Commonality Augmented Disentanglement for Multimodal Crowdfunding Success Prediction

Jiayang Li*†, Xovee Xu*†, Yili Li*, Ting Zhong*, Kunpeng Zhang‡ and Fan Zhou*§¶
*University of Electronic Science and Technology of China, Chengdu, Sichuan, China
‡University of Maryland, College park, MD, USA
§Kash Institute of Electronics and Information Industry, Kashi, Xinjiang, China
jy.li@std.uestc.edu.cn, xovee.xu@gmail.com, {zhongting, lylxzr}@uestc.edu.cn, kpzhang@umd.edu, fan.zhou@uestc.edu.cn
†Equal Contribution, ¶Corresponding

***Abstract*—Online crowdfunding platforms have been gaining increasing popularity due to their convenience in soliciting social capital from the public. These platforms offer valuable opportunities for fundraisers to bring their creative products to life and support pro-social projects. However, the relatively low success rate of crowdfunding campaigns highlights the need for better strategies. While existing studies have explored various factors that contribute to crowdfunding success, they often overlook the intricate relationships between different aspects of crowdfunding. In this paper, we propose a novel model called Commonality Augmented Multimodal Disentanglement (CAMD) for predicting crowdfunding success. It can disentangle the roles of different data modalities, separating common factors from specific ones. Besides, we enhance the disentangled commonality using an augmentation network to achieve balanced representation of the skewed interrelations between different modalities. At last, we introduce a cross-attention-based multimodal fusion mechanism that further improves model performance by highlighting the crucial role of crowdfunding attributes. Experiments conducted on two large-scale crowdfunding datasets demonstrate the effectiveness and generalizability of our model.**

***Index Terms*—Crowdfunding success prediction, multimodal disentanglement, data augmentation.**

I. Introduction

Crowdfunding presents valuable opportunities for both individuals and entrepreneurs to raise capital from the public. However, crowdfunding projects face challenges such as a low success rate and difficulty in reaching their funding goals [1], [2]. This issue hinders the overall growth and potential of crowdfunding as a financing method. To address this challenge, the concept of crowdfunding success prediction has emerged, aiming to help fundraisers organize their campaigns more effectively [3], [4]. In addition, it can provide valuable insights to crowdfunding platforms, enabling them to make more accurate and strategic recommendations, ultimately enhancing their social impact and financial success [5], [6].

This work is supported in part by the National NSF of China under Grant Nos. 62072077 and 62176043, and by the Kashgar Science & Technology Bureau under Grant No. KS2023025.

This is the author's version of the paper presented at ICASSP '25 in Hyderabad, India, on Apr 6–11, 2025.

doi: 10.1109/ICASSP49660.2025.10889564

Code & Data: https://github.com/Jiayang-L1/camd

Previous research has delved into the determinants of crowdfunding success from various perspectives. Several studies have explored the impact of linguistic style on crowdfunding success [7], [8]. Others have investigated the role of visual cues in attracting investors [9], [10]. Beyond individual components, some studies have taken a holistic approach, combining textual, visual, and social information to predict crowdfunding success based on social capital, human capital, and processing level theories [4], [11]. Furthermore, models have been built to predict crowdfunding success using deep neural networks, such as MDL [12] and DCAN [13].

However, existing methods face several notable limitations. Visual and textual cues are two critical content modalities used to convey the appeal and quality of the crowdfunding project. Each modality contributes uniquely, e.g., text presents detailed background information and desired outputs, while images vividly showcase products or crowdfunding causes. Besides, the synergy between modalities manifest through the latent dependency between them, e.g, the impact of tone could be amplified by using complementary hues. While it is hard for hand-crafted feature engineering methods to capture such complicated interactions, existing deep crowdfunding methods also fail to capture the latent relation of modalities. For example, the oversimplified structure of MDL [12] struggle to mining both the unique and synergetic effect of modalities, and DCAN [13] fuses modalities without consideration of synergy.

Disentangling multimodal data is an effective approach for thoroughly uncovering the latent structure within inherently heterogeneous modalities [14]–[17]. For instance, in [14] the researchers disentangles the multimodal data into modality-invariant and modality-specific space, corresponding to synergetic (commonality) and complementary (specificity) effects, respectively. The synergetic effect, often referred to as commonality, is readily apparent in typical applications of disentanglement, such as video sentiment analysis. In such cases, textual, visual, and acoustic modalities are all derived from the same video segment, yielding to aligned semantics. However, in the context of crowdfunding, modalities have relatively weak dependencies, without mandatory restrictions on their semantics. The lack of inherent alignment makes it difficult to identify and extract commonalities, posing a unique

challenge in crowdfunding disentanglement.

