2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

CaMR: Towards Connotation-aware Music Retrieval on Social Media with Visual Inputs

Lanyu Shang*, Daniel (Yue) Zhang*, Siamul Karim Khan*, Jialie Shen†, Dong Wang*
*Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
{lshang, yzhang40, skhan22, dwang5}@nd.edu
†School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, UK
j.shen@qub.ac.uk

***Abstract*—With the ubiquitous network connectivity and the proliferation of mobile devices, people are increasingly consuming digital contents from social media driven music sharing platforms (e.g., YouTube, Soundcloud). In this paper, we study a novel problem of connotation-aware music retrieval that focuses on the connotation which expresses the implicit feeling or emotion beyond the explicit content in artworks. Our goal is to automatically retrieve relevant music on social media based on the connotation of visual inputs (e.g., images, photos) provided by the users. The problem is challenging as it requires the accurate identification of the implicit connotation from both images and music pieces, and the precise matching of the identified connotation across different data modalities. We develop a connotation-aware music retrieval (CaMR) framework to address the above challenges. Evaluation results from a real-world social media dataset demonstrate that the CaMR framework can retrieve music that is highly relevant to the connotation of the input image.**

I. Introduction

With the ubiquitous network connectivity and the proliferation of mobile devices, people are increasingly consuming digital contents from social media driven music sharing platforms (e.g., YouTube, Soundcloud) [1], [2]. Many solutions have been developed to recommend relevant music with various inputs, including musical contents (e.g., audio, lyrics) and contextual information (e.g., users' social behavior, locations). Connotation is one of the most valuable components in an artwork (e.g., an image or a music piece) that implicitly expresses the abstract idea or inherent emotion beyond the explicit content [3], [4]. For example, people often feel very inspired when they see a high mountain in a photo. Such an inspiring emotion can also be perceived from the musical contents (e.g., an energetic music piece). In this paper, we study a novel research problem of connotation-aware music retrieval on social media with visual inputs. The goal of our problem is to automatically retrieve relevant music on social media based on the connotation of visual inputs (e.g., images, photos) provided by the users. However, it is a challenging task to understand the implicit connotation that might be deeply embedded in both images and music, and effectively establish the connection between them. To address this problem, we develop a Connotation-aware Music Retrieval (CaMR) framework to effectively retrieve music that can reflect the implicit connotation of the input image.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

A significant amount of efforts have been made to address the context-aware challenge in retrieving relevant music on social media, and many contextual factors have been studied by existing work, including locations, music social tags, and users' social behavior [5]. However, these solutions did not explicitly study the implicit connotation expressed in the visual and musical modalities, and cannot be applied to solve our connotation-aware music retrieval problem with visual inputs. Our problem is non-trivial because it requires the accurate identification of the implicit connotation from both images and music pieces as well as the accurate matching of the identified connotation across different data modalities.

Metaphor is often known as a common vehicle to implicitly express the connotation in an artwork (e.g., literature, music, photography). For example, "dove" is often used as a metaphor for "peace and love", and "stars" is often used to refer to "inspiration" in literature. Existing solutions on image-related music identification primarily focus on matching visual objects in the image with keywords in the lyrics, but ignore a fundamental component, metaphor, in both the image and music [6]. Figure 1 shows an image of a tree next to an abandoned house in winter, and the lyric snippets of four candidate songs. Song A and B are relevant to the objects (i.e., "house", "snow") in the given image. However, they do not quite match the metaphorical meaning (connotation) of the input image. For example, Song A shares an ambiance of "peace" and "pleasure", and Song B expresses an emotion of "joy" during Christmas time. Both of them miss the connotation of "sadness" and "depression" that the abandoned house and the bare tree in the image express. In contrast, Song C and D do not contain any keywords matching the objects in the image. However, they both capture the connotation of desolation expressed in the input image and should be considered as relevant in the retrieved music.

