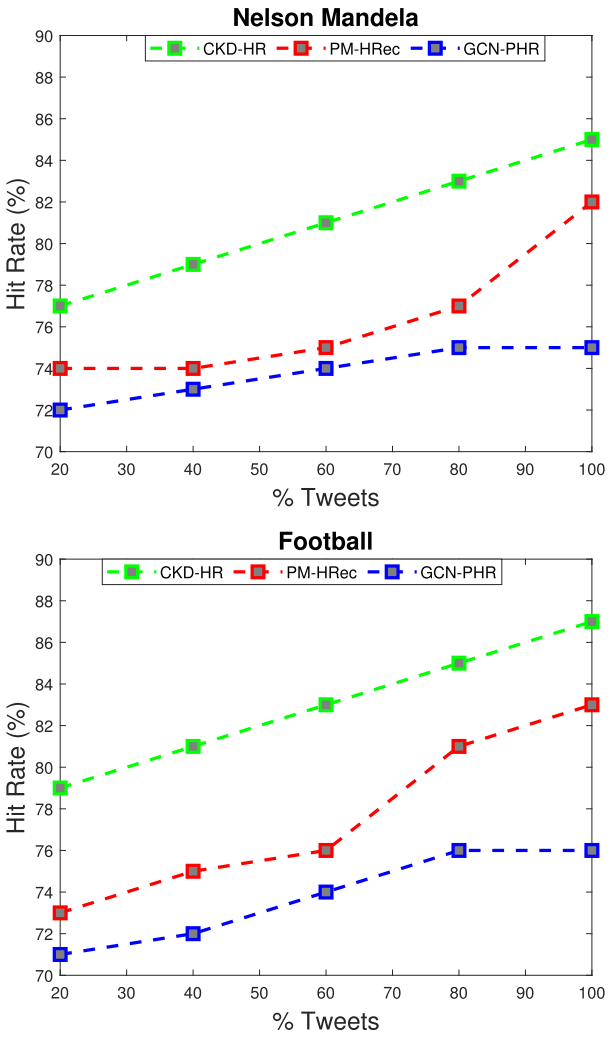

Fig. 4. CKD-HR versus state-of-the-art solutions: accuracy on large data.

recommendation collections.[1] The training, validation, and testing sets are selected sequentially (70% for training, 10% for validation, and 20% for testing). All algorithms have been implemented in Python 3.7 using the Keras library for deep learning. We also implement the community API for community detection and Pykg2vec for knowledge graph.

For proper evaluation of recommended hashtags, both computational time and accuracy are calculated. The runtime is measured in seconds, and the accuracy is determined by *hit_rate* measure, which is defined as

$$\text{hit\_rate} = \frac{\sum_{\mathcal{T}_i \in \mathcal{T}_{\text{test}}} \text{Correct}(\mathcal{T}_i)}{|\mathcal{T}_{\text{test}}|} \tag{2}$$

where $\text{Correct}(\mathcal{T}_i)$ is set to 1 if the set of the recommended hashtag of $\mathcal{T}_i$ contains the standard hashtags of $\mathcal{T}_i$. Otherwise, its value is 0. $\mathcal{T}_{\text{test}}$ is the set of the testing tweets.

Different tweets collections are used in the experiments (See Table I for more details). These databases are varied from small, large, sparse, and dense collections. Thus, some tweet collections contain a high number of tweets, some collections

[1]https://www.aminer.org/data-sna

TABLE I
TWEETS COLLECTION DESCRIPTION

Category	Corpus	# Tweets	# Hashtags
Small	Wikepedia1	81,270	13,156
	Wikepedia2	86,929	19,124
	Wikepedia3	168,199	32,280
Medium	Sewol ferry	239,117	723
	TREC2015	250,306	66,384
	TREC2011	333,491	106,682
Large	Nelson Mandela	2,813,461	50,425
	Football	3,000,000	90,660

contain a high number of hashtags, and some others contain both a high number of tweets and hashtags. The state-of-the-art hashtag recommendation approaches implemented in this study are pattern mining for hashtag recommendation (PM-HRec) [11], and graph convolution network-based personalized hashtag recommendation (GCN-PHR) [37]. The first approach studied the correlation by using pattern discovery in the hashtag recommendation process. The second approach is a graph convolution neural network for learning the hashtag recommendation from the tweets collection.