To address the above two challenges, we propose a novel commonality augmented approach aimed at magnifying observations related to disentangled commonality, which plays an important role in enhancing the efficiency of AI models while also improving their robustness [18]–[20]. For example, LeMDA [18] suggests augmenting multimodal representation in latent feature spaces without constraints on the identities and associations of modalities. In light of this, we incorporate variational autoencoders (VAE) into multimodal disentangle learning, which augments the commonality demonstrated in the latent modality-invariant space. It allows us to strike a subtle balance between dominant specificity and relatively insufficient commonality, making the disentangled process more effective in context of complex crowdfunding dynamics. Besides, we design a cross-attention-based multimodal fusion method to learn crowdfunding modalities as well as their interactions and complementarity, enabling us to obtain interactive and comprehensive representations of crowdfunding data, and achieve improved success prediction performance.

In summary, we make the following contributions: (1) We propose a novel **C**ommonality **A**ugmented **M**ultimodal **D**isentanglement framework (CAMD) for improving the crowdfunding success prediction performance. CAMD disentangles the intricate effects within crowdfunding data into commonality and specificity components. (2) Furthermore, it simultaneously magnifies the disentangled commonality through a commonality-augmented network within an invariant space, thereby facilitating balanced disentanglement learning. (3) We introduce a cross-attention mechanism to fuse metadata and modality representations, allowing for a comprehensive exploration of the interaction and complementarity of metadata and content modalities on crowdfunding success prediction. (4) Extensive experiments conducted on two large-scale crowdfunding datasets demonstrate the effectiveness and generalizability of CAMD.

II. Methodology

Problem Statement. Online crowdfunding operates primarily in two major modes, namely "All-Or-Nothing" and "Keep-It-All" [21]. The former mandates reaching the fundraising goal to receive funds, while the latter permits fund withdrawal regardless of fundraising progress. Previous research on crowdfunding success prediction has predominantly focused on binary classification – classifying projects as either *successful* or *failed* based on whether fundraising goals are met within a predefined time period. However, this binary setting does not adequately capture the dynamics of "Keep-It-All" projects and introduces biases towards overfunded or nearly successful projects. We formulate crowdfunding success prediction as a regression problem, where the goal is to predict the amount of funds raised. Specifically, for N crowdfunding projects denoted as $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$, each project c_i is associated with a variety of multimedia content. This includes textual project description T_i, visual campaign photo P_i, and project metadata M_i. Our primary objective is to build a prediction model capable of learning from multimodal crowdfunding content to estimate the amount of money raised, y_i.

Framework Overview. The proposed CAMD model consists of four main components: (1) feature extraction from multimodal crowdfunding content; (2) disentangled learning on modality-invariant and modality-specific spaces; (3) cross-attention-based multimodal fusion; and (4) invariant augmentation on the commonality representation. At last, we optimize CAMD via a task loss, a disentanglement loss, and an augmentation loss. An overview of CAMD is illustrated in Fig. 1.

A. Crowdfunding Feature Extraction

Crowdfunding projects typically have three types of content: (1) **metadata**: This includes essential project attributes such as fundraising goal, category, country, number of words, number of photos, and crowdfunding mode. The categorical features are processed via one-hot encoding; (2) **textual content**: The textual content of crowdfunding projects is derived from the project description. Following [12], [13], we use GloVe [22] to generate a 300-dim embedding for each word in T_i. These embeddings are then aggregated into a comprehensive textual embedding $\mathbf{t}_i$ by applying an average pooling operation; (3) **visual signals**: Visual cues of crowdfunding projects are mainly conveyed through the cover photos. To extract semantic information from these images, we employ a pre-trained ResNet-152 [23], which generates a 2048-dimensional embedding $\mathbf{p}_i$ for each cover photo P_i.