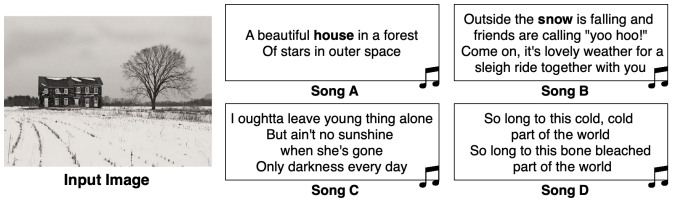

Figure 1: Example of Music Retrieval

In this paper, we develop CaMR, a connotation-aware music retrieval framework that can retrieve music of relevant connotation for a given image. In particular, we develop a metaphor-enriched connotation extraction module to explicitly identify metaphors through a set of semantic and emotion entities extracted from both the image and music. We design a hybrid meta-path learning scheme to retrieve relevant music based on the consistency of the learned connotation between the input image and candidate music. To the best of our knowledge, CaMR is the first *connotation-aware* music retrieval framework for visual inputs from social media users. We evaluate the CaMR framework on a real-world dataset from social media, and the results show that CaMR significantly outperforms state-of-the-arts in terms of the relevance of connotation between the retrieved music and the input image.

II. Related Work

A. Information Network Analysis

Information networks have emerged as an effective scheme to facilitate information retrieval and recommendation [7], [8], [9]. Many recommendation methods have been developed to recommend users, products, places, and explore their relations using information networks [10], [11], [12]. For example, Palumbo *et al.* utilized node2vec to recommend movies by learning knowledge graph embeddings [13]. Wang *et al.* designed a deep-knowledge aware network that utilizes a knowledge graph for news retrieval and recommendations [14]. Zhang *et al.* developed the iPoemRec system that utilizes heterogeneous information networks and neural embedding techniques to perform poetry recommendations [15]. However, none of the existing information network solutions is devoted to studying the problem of connotation-aware music retrieval with visual inputs. In this paper, we develop a novel connotation-aware retrieval framework to solve this problem.

B. Music Recommendation

Existing techniques for music recommendation can be mainly classified into collaborative filtering methods, content-based methods, context-based methods, and hybrid methods [16]. For example, Sánchez-Moreno *et al.* designed a collaborative filtering method to recommend music by learning users' preference of musical artists [17]. Patra *et al.* developed a content-based method to recommend music by exploring the similarity of lyrics [18]. Wang *et al.* developed a context-aware method that utilizes an information network to encode heterogeneous information about users and music objects [19]. However, existing solutions do not explicitly explore the connotation of music pieces and cannot be directly applied to solve the connotation-aware music retrieval problem. In this paper, we propose a novel CaMR framework to retrieve music pieces with relevant connotations to the input image.

III. Problem Formulation

In this section, we formally present the problem of connotation-aware music retrieval based on an input image from the user. We first define a few key terms that will be used in the problem formulation.

Definition 1. ***Image*** *(I): the input image from a user to the music retrieval application.*

Definition 2. ***Music Set*** *(M): a set of N music pieces $M = \{M_1, M_2, \cdots, M_N\}$ from which the connotation-aware music is retrieved. Each music piece M_i contains a set of elements including lyric, audio track, and metadata information.*

Definition 3. ***Connotation****: the abstract idea and inherent emotion conveyed by the content of an artwork [3].*

In this paper, our goal is to find the music pieces that are relevant to the input image in terms of the connotation. The connotative association between an image and a music piece often relates in three aspects: *semantic concepts*, *lyric sentiment*, and *auditory emotion*. In this paper, we define three consistency criteria for the connotative association between the image and music.

Definition 4. ***Semantic Consistency*** *(C^S_{I,M_i}): the consistency of the semantic concepts associated with the connotation in the image I and music piece M_i.*

Definition 5. ***Lyric Consistency*** *(C^L_{I,M_i}): the consistency of the sentiments expressed in the image I and the lyric of music piece M_i.*

Definition 6. ***Audio Consistency*** *(C^A_{I,M_i}): the consistency of the emotions expressed in the image I and the audio of music piece M_i.*

Given the above definitions, our objective is to retrieve a list of music pieces on social media that are relevant to the input image in terms of the overall connotative association (i.e., the sum of semantic consistency, lyric consistency, and audio consistency). In particular, for each input image I, the CaMR framework will output a ranked list of K retrieved music pieces, defined as $D = \{\bar{M}_1, \bar{M}_2, \cdots, \bar{M}_K\}$, where $\bar{M}_1$ represents the most relevant music piece. The connotation-aware music retrieval problem can be formally formulated as:

$$\arg\max_{D} \sum_{k=1}^{K} \left(C^S_{I,\bar{M}_k} + C^L_{I,\bar{M}_k} + C^A_{I,\bar{M}_k} \right) \tag{1}$$

IV. Solution

In this section, we present the Connotation-aware Music Retrieval (CaMR) framework to solve the problem defined in the previous section.