B. CKD-HR Versus State-of-the-Art Solutions

This part of the analysis aims to validate the usage of CKD-HR against the state-of-the-art hashtag recommendation approaches. Figs. 2–4 evaluate the accuracy of CKD-HR determined by the rate value with the PM-HRec, and GCN-PHR using different tweet collections. The results show the large superiority of CKD-HR against the two baseline algorithms. By varying the percentage of tweets from 20% to 100%, the hit rate of the CKD-HR is higher than the two other algorithms, whatever the scenario used as input. For instance, with 100% of tweets on Wikepedia3 collection, the hit rate of the baseline algorithms does not exceed 88%, whereas the hit rate of CKD-HR is up to 92%. Figs. 3, 5, 6, and 7 demonstrate the computational processing time of CKD-HR utilizing baseline hashtag recommendation methods and various twitter collections. The findings show that CKD-HR outperforms the baseline algorithms. By varying the percentage of tweets from 20% to 100%, the processing runtime of the CKD-HR is less than the two baseline approaches, where a clear superiority against PM-HRec is observed. For instance, with 100% of tweets on Wikepedia3 collection, the runtime of PM-HRec exceeds 32 s, whereas the runtime of CKD-HR does not exceed 10 s. The following reasons contributed to the achievement of these results.

1) Using community discovery to learn from tweets with similar hashtags that are strongly connected.
2) Using a knowledge graph makes it possible to deal with multiple concepts that have the same meaning.
3) Implementing embedded deep learning on each community of tweets allows for rapid learning of the hashtag recommendation method.

C. Performance on Large and Big Scale Tweets Collection

Using a huge corpus of 40 000 000 tweets, Fig. 8 shows both the accuracy and the runtime. We can conclude that the

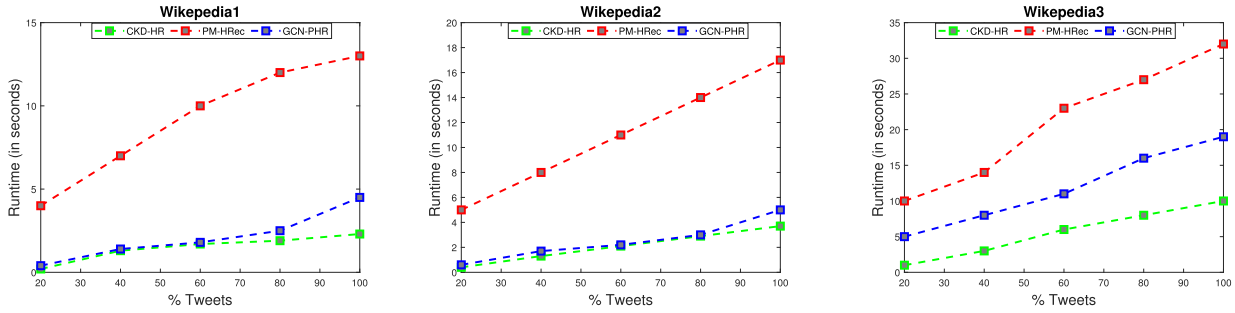

Fig. 5. CKD-HR versus state-of-the-art solutions: runtime on small data.