B. Multimodal Content Disentanglement

Visual and textual signals play crucial roles in crowdfunding campaigns, serving as essential means to convey information to potential investors and donors [24]. These content elements significantly impact the success of crowdfunding campaigns, both within their respective modalities and through their cooperative interactions. We disentangle the extracted multimodal features into two distinct spaces: modality-invariant and modality-specific. The modality-invariant space captures the commonality between visual and textual crowdfunding content. Meanwhile, modality-specific spaces are designed to encapsulate the unique characteristics of each modality, contributing to the overall success of the crowdfunding campaign.

To achieve the abovementioned disentanglement, we first unify the dimensionality of visual and textual embeddings. For high-dimensional visual embedding $\mathbf{p}_i$, we map them to a lower-dimensional space $\mathbf{p}'_i = \texttt{Conv1d}(\texttt{AE}(\mathbf{p}_i))$ using a combination of an autoencoder and a 1D convolution layer. For textual embedding $\mathbf{t}_i$, we directly feed them to a fully-connected (FC) layer: $\mathbf{t}'_i = \texttt{FC}(\mathbf{t}_i)$. After the dimension alignment for both content modalities, we proceed to project them into modality-invariant and modality-specific spaces. This is done through a commonality encoder $\texttt{Enc}_c$ and two specificity encoders $\texttt{Enc}_t$ and $\texttt{Enc}_p$. The commonality encoder aims to capture the consistent crowdfunding semantics in both visual and textual modalities, while the specificity encoders are more focused on individual modalities and their relations to crowdfunding success.

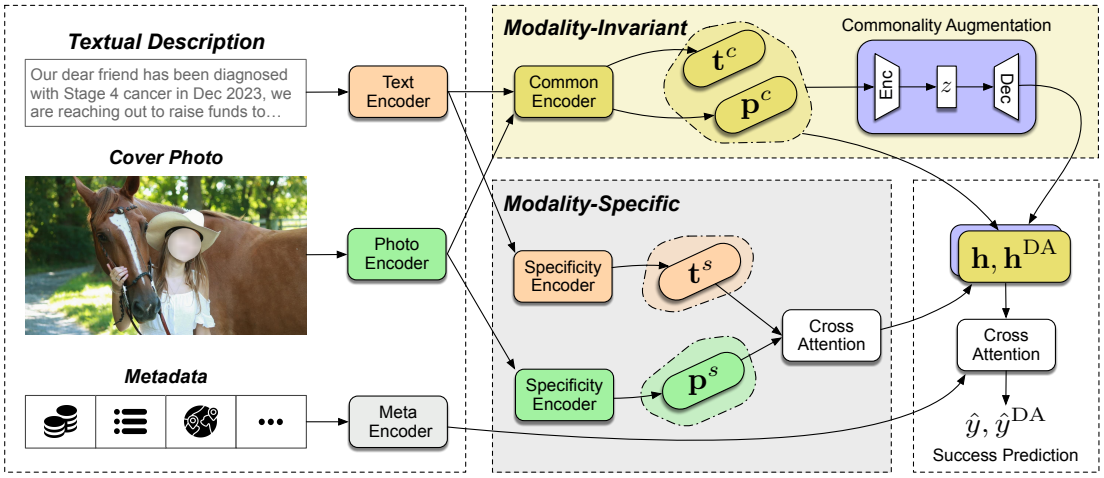


Fig. 1: Overview of our proposed model.

Commonality and specificity representations are defined as:

$$\mathbf{t}^c = \texttt{Enc}_c(\mathbf{t}'; \theta_c), \quad \mathbf{p}^c = \texttt{Enc}_c(\mathbf{p}'; \theta_c) \tag{1}$$

$$\mathbf{t}^s = \texttt{Enc}_t(\mathbf{t}'; \theta_t), \quad \mathbf{p}^s = \texttt{Enc}_p(\mathbf{p}'; \theta_p) \tag{2}$$

where the common encoder $\texttt{Enc}_c(\cdot; \theta_c)$ shares the same parameters θ_c through two modalities, while specific encoders have different parameters θ_t and θ_p for each modality.

C. Fusion & Prediction

Now the disentangled representations are fused together to generate a joint modality representation. We obtain commonality representation $\mathbf{h}_c = \mathbf{t}^c \oplus \mathbf{p}^c$ by concatenating representations in modality-invariant subspace, considering the semantic similarities they demonstrated. Specificity representations, on the other hand, capture unique characteristics of each modality. We implement a bidirectional cross-attention layer between them to exploit the complementary information, $\mathbf{h}_s = \texttt{CrossAttn}(\mathbf{t}^s, \mathbf{p}^s)$. The final representation of modalities is obtained as $\mathbf{h} = \mathbf{h}_c \oplus \mathbf{h}_s$.