A. Metaphor-enriched Connotation Extraction (MCE) Module

The metaphor-enriched connotation extraction (MCE) module is designed to effectively extract the connotation of image and music by exploring the metaphors and characterizing their relation to semantic concepts, lyric sentiment, and auditory emotion. In particular, we design a metaphor network (MNet) to explicitly extract a set of metaphorical entities that are related to the connotations in image and music, and identify

the metaphors by exploring the links between the entities. We formally define the metaphor network (MNet) as follows.

Definition 7. ***Metaphor Network (MNet):*** *it is defined as an undirected graph* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$*, where* $\mathbf{V}$ *is a set of metaphorical entities, and* $\mathbf{E}$ *is a set of links connecting the entities. The entities and links are described below.*

1) Metaphorical Entity Extraction: First, we extract a set of metaphorical entities in MNet as follows.

- **Semantic Entity** ($\mathbf{V}_s$): $\mathbf{V}_s \subseteq \mathbf{V}$ is the set of entities representing meaningful objects or concepts (e.g., lonely tree, mighty oak) in lyrics.
- **Sentiment Entity** ($\mathbf{V}_t$): $\mathbf{V}_t \subseteq \mathbf{V}$ is the set of entities representing the overall sentiment (e.g., joy, sad, neutral) conveyed in the lyric of a music piece.
- **Audio Entity** ($\mathbf{V}_a$): $\mathbf{V}_a \subseteq \mathbf{V}$ is the set of entities representing the overall acoustic feeling (e.g., relaxed, depressed) of the auditory content in a music piece.
- **Metadata Entity** ($\mathbf{V}_m$): $\mathbf{V}_m \subseteq \mathbf{V}$ is the set of entities representing the metadata information (e.g., genre, theme) of a music piece.

We observe that the above heterogeneous entities are closely related to the metaphors in artworks, which serve as a key vehicle to express the connotation. We extract these metaphorical entities from multiple data modalities to effectively capture the implicit connotation conveyed in the artworks. In particular, we extract semantic entities in lyrics by identifying the descriptive adjective-noun phrases with the part-of-speech tagger and extract sentiment entities using the Watson Tone Analyzer. To extract the audio entities from the musical content, we leverage a set of audio features (e.g., positiveness, energy, loudness, danceability) from Spotify and map them to the corresponding audio entities (e.g., "excited", "relaxed", "stressed", and "depressed"). Additionally, we collect a set of categorical metadata features (e.g., genre, theme) that are found to be relevant in identifying the connotation in musical artworks.

2) Metaphorical Relation Extraction: We observe that the connections between the heterogeneous entities contain rich information in identifying the metaphorical relation between entities, which are essential in capturing the implicit connotation. For example, the lyric semantic entity "lonely shadow" often relates to the "negative" sentiment of loneliness and is accompanied by a melody of "depressed" auditory emotion. Therefore, we explicitly model the metaphorical relations as links between entities in MNet. Specifically, we use $e_{v,v'} \in \mathbf{E}$ to denote the link between entity v and v'. For each link $e_{v,v'}$, we define a weight factor $\omega_{v,v'} \in (0, 1]$ to indicate the closeness between entity v and v'. In particular, we focus on the following five types of links that are closely related to the metaphorical relation between entities: *semantic-semantic*, *semantic-sentiment*, *semantic-audio*, *semantic-metadata*, *sentiment-audio*. We also define the concept of meta-path (i.e., a collection of consecutive links between two entities) to capture the indirect connection between entities.

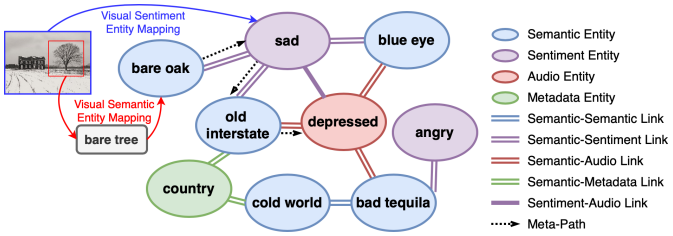

Figure 2: An example of MNet

Figure 2 shows an example of the constructed MNet. In this network, we can capture not only the direct relationship between a pair of entities, but also the hidden connections between any pair of entities using the meta-paths which are essential in characterizing the connotation of both image and music. For example, a direct link can capture the sentiment (e.g., "sad") of a semantic entity "bare oak", and a meta-path can further capture the hidden relationship between the semantic entity "bare oak" and the audio entity "depressed".