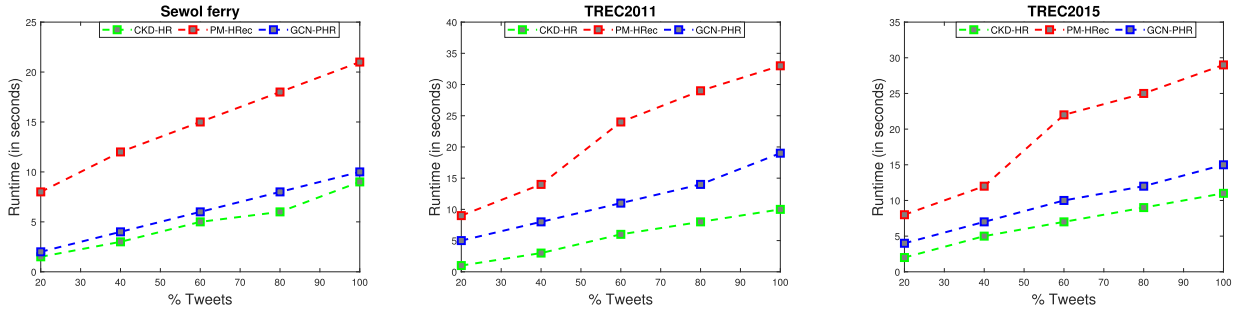

Fig. 6. CKD-HR versus state-of-the-art solutions: runtime on medium data.

baseline solutions take more time to process 40 000 000 tweets than our approach. This is especially true for the pattern discovery-based approach, which needs to extract all relevant patterns to explore the correlation between tweets, while our approach only needs to apply community detection to find the correlated tweets. Moreover, regardless of the circumstances used in the experiment, our solution outperforms both methods in terms of accuracy. These results support the application of the proposed method to the current hashtag suggestion problem.

D. Case Study

This study focuses on the output results showing the hashtags recommended by CKD-HR. This case study focuses on three topics: health, movies, and sports-related tweets. Table II shows that the proposed approach can recommend interesting hashtags, such as #afl15 for sports, which interprets the performance of the Arizona Fall League (AFL) baseball team in 2015. Our approach derives important knowledge from tagged tweets and computes semantic similarity using the knowledge graphs of the created communities, which explains these results.

TABLE II
CASE STUDY OF CKD-HR APPROACH

Topic	Keyword	Top two relevant recommended hashtags
Health	Pharmacy	#Nursing, #Food
Cinema	Movies	#Geek, #Western
Sport	Basketball	#NBA, #alf15

V. FUNDING, LIMITATION, AND FUTURE PERSPECTIVES

This section presents the primary funding of this article and its shortcomings and future prospects. The following is a summary of the primary funding for the combination of community detection, knowledge graphs, and deep learning:

1) Studying correlation analysis between the collection of tweets is beneficial for the entire hashtag recommendation pipeline. This allows us to form communities that are homogeneous and strongly connected, which is beneficial for the learning process. By simply examining comparable concepts, the deep learning model can be easily trained. For example, if the entire network contains tweets about sports, politics, and science, it is good to identify three communities, each describing a particular concept (sports, politics, and science).
2) The use of knowledge graphs and a common knowledge base between communities enables the selection of the most relevant tweets and the use of deep models to recommend appropriate hashtags for tweets that are not labeled. In a heterogeneous environment, this gives the most relevant tweets and models.
3) This research improves both the semantic and deep learning sides. We can accurately identify the tweets and modes needed for the recommendation process using the knowledge graph created by removing irrelevant tweets and models. In addition, community building helps to improve the performance of the training processes of the different deep learning models.

In the future, several directions could be explored to address the CKD-HR solution and further improve the hashtag recommendation process.

1) **Improving the Community Detection Step:** Community detection aims to classify highly correlated tweets