Furthermore, a cross-attention between metadata $\mathbf{m}$ and $\mathbf{h}$ is performed to explore the synergetic effects between crowdfunding content and metadata: $\mathbf{h}' = \texttt{CrossAttn}(\mathbf{h}, \mathbf{m})$. At last, the task prediction is obtained by $\hat{y} = \texttt{FC}(\mathbf{h}')$.

D. Invariant Augmentation

Compared with the well-captured specificity, the model's observation towards commonality is relatively suppressed. To ensure fair disentanglement on skewed modalities, we propose to use another data augmentation network DA in parallel with disentanglement network to magnify the synergetic effect within modalities. The augmentation network DA takes the commonality representation $\mathbf{h}_c$ as input,and outputs an augmented commonality representation $\mathbf{h}_c^{\mathrm{DA}}$. Specifically, the network is constructed in a variational inference manner [25], where we have FC-based encoder and decoder for learning an augmented representation $\mathbf{h}_c^{\mathrm{DA}} = \texttt{DA}(\mathbf{h}_c)$. The commonality within modalities is augmented and the imbalance between two spaces is better controlled.

During training, we first obtain two commonality representations $\mathbf{h}_c$ and $\mathbf{h}_c^{\mathrm{DA}}$ for each project c_i. Then we feed them through the rest of the fusion network and obtain two predictions $\hat{y}$ and $\hat{y}^{\mathrm{DA}}$, which are evaluated together with y.

E. Optimization

The overall optimization of CAMD is by minimizing:

$$\mathcal{L}_{\texttt{all}} = \mathcal{L}_{\texttt{task}} + \mathcal{L}_{\texttt{disentangle}} + \mathcal{L}_{\texttt{augment}}. \tag{3}$$

Task Loss. A standard regression loss – mean squared error (MSE) – is used to evaluate model performance:

$$\mathcal{L}_{\texttt{task}} = \texttt{MSE}(\hat{y}, y) + \texttt{MSE}(\hat{y}^{\mathrm{DA}}, y). \tag{4}$$

Disentanglement Loss. This loss ensures the disentangled encoders to correctly capture decoupled modality information:

$$\mathcal{L}_{\texttt{disentangle}} = \alpha(\mathcal{L}_{\texttt{sim}} + \mathcal{L}_{\texttt{diff}} + \mathcal{L}_{\texttt{recon}}). \tag{5}$$

Specifically, $\mathcal{L}_{\text{sim}}$ is the consistency constraint ensuring generated common representations capture similar semantics, $\mathcal{L}_{\text{diff}}$ ensures that the information learned by commonality and specificity representations are not overlapping, and $\mathcal{L}_{\text{recon}}$ prevent trivial information learning. We have them defined as:

$$\mathcal{L}_{\text{sim}} = 1 - \cos(\mathbf{t}^c, \mathbf{p}^c), \quad \mathcal{L}_{\text{recon}} = \frac{1}{2} \sum_{\mathbf{u} \in \{\mathbf{t}, \mathbf{p}\}} \texttt{MSE}(\mathbf{u}', \hat{\mathbf{u}}),$$

$$\mathcal{L}_{\text{diff}} = \frac{1}{3}(\max(0, \cos(\mathbf{t}^s, \mathbf{p}^s)) + \sum_{\mathbf{u} \in \{\mathbf{t}, \mathbf{p}\}} \max(0, \cos(\mathbf{u}^c, \mathbf{u}^s))) \tag{6}$$

where $\mathcal{L}_{\text{sim}}$ and $\mathcal{L}_{\text{diff}}$ are realized by minimizing the cosine embedding loss, $\mathcal{L}_{\text{recon}}$ minimizes the MSE between raw modality representation $\mathbf{u}'$ and disentangled representation $\hat{\mathbf{u}} = \texttt{Dec}(\mathbf{u}^c \oplus \mathbf{u}^s; \theta)$, where $\mathbf{u} \in \{\mathbf{t}, \mathbf{p}\}$, $\texttt{Dec}(\cdot; \theta)$ is a decoder.