B. Visual Entity Mapping (VEM) Module

In this section, we present the visual entity mapping module that is designed to accurately map visual entities extracted from the input image to the relevant entities in MNet. In particular, we focus on two types of visual entities: 1) the *visual semantic entity* that represents a visual object in the image, and 2) the *visual sentiment entity* that represents the overall sentiment of an image.

To identify the visual semantic entities, we first detect the visual objects in the input image and extract a description for each visual object and map the extracted visual descriptions to the relevant semantic entities in MNet. For example, the descriptions of the visual objects extracted for the input image shown in Figure 1 include "snowy field", "old house", and "bare tree". After that, we perform the mapping between the extracted visual semantic entities and the semantic entities in MNet by comparing their linguistic similarity (i.e., cosine similarity of the word embeddings). Our mapping strategy is designed to curb the "unknown entity" issue by allowing a certain degree of ambiguity during the mapping process (e.g., mapping "bare tree" to "bare oak"). Moreover, we also extract the visual sentiment entity of the input image to help us better connect the input image with music pieces that share similar connotations. In particular, we take the sentiment of the description of the visual objects as the visual sentiment entity of the image and map it to the corresponding sentiment entity in MNet. After the mapping process, the input image is connected to MNet via a set of extracted visual entities. For example, we can connect the input image in Figure 2 with the "bare oak" semantic entity and the "sad" sentiment entity in MNet through the VEM.

C. Hybrid Meta-Path Learning (HML) Module

1) Connotation Representation: In the HML module, we capture the connotation of an image and a music piece as a collection of meta-paths in MNet. First, we develop a weighted Random Walk scheme to collect the meta-paths for

an input image. Random Walk is a method commonly used to retrieve information from graph-structured data [20]. In the CaMR framework, we perform two random walk processes to explore and record meta-paths for the visual entities extracted by the VEM module. In particular, we define the semantic-oriented meta-path (MP_{sm}) and the sentiment-oriented meta-path (MP_{st}) as follows.

Definition 8. ***Semantic-oriented Meta-Path*** *(MP_{sm}): a random walk that traverses MNet M_{sm} times from each visual semantic entity of the image and records at most N_{sm} entities.*

Definition 9. ***Sentiment-oriented Meta-Path*** *(MP_{st}): a random walk that traverses MNet M_{st} times from each visual sentiment entity of the image and records at most N_{st} entities.*

In particular, we design a weighted random walk process in each traversal to efficiently retrieve the relevant connotation-aware meta-paths for a given input image. The goal of the weighted random walk process is to allow the neighborhood entity with a closer metaphorical relation to the current entity to be captured with a higher probability in the traversals. Formally, let v_i be the current entity and $V_{neighbor}$ be the set of entities directly connected with v_i. For any $v_j \in V_{neighbor}$, the entity-specific probability p_{v_i,v_j} of selecting v_j to be the next entity in the traversal is computed as:

$$p_{v_i,v_j} = \frac{n_{i,j} \cdot \omega_{v_i,v_j}}{\sum_{k=1}^{m} n_{i,k} \cdot \omega_{v_i,v_k}} \tag{2}$$

where $n_{i,j}$ is the number of co-appearances of entity v_i and v_j in the same data source, m is the total number of entities in the neighborhood of v_i, and ω_{v_i,v_j} is the weight factor indicating the closeness between entity v_i and v_j.

To learn the representation of connotation in music pieces, we follow the same random walk mechanism and represent each music piece as a collection of meta-paths retrieved from MNet. In particular, we use the semantic and sentiment entities extracted from each music piece as entities to start the traversal, and collect the semantic-oriented and sentiment-oriented meta-paths. Finally, we obtain two sets of meta-paths that represent the connotation of both the input image and each candidate music piece.