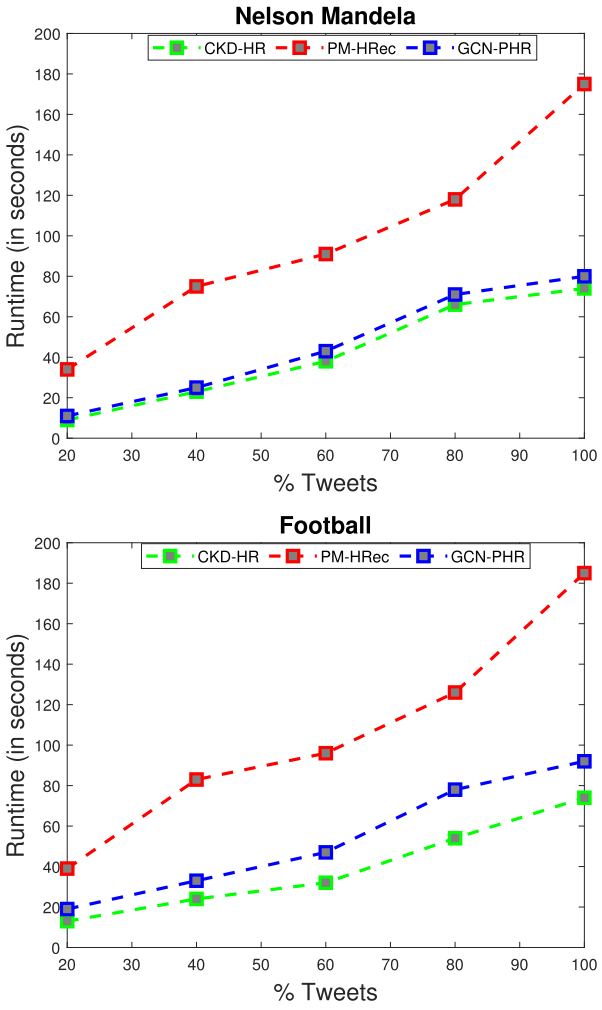

Fig. 7. CKD-HR versus state-of-the-art solutions: runtime on large data.

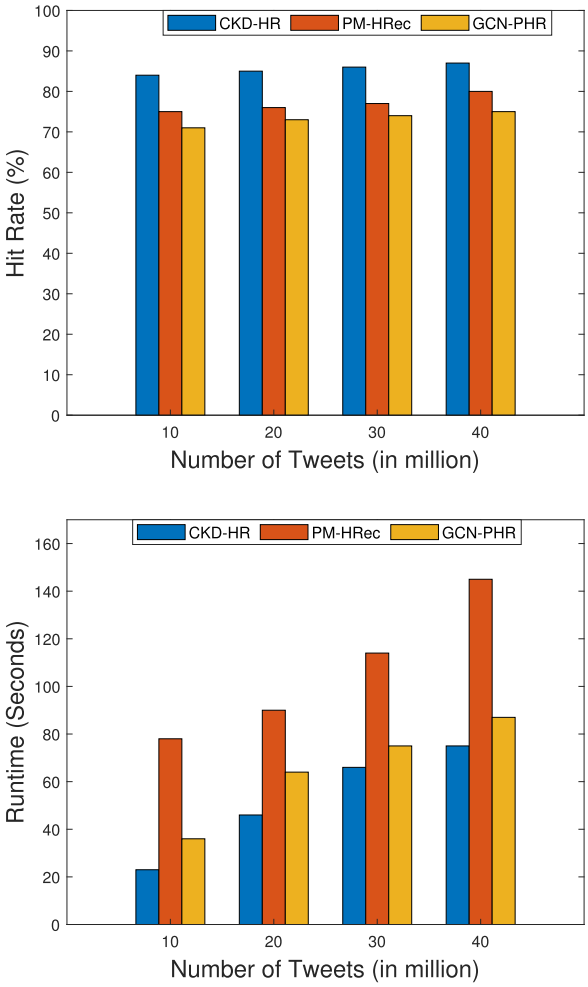

Fig. 8. Performance analysis on large-scale tweets collection.

into the same community. Even though Louvain's algorithm is considered one of the state-of-the-art community detection algorithms, this algorithm is not suitable for processing large-scale graph data and is very sensitive to the number of communities. Moreover, it cannot handle overlapping communities where each node/user may be in more than one community. To overcome these limitations, additional techniques can be used to identify homogeneous communities. In the future, it may be interesting to merge different decomposition techniques with the one presented here. Some examples are pure partitioning [38], fuzzy partitioning [39], or pattern mining [11]. Another option is to develop a method to automatically regulate the number of communities. Using multiple runs to find the best value for the number of communities is not very efficient in practice. One method to solve this problem is to create a knowledge base that contains each social training network configuration with the best value for the number of communities and then analyze the meta-features of the social network configuration (number of users, number of hashtags, number of tweets, times of published tweets, and so on) and the best values for the number of communities. In this way, the optimal number of communities for a new social network can be automatically predicted.

