MDCC: A Multimodal Dynamic Dataset for Donation-based Crowdfunding Campaigns

Xovee Xu[*]
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Kash Institute of Electronics and Information Industry
Kashi, Xinjiang, China
xovee@std.uestc.edu.cn

Jiayang Li[*]
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
jy.li@std.uestc.edu.cn

Fan Zhou[†]
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Kash Institute of Electronics and Information Industry
Kashi, Xinjiang, China
fan.zhou@uestc.edu.cn

ABSTRACT

Crowdfunding platforms have become pivotal financial support avenues for diverse causes, yet the success rates are surprisingly low. Previous research has largely focused on reward-based crowdfunding, leaving donation-based platforms under-studied. In addition, the roles of multimodal data (e.g., textual descriptions and visual photos) and dynamic elements (e.g., sequences of donations, project updates, and comments) in influencing campaign success have been largely overlooked. This paper introduces MDCC, a Multimodal Dynamic dataset for donation-based Crowdfunding Campaigns, collected from 14,961 projects on GoFundMe, incorporates multimodal project information and captures project dynamics, thus providing a comprehensive tool for analyzing donation-based crowdfunding. The dataset is expected to inspire innovative methodologies and facilitate understanding of project success determinants. Our preliminary experiments demonstrate the significance of multimodal and dynamic crowdfunding data on predicting the success of donation-based projects.

CCS CONCEPTS

• **Human-centered computing** → *Web-based interaction*; • **Information systems** → **World Wide Web**; *Web mining*.

KEYWORDS

Donation-based crowdfunding, dataset, time series, multimodal

ACM Reference Format:
Xovee Xu, Jiayang Li, and Fan Zhou. 2023. MDCC: A Multimodal Dynamic Dataset for Donation-based Crowdfunding Campaigns. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615124

1 INTRODUCTION

Crowdfunding, as a new way of raising money from the Internet, enabling individuals and groups to achieve a variety of creative and social entrepreneurship activities. Leveraging the ubiquity and reach of the Internet, crowdfunding platforms such as Kickstarter and GoFundMe have emerged as popular avenues for soliciting financial support for various causes, from product development to medical expenses and community-oriented projects [5].

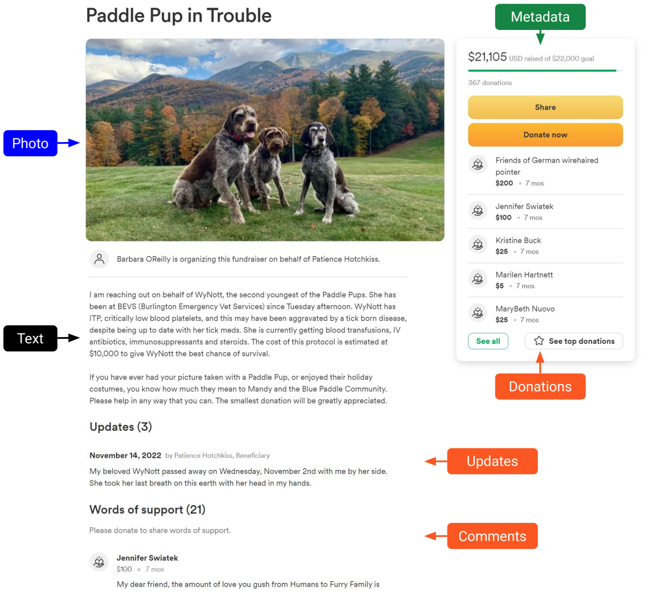
While these platforms differ in their operating models – for example, Kickstarter and Indiegogo offer rewards to backers, whereas GoFundMe and WaterDrop primarily drive donation-based crowdfunding – they share a common challenge: the success of a crowdfunding campaign is far from guaranteed. Studies reveal that more than half of all reward-based crowdfunding projects did not achieve their goals [11, 16], and the success rate for donation-based projects is even lower, with only about one in six campaigns reaching its target [9, 10, 12]. This underscores the crucial, yet challenging task of understanding and predicting crowdfunding success for both fundraisers and platforms.

Over the last decade, extensive research has sought to identify the determinants of crowdfunding success and propose many dedicated prediction methods [5]. However, many previous studies are limited to one or more drawbacks listed below: (*i*) A considerable research focus is directed towards reward-based crowdfunding, leaving donation-based crowdfunding projects – which are more difficult to succeed – underexplored. As suggested by [5], more research attention is needed for "Keep-it-All" model and for donation-based projects. (*ii*) The complexities of a crowdfunding project are manifold, encompassing diverse factors like textual descriptions and visual photos. The contributions of these different data modalities to crowdfunding success, and their potential synergistic effects, have often been overlooked in existing works that typically focus on single aspect of the project. (*iii*) The dynamics of crowdfunding projects reveal pivotal patterns and insights for fundraisers and platforms to adjust and optimize the projects [22]. However, research into the dynamic aspects of crowdfunding campaigns, such as the impact of sequential project adoptions, updates, or comments on the project success, remains scarce [17].

This paper aims to address these shortcomings by collecting and formulating a new Multimodal Dynamic dataset for donation-based Crowdfunding Campaigns – MDCC. We have tracked crowdfunding projects on gofundme.com within two months, and recorded

[*] Both authors contributed equally to this research
[†] Corresponding author

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615124



Paddle Pup in Trouble

Barbara OReilly is organizing this fundraiser on behalf of Patience Hotchkiss.

I am reaching out on behalf of WyNott, the second youngest of the Paddle Pups. She has been at BEVS (Burlington Emergency Vet Services) since Tuesday afternoon. WyNott has ITP, critically low blood platelets, and this may have been aggravated by a tick born disease, despite being up to date with her tick meds. She is currently getting blood transfusions, IV antibiotics, immunosuppressants and steroids. The cost of this protocol is estimated at $10,000 to give WyNott the best chance of survival.

If you have ever had your picture taken with a Paddle Pup, or enjoyed their holiday costumes, you know how much they mean to Mandy and the Blue Paddle Community. Please help in any way that you can. The smallest donation will be greatly appreciated.

Updates (3)

November 14, 2022 by Patience Hotchkiss, Beneficiary

My beloved WyNott passed away on Wednesday, November 2nd with me by her side. She took her last breath on this earth with her head in my hands.

Words of support (21)

Please donate to share words of support.

Jennifer Swiatek
$100 • 7 mos

My dear friend, the amount of love you gush from Humans to Furry Family is

$21,105 USD raised of $22,000 goal

367 donations

Friends of German wirehaired pointer
$200 • 7 mos

Jennifer Swiatek
$100 • 7 mos

Kristine Buck
$25 • 7 mos

Marilen Hartnett
$5 • 7 mos

MaryBeth Nuovo
$25 • 7 mos

Figure 1: An example of GoFundMe crowdfunding campaign, which contains a variety of multimodal and dynamic data, including project description, campaign photos, project metadata, and sequences of donations, updates, and comments.

14,961 projects with 21,029 photos, 1.2M donations, 17K updates, and 55K comments. All projects raised 114M dollars in