2) Music Identification for Visual Inputs: Next, we leverage the retrieved set of meta-paths for both the image and music to retrieve music pieces based on the overall connotation consistency (i.e., the sum of semantic consistency, lyric consistency, and audio consistency defined in Section III) between them. Specifically, let $\mathbf{MP_I}$ and $\mathbf{MP_{M_i}}$ be the retrieved meta-paths for an image I and a music piece M_i of interest, respectively. For example, the image in Figure 1 can be represented as a meta-path of "bare oak $\rightarrow$ sad $\rightarrow$ old interstate $\rightarrow$ depressed", and Song D in Figure 1 can be represented as "cold world $\rightarrow$ country $\rightarrow$ old interstate $\rightarrow$ sad" as shown in Figure 2. We then compute the semantic consistency (C^S_{I,M_i}) as the cosine similarity between entities in $\mathbf{MP_I}$ and $\mathbf{MP_{M_i}}$ as follows:

$$C^S_{I,M_i} = \cos\left(\mathbf{T_s}(\mathbf{MP_I}), \mathbf{T_s}(\mathbf{MP_{M_i}})\right) \tag{3}$$

where $\mathbf{T_s}(\cdot)$ is the average of the linguistic embeddings.

Moreover, the lyric consistency and audio consistency are closely related to the sentiment entities and audio entities in MNet which contain a set of pre-defined text sentiment and audio emotion categories. We measure the lyric consistency (C^L_{I,M_i}) and audio consistency (C^A_{I,M_i}) using the Jaccard coefficient between $\mathbf{MP_I}$ and $\mathbf{MP_{M_i}}$ as follows:

$$C^L_{I,M_i} = \frac{|\mathbf{MP_I} \cap \mathbf{MP_{M_i}} \cap \mathbf{V_t}|}{|(\mathbf{MP_I} \cup \mathbf{MP_{M_i}}) \cap \mathbf{V_t}|} \tag{4}$$

$$C^A_{I,M_i} = \frac{|\mathbf{MP_I} \cap \mathbf{MP_{M_i}} \cap \mathbf{V_a}|}{|(\mathbf{MP_I} \cup \mathbf{MP_{M_i}}) \cap \mathbf{V_a}|} \tag{5}$$

where $\mathbf{V_t}$ and $\mathbf{V_a}$ are the sets of sentiment entities and audio entities in MNet, respectively.

Finally, we rank the candidate music pieces based on the sum of these three consistency measurements (i.e., $C^S_{I,M_i} + C^L_{I,M_i} + C^A_{I,M_i}$), and output the top K music pieces with the maximum overall connotation consistency.

V. Evaluation

A. Dataset and Experiment Setup

1) Dataset: In the evaluation, we collect a set of ground truth labels of 5250 image-to-song pairs using a crowdsourcing approach. For each song, we also collect its lyrics from genius.com and the corresponding music video link on YouTube. We then send the images and songs to Amazon Mechanical Turk (MTurk), one of the largest crowdsourcing platforms, to collect the ground truth labels (for the purpose of evaluation only). In the task, an MTurk worker is asked to annotate three image-to-song pairs on whether they feel each image-to-song pair is relevant after carefully reviewing the content of the image, and the lyric and music video of the song.

2) Baselines: We compare the CaMR with the following state-of-the-art baseline methods from recent literature.

- **ASONAM19**: a poetry recommender system recommending poems using visual input from social media [15]. We adapt it to retrieve music by comparing the lyrics with the recommended poem for the given input image.
- **RecSys18**: an image-based music retrieval system that retrieves relevant songs of a given image using a representation learning framework [6].
- **Im2P**: a neural network scheme that generates the description of an input image [21]. The music is ranked based on the linguistic similarity between the description of the image and the lyric of the song.

Please note that none of the above baselines solve the problem of connotation-aware music retrieval. So we have to adapt these methods to perform the corresponding music retrievals for the purpose of comparison. To ensure a fair comparison with CaMR, we carefully tune the corresponding parameters of each baseline to achieve their best performance.

3) Evaluation Metrics: We adopt the following evaluation metrics that are commonly used in evaluating the performance of information retrieval and recommendation.

- **Precision (Pre@K)**: is the percentage of songs in the top-K retrieval list that are actually relevant to the input image. Formally, $Pre@K = \frac{\{Retrieved\ Songs\} \cap \{Relevant\ Songs\}}{K}$
- **Recall (Recall@K)**: is the percentage of relevant songs that are identified in the top-K retrieval list. Formally, $Recall@K = \frac{\{Retrieved\ Songs\} \cap \{Relevant\ Songs\}}{\{Relevant\ Songs\}}$

B. Music Retrieval Performance

We first evaluate the music retrieval performance of the CaMR scheme in comparison to the aforementioned baseline methods. The evaluation results are summarized in Table I. We observe that the CaMR scheme significantly outperforms all compared baselines on all evaluation metrics. For example, CaMR achieves 5.2% and 2.4% performance gains compared to the best-performing baseline (i.e., Im2P) in terms of Pre@3, and Pre@5, respectively. Such performance gains show that CaMR can retrieve music pieces that are more relevant to the implicit connotation expressed in the image. We also note that the ASONAM19 approach does not perform well because it mainly relies on the information network constructed from classic poems and ignores the gap between classic literature and modern lyrics. We further observe that object similarity based approaches (e.g., RecSys18 and Im2P) fail to effectively retrieve relevant songs because they primarily focus on matching the object in the image with the keywords in the lyrics and ignore the connotation of both images and lyrics.