2) **Improving the Recommendation Step:** Even though CKD-HR has improved runtime performance, processing large collections of tweets still takes a lot of time. To address this limitation, we plan to integrate high-performance computing, such as graphical processor units [40] and supercomputing [41], as future work. The community discovery step creates communities that can be handled by separate jobs while meeting high-performance computing requirements, such as synchronization, communication, memory management, and load balancing. Therefore, developing effective load balancing solutions is an important issue. One way to address this problem is to develop tools that allow the identification of homogeneous groups based on the number of users and tweets. Another way is to develop new community matching tools to identify groupings with a similar number of users and tweets. Another way to optimize the whole recommendation process is to use the proposed framework with MapReduce.

VI. Conclusion

This research presents a cognitive-based deep learning approach for hashtag recommendations based on distributed knowledge graphs. Community detection, knowledge graphs, and deep learning are used in the developed methodology. It starts with the identification of homogeneous communities in the social network. Each community contains users who share similar tweets and hashtags. The hashtag recommendation process is then learned for each community using deep learning. The knowledge graphs with shared knowledge bases between communities are constructed with the goal of highlighting the semantic relationships between user communities on the Twitter platform while accurately selecting relevant groups for recommending hashtags for the new user's nontagged tweets. Extensive experiments were conducted to thoroughly demonstrate the usefulness of our method using different collections of tweets. The experimental results show the efficiency of the proposed approach compared with the baseline approaches in terms of both runtime and accuracy.

References

[1] Q. Zheng *et al.*, "Exploring Weibo users' attitudes toward lesbians and gays in mainland China: A natural language processing and machine learning approach," *Comput. Hum. Behav.*, vol. 127, Feb. 2022, Art. no. 107021.

[2] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3417–3455, Sep. 2020.

[3] G. Wu and H. Ji, "Short-term memory neural network-based cognitive computing in sports training complexity pattern recognition," *Soft Comput.*, pp. 1–16, Jan. 2022.

[4] K. Z. Khanam, G. Srivastava, and V. Mago, "The homophily principle in social network analysis: A survey," *Multimedia Tools Appl.*, pp. 1–44, Jan. 2022.

[5] K. Petersen and J. M. Gerken, "#COVID-19: An exploratory investigation of hashtag usage on Twitter," *Health Policy*, vol. 125, no. 4, pp. 541–547, Apr. 2021.

[6] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, and A. Cano, "Exploring pattern mining algorithms for hashtag retrieval problem," *IEEE Access*, vol. 8, pp. 10569–10583, 2020.

[7] M. D. Garvey, J. Samuel, and A. Pelaez, "Would you please like my tweet?! An artificially intelligent, generative probabilistic, and econometric based system design for popularity-driven tweet content generation," *Decis. Support Syst.*, vol. 144, May 2021, Art. no. 113497.

[8] S. Zhang, H. Liu, J. He, S. Han, and X. Du, "A deep bi-directional prediction model for live streaming recommendation," *Inf. Process. Manage.*, vol. 58, no. 2, Mar. 2021, Art. no. 102453.

[9] P. Dey and S. Roy, "Governance in smart city: An approach based on social network," in *Smart Cities: A Data Analytics Perspective*. Springer, 2021, pp. 63–87.

[10] A. Javari, Z. He, Z. Huang, R. Jeetu, and K. C.-C. Chang, "Weakly supervised attention for hashtag recommendation using graph data," in *Proc. Web Conf.*, Apr. 2020, pp. 1038–1048.

[11] 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.

[12] N. Kumar, E. Baskaran, A. Konjengbam, and M. Singh, "Hashtag recommendation for short social media texts using word-embeddings and external knowledge," *Knowl. Inf. Syst.*, vol. 63, no. 1, pp. 175–198, Oct. 2020.

[13] K. Lei, Q. Fu, M. Yang, and Y. Liang, "Tag recommendation by text classification with attention-based capsule network," *Neurocomputing*, vol. 391, pp. 65–73, May 2020.

[14] Y. Gong, Q. Zhang, and X. Huang, "Hashtag recommendation for multimodal microblog posts," *Neurocomputing*, vol. 272, pp. 170–177, Jan. 2018.

[15] Z. Wang, H. Xia, S. Chen, and G. Chun, "Joint representation learning with ratings and reviews for recommendation," *Neurocomputing*, vol. 425, pp. 181–190, Feb. 2021.