Augmentation Loss. Augmentation network is optimized by:

$$\mathcal{L}_{\texttt{augment}} = -\beta_1 \mathcal{L}_{\texttt{task}}^{\mathrm{DA}} + \beta_2(\mathcal{L}_{\texttt{consist}} + \mathcal{L}_{\texttt{VAE}}), \tag{7}$$

where β_1, β_2 are hyper-parameters controlling augmentation effect. The first term $-\mathcal{L}_{\texttt{task}}^{\mathrm{DA}}$ maximizes the task loss made

TABLE I: Descriptive statistics of datasets.

Dataset	# projects	"Keep-It-All"	avg goal	avg raised
GoFundMe	14,203	100%	7,385	3,672
Indiegogo	5,804	91.4%	7,435	2,230

TABLE II: Ablation Results (RMSE)

Dataset	CAMD	w/o fusion	w/o disen.	w/o augment
GoFundMe	0.7160	0.8646	0.7219	0.7199
Indiegogo	1.6140	1.7754	1.6498	1.6449

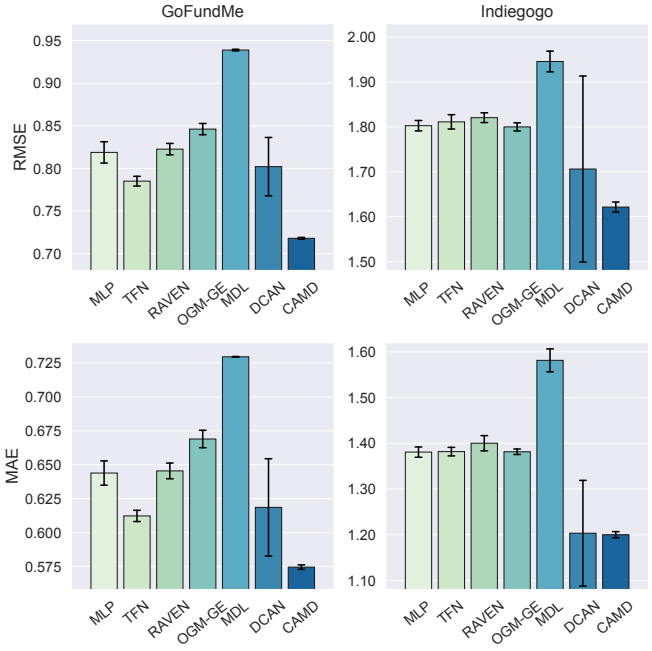
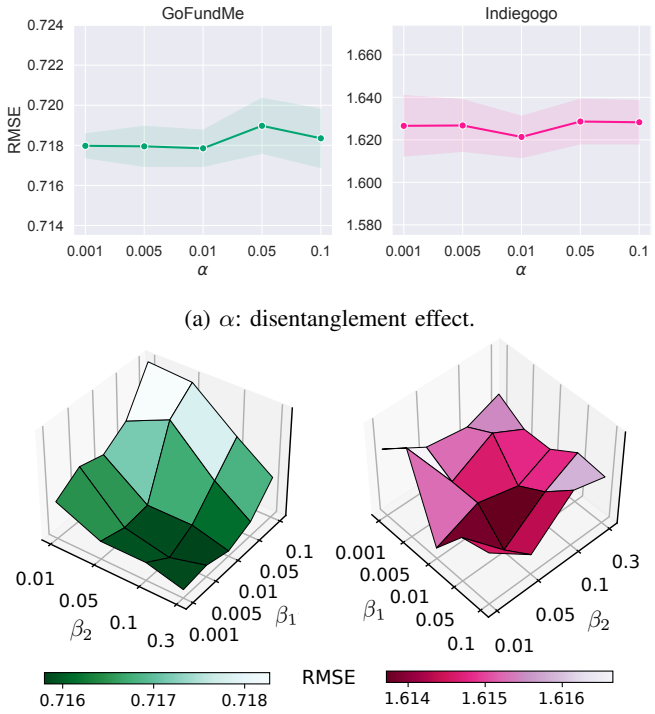


Fig. 2: Crowdfunding success prediction results. We run each model five times and report mean values with std.



(a) α: disentanglement effect.

(b) β_1 and β_2: augmentation effect.