Table I: Precision and Recall

	Pre@3	**Pre@5**	**Recall@3**	**Recall@5**
CaMR	**0.626**	**0.596**	**0.033**	**0.052**
ASONAM19	0.547	0.540	0.028	0.047
RecSys18	0.526	0.520	0.027	0.045
Im2P	0.595	0.582	0.032	0.041

VI. Conclusion

In this paper, we develop CaMR, a connotation-aware music retrieval framework to retrieve the music of relevant connotation for a given image. The CaMR framework explicitly extracts metaphors in the image and music and establish the connection between them through a metaphor information network. We evaluate the CaMR on a real-world dataset, and the results show that CaMR achieves significant performance gains compared to the state-of-the-art baselines by retrieving more connotatively relevant music for the visual inputs.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CNS-1845639, CNS-1831669, Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] L. Shang, D. Y. Zhang, M. Wang, and D. Wang, "Vulnercheck: a content-agnostic detector for online hatred-vulnerable videos," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 573–582.

[2] D. Y. Zhang, L. Song, Q. Li, Y. Zhang, and D. Wang, "Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 901–910.

[3] C. C. Liem, M. Larson, and A. Hanjalic, "When music makes a scene," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 15–30, 2013.

[4] J. Marshall and D. Wang, "Mood-sensitive truth discovery for reliable recommendation systems in social sensing," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 2016, pp. 167–174.

[5] D. Zhang, Y. Zhang, Q. Li, and D. Wang, "Sparse user check-in venue prediction by exploring latent decision contexts from location-based social networks," *IEEE transactions on Big Data*, 2019.

[6] C.-C. Hsia, K.-H. Lai, Y. Chen, C.-J. Wang, and M.-F. Tsai, "Representation learning for image-based music recommendation," *arXiv preprint arXiv:1808.09198*, 2018.

[7] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.

[8] M. T. Rashid and D. Wang, "Covidsens: a vision on reliable social sensing for covid-19," *Artificial Intelligence Review*, pp. 1–25, 2020.

[9] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, "On credibility estimation tradeoffs in assured social sensing," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1026–1037, 2013.

[10] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *arXiv preprint arXiv:2003.00911*, 2020.

[11] L. Shang, D. Y. Zhang, M. Wang, S. Lai, and D. Wang, "Towards reliable online clickbait video detection: A content-agnostic approach," *Knowledge-Based Systems*, vol. 182, p. 104851, 2019.

[12] L. Shang, Y. Zhang, D. Zhang, and D. Wang, "Fauxward: a graph neural network approach to fauxtography detection using social media comments," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–16, 2020.

[13] E. Palumbo, G. Rizzo, R. Troncy, E. Baralis, M. Osella, and E. Ferro, "Knowledge graph embeddings with node2vec for item recommendation," in *European Semantic Web Conference*. Springer, 2018.

[14] H. Wang, F. Zhang, X. Xie, and M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1835–1844.

[15] D. Zhang, B. Ni, Q. Zhi, T. Plummer, Q. Li, H. Zheng, Q. Zeng, Y. Zhang, and D. Wang, "Through the eyes of a poet: classical poetry recommendation with visual input on social media," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 333–340.

[16] D. Das, L. Sahoo, and S. Datta, "A survey on recommendation system," *International Journal of Computer Applications*, vol. 160, no. 7, 2017.

[17] D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García, "A collaborative filtering method for music recommendation using playing coefficients for artists and users," *Expert Systems with Applications*, vol. 66, pp. 234–244, 2016.

[18] B. G. Patra, D. Das, and S. Bandyopadhyay, "Retrieving similar lyrics for music recommendation system," in *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, 2017.

[19] D. Wang, G. Xu, and S. Deng, "Music recommendation via heterogeneous information graph embedding," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 596–603.

[20] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.

[21] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *IEEE CVPR*, 2017, pp. 317–325.