[16] B. Shi, G. Poghosyan, G. Ifrim, and N. Hurley, "Hashtagger+: Efficient high-coverage social tagging of streaming news," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 43–58, Jan. 2018.

[17] R. Makki *et al.*, "ATR-Vis: Visual and interactive information retrieval for parliamentary discussions in Twitter," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 1, pp. 1–33, Feb. 2018.

[18] S. Sedhai and A. Sun, "Hashtag recommendation for hyperlinked tweets," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 831–834.

[19] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, "Using topic models for Twitter hashtag recommendation," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 593–596.

[20] S. Zhang and H. Cheng, "Exploiting context graph attention for poi recommendation in location-based social networks," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2018, pp. 83–99.

[21] F. Zhao, Y. Zhu, H. Jin, and L. T. Yang, "A personalized hashtag recommendation approach using LDA-based topic model in microblog environment," *Future Gener. Comput. Syst.*, vol. 65, pp. 196–206, Dec. 2016.

[22] Y. Li, J. Jiang, T. Liu, M. Qiu, and X. Sun, "Personalized microtopic recommendation on microblogs," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 6, p. 77, 2017.

[23] Y. Xie, S. Tong, P. Zhou, Y. Li, and D. Feng, "Efficient storage management for social network events based on clustering and hot/cold data classification," *IEEE Trans. Computat. Social Syst.*, early access, Feb. 14, 2022, doi: 10.1109/TCSS.2022.3146310.

[24] J. Liu, Z. He, and Y. Huang, "Hashtag2Vec: Learning hashtag representation with relational hierarchical embedding model," in *Proc. IJCAI*, Jul. 2018, pp. 3456–3462.

[25] L. Liu, Z. Cao, P. Zhao, P. J.-H. Hu, D. D. Zeng, and Y. Luo, "A deep learning approach for semantic analysis of COVID-19-related stigma on social media," *IEEE Trans. Computat. Social Syst.*, early access, Feb. 17, 2022, doi: 10.1109/TCSS.2022.3145404.

[26] L. Almuqren and A. I. Cristea, "Predicting STC customers' satisfaction using Twitter," *IEEE Trans. Computat. Social Syst.*, early access, Jan. 10, 2022, doi: 10.1109/TCSS.2021.3135719.

[27] J. Gao, C. Zhang, Y. Xu, M. Luo, and Z. Niu, "Hybrid microblog recommendation with heterogeneous features using deep neural network," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114191.

[28] Y.-C. Chen, K.-T. Lai, D. Liu, and M.-S. Chen, "TAGNet: Triplet-attention graph networks for hashtag recommendation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1148–1159, Mar. 2022.

[29] J. C.-W. Lin, Y. Shao, Y. Djenouri, and U. Yun, "ASRNN: A recurrent neural network with an attention model for sequence labeling," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106548.

[30] H. Jelodar *et al.*, "Recommendation system based on semantic scholar mining and topic modeling on conference publications," *Soft Comput.*, vol. 25, no. 5, pp. 3675–3696, Mar. 2021.

[31] E. Sert, O. Okan, A. Özbilen, Ş. Ertekin, and S. Özdemir, "Linking COVID-19 perception with socioeconomic conditions using Twitter data," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 394–405, Apr. 2022.

[32] N. Saini, S. Saha, P. Bhattacharyya, S. Mrinal, and S. K. Mishra, "On multimodal microblog summarization," *IEEE Trans. Computat. Social Syst.*, early access, Oct. 22, 2021, doi: 10.1109/TCSS.2021.3110819.

[33] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 4, pp. 1003–1015, Aug. 2021.

[34] T. Shi, S. Ding, X. Xu, and L. Ding, "A community detection algorithm based on quasi-Laplacian centrality peaks clustering," *Appl. Intell.*, vol. 51, pp. 7917–7932, Mar. 2021.

[35] S. Coniglio, F. Furini, and P. S. Segundo, "A new combinatorial branch-and-bound algorithm for the knapsack problem with conflicts," *Eur. J. Oper. Res.*, vol. 289, no. 2, pp. 435–455, 2021.