Fig. 3: Hyperparameter analysis.

by augmentation network, encouraging original network to be updated through augmented representations in an adversarial way. $\mathcal{L}_{\texttt{consist}}$ preserves semantic structure of network input by restricting distributional discrepancies of original and augmented representations, where we achieve it through minimizing the L_2 distance between $\mathbf{h}_c$ and $\mathbf{h}_c^{\text{DA}}$. $\mathcal{L}_{\text{VAE}}$ is used to regularize augmentation network, we only adopt the KL regularizer on the encoder distribution.

III. Experiments

Experimental Setup. We evaluate our model on GoFundMe [26] and Indiegogo [27]. GoFundMe's backers contribute to projects without expecting any financial return and all the projects are following the "Keep-It-All" mode. Moreover, GoFundMe projects have no specified time limit and users are at liberty to contribute for as long as the project remains active. Indiegogo is a leading reward-based platform, where backers contribute to projects in exchange for tangible or intangible rewards. Yet Indiegogo provides both "Keep-It-All" and "All-Or-Nothing" modes, and the former is more favored by fundraisers. Projects without description or cover photo are removed, and the prediction time is 30 days, a typical deadline for crowdfunding projects [1]. See data statistics in Table I.

To demonstrate the effectiveness of our proposed model, we compare it with the following baselines: (1) an MLP model with textual, visual, and metadata as input features concatenated; (2) deep multimodal learning methods: **TFN** [28], **RAVEN** [29] and **OGM-GE** [30]; (3) dedicated multimodal crowdfunding success prediction methods: **MDL** [12] and **DCAN** [13]. For a fair comparison, all baselines use the same input features as our model. We use Rooted MSE and mean absolute error (MAE) as metrics.

Performance Comparison. The prediction performance comparison results on the two datasets are shown in Fig. 3. Our proposed CAMD model consistently outperforms all baselines across datasets, in terms of both RMSE and MAE metrics. Upon which, we argue that the deficiency of baselines is two-fold. First, our commonality augmented disentanglement learning framework is more suitable and effective for crowdfunding success prediction than other methods. MDL simply concatenates features from different sources, failing to capture intricate crowdfunding multimodal interrelations; the cross-attention between textual and visual information implemented by DCAN is unable to extract balanced multimodal representations. Second, the importance of metadata features is underestimated by previous works, which simply fuse metadata with other modalities or disengage metadata directly. By contrast,

our proposed CAMD not only achieved balanced multimodal representations via modality-invariant and modality-specific spaces, but also introduced an invariant augmentation network to further enhance prediction performance.

Ablation Study. To investigate the roles of the important components in CAMD, we conduct ablation studies by removing one component at a time and observing performance changes. Specifically, (1) *CAMD w/o fusion*: we replace the fusion module with naïve concatenation between textual, visual, and metadata features; (2) *CAMD w/o disentangle*: the multimodal content disentanglement module is replaced by a cross-attention network; (3) *CAMD w/o augment*: the data augmentation network on commonalities is removed, retaining only disentanglement. Table II shows that the fusion network is critical for prediction. It is evident that different modalities between multimodal content exhibit complex interrelations and simple fusion networks are insufficient to capture rich semantics behind crowdfunding content. In addition, our disentanglement module outperforms the tangled cross-attention network, highlighting the need to separately model commonality and specific representations. The augmentation network reconciles skewed modalities, improving multimodal representation.

Hyperparameter Analysis. To analyze the effects of balancing hyperparameters α, β_1, and β_2, we tested various values, as shown in Fig. 3. The optimal values are $\alpha = 0.01$, $\beta_1 = 0.05$, and $\beta\_2 = 0.1$.

IV. Conclusion

In this work, we introduce a novel multimodal disentanglement framework CAMD that features a commonality-specificity disentangled network within an invariant augmentation for crowdfunding success prediction. Our model disentangles the multimodal crowdfunding content into modality-invariant and modality-specific spaces and incorporates a cross-attention network to fuse metadata and multimodal representations. Extensive experiments on two large-scale datasets verified the effectiveness and generalizability of CAMD.

References

[1] E. Mollick, "The dynamics of crowdfunding: An exploratory study," *Journal of Business Venturing*, vol. 29, no. 1, pp. 1–16, 2014.

[2] J. Huang, H. Shen, Q. Cao, L. Cai, and X. Cheng, "How medical crowdfunding helps people? A large-scale case study on the Waterdrop fundraising," in *ICWSM*, 2021, pp. 220–229.