[36] Z. Bu and R. E. Korf, "A*+BFHS: A hybrid heuristic search algorithm," 2021, *arXiv:2103.12701*.

[37] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, "Personalized hashtag recommendation for micro-videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1446–1454.

[38] Y. Djenouri, A. Belhadi, D. Djenouri, and J. C.-W. Lin, "Cluster-based information retrieval using pattern mining," *Appl. Intell.*, vol. 51, pp. 1888–1903, Oct. 2020.

[39] J. C.-W. Lin, J. M.-T. Wu, Y. Djenouri, G. Srivastava, and T.-P. Hong, "Mining multiple fuzzy frequent patterns with compressed list structures," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–8.
[40] B. Schifferer *et al.*, "GPU accelerated feature engineering and training for recommender systems," in *Proc. Recommender Syst. Challenge*, Sep. 2020, pp. 16–23.
[41] V. Gupta and R. Hewett, "Real-time tweet analytics using hybrid hashtags on Twitter big data streams," *Information*, vol. 11, no. 7, p. 341, Jun. 2020.

Youcef Djenouri (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Science and Technology Houari Boumediene, Algiers, Algeria, in 2014.

He is currently a Research Scientist with SINTEF Digital, Oslo, Norway. He is working on topics related to artificial intelligence and data mining, with a focus on association rules mining, frequent itemsets mining, parallel computing, swarm and evolutionary algorithms, and pruning association rules. He has authored or coauthored more than 100 refereed research papers in the areas of data mining, parallel computing, and artificial intelligence.

Asma Belhadi received the Ph.D. degree in computer engineering from the University of Science and Technology Houari Boumediene, Algiers, Algeria, in 2016.

She is currently a Post-Doctoral Researcher with the Kristiania University College, Oslo, Norway. She is working on topics related to artificial intelligence and data mining, with a focus on logic programming. She has authored or coauthored over 50 refereed research articles in the areas of artificial intelligence and smart city applications.

Gautam Srivastava (Senior Member, IEEE) received the B.Sc. degree from Briar Cliff University, Sioux City, IA, USA, in 2004, and the M.Sc. and Ph.D. degrees from the University of Victoria, Victoria, BC, Canada, in 2006 and 2012, respectively.

He is active in research in the field of cryptography, data mining, security and privacy, and blockchain technology. In his five years as a research academic, he has authored or coauthored a total of 200 papers in high-impact conferences in many countries and in high-status journals, including Science Citation Index (SCI) and SCI Expanded (SCIE).

Jerry Chun-Wei Lin (Senior Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, in 2010.

He is currently a Full Professor with the Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway. He has authored or coauthored more than 500 research articles in refereed journals (with more than 60 ACM/IEEE journals) and international conferences, including the IEEE International Conference on Data Engineering (ICDE), International Conference on Data Mining (ICDM), Principle Knowledge Data Discovery (PKDD), and Pacific–Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 16 edited books, and 33 patents (held and filed, three U.S. patents). His research interests include data mining and analytics, natural language processing, soft computing, the IoTs, bioinformatics, artificial intelligence/machine learning, and privacy-preserving and security technologies.

Dr. Lin is a fellow of IET and an ACM Distinguished Member (Scientist). He is the Editor-in-Chief of the *International Journal of Data Science and Pattern Recognition*, the Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), *Information Science* (INS), *Journal of Information Technology* (JIT), *Journal of Ambient Intelligence and Humanized Computing* (AIHC), *The International Journal of Interactive Multimedia and Artificial Intelligence* (IJIMAI), *Human-centric Computing and Information Sciences* (HCIS), *Intelligent Data Analysis* (IDA), PlosOne, and IEEE ACCESS, and the Guest Editor of several IEEE/ACM journals, such as the IEEE TRANSACTIONS ON FUZZY SYSTEMS (TFS), IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), IEEE TIST, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS (JBHI), *ACM Transactions on Management Information Systems* (TMIS), *ACM Transactions on Internet Technology* (TOIT), *ACM Transactions on Asian and Low-Resource Language Information Processing* (TALLIP), and *ACM Journal of Data and Information Quality* (JDIQ). He was recognized as the most cited Chinese Researcher, respectively, in 2018–2021 by Scopus/Elsevier.