[3] C. Lagazio and F. Querci, "Exploring the multi-sided nature of crowdfunding campaign success," *Journal of Business Research*, vol. 90, pp. 318–324, 2018.

[4] J.-Y. Yeh and C.-H. Chen, "A machine learning approach to predict the success of crowdfunding fintech project," *Journal of Enterprise Information Management*, vol. 35, no. 6, pp. 1678–1696, 2022.

[5] Y. Xiao, C. Liu, W. Zheng, H. Wang, and C.-H. Hsu, "A feature interaction learning approach for crowdfunding project recommendation," *Applied Soft Computing*, vol. 112, p. 107777, 2021.

[6] V. Rakesh, W.-C. Lee, and C. K. Reddy, "Probabilistic group recommendation model for crowdfunding domains," in *WSDM*, 2016, pp. 257–266.

[7] T. Mitra and E. Gilbert, "The language that gets people to give: Phrases that predict success on Kickstarter," in *CSCW*, 2014, pp. 49–61.

[8] H. Yuan, R. Y. Lau, and W. Xu, "The determinants of crowdfunding success: A semantic text analytics approach," *Decision Support Systems*, vol. 91, pp. 67–76, 2016.

[9] S. Dey, B. Duff, K. Karahalios, and W.-T. Fu, "The art and science of persuasion: Not all crowdfunding campaign videos are the same," in *CSCW*, 2017, pp. 755–769.

[10] J. R. Hou, J. J. Zhang, and K. Zhang, "Can title images predict the emotions and the performance of crowdfunding projects?" in *HICSS*, 2019, pp. 4439–4448.

[11] F. Zhou, X. Xu, G. Trajcevski, and K. Zhang, "A survey of information cascade analysis: Models, predictions, and recent advances," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–36, 2021.

[12] C. Cheng, F. Tan, X. Hou, and Z. Wei, "Success prediction on crowdfunding with multimodal deep learning," in *IJCAI*, 2019, pp. 2158–2164.

[13] Z. Tang, Y. Yang, W. Li, D. Lian, and L. Duan, "Deep cross-attention network for crowdfunding success prediction," *IEEE Transactions on Multimedia*, vol. 25, pp. 1306–1319, 2023.

[14] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *ACM MM*, 2020, pp. 1122–1131.

[15] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *ACM MM*, 2022, pp. 1642–1651.

[16] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *CVPR*, 2023, pp. 6631–6640.

[17] Q. Gao, J. Hong, X. Xu, P. Kuang, F. Zhou, and G. Trajcevski, "Predicting human mobility via self-supervised disentanglement learning," *TKDE*, vol. 36, no. 5, pp. 2126–2141, 2024.

[18] Z. Liu, Z. Tang, X. Shi, A. Zhang, M. Li, A. Shrivastava, and A. G. Wilson, "Learning multimodal data augmentation in feature space," in *ICLR*, 2022.

[19] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, and Q. V. H. Nguyen, "Are graph augmentations necessary? Simple graph contrastive learning for recommendation," in *SIGIR*, 2022, pp. 1294–1303.

[20] X. Hao, Y. Zhu, S. Appalaraju, A. Zhang, W. Zhang, B. Li, and M. Li, "MixGen: A new multi-modal data augmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 379–389.

[21] D. J. Cumming, G. Leboeuf, and A. Schwienbacher, "Crowdfunding models: Keep-it-all vs. all-or-nothing," *Financial Management*, vol. 49, no. 2, pp. 331–360, 2020.

[22] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] J. C. Kaminski and C. Hopp, "Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals," *Small Business Economics*, vol. 55, pp. 627–649, 2020.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *ICLR*, 2014.

[26] X. Xu, J. Li, and F. Zhou, "MDCC: A multimodal dynamic dataset for donation-based crowdfunding campaigns," in *CIKM*, 2023, pp. 5417–5421.

[27] Q. Liu, G. Wang, H. Zhao, C. Liu, T. Xu, and E. Chen, "Enhancing campaign design in crowdfunding: A product supply optimization perspective." in *IJCAI*, 2017, pp. 695–702.

[28] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017, pp. 1103–1114.

[29] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *AAAI*, 2019, pp. 7216–7223.

[30] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8238–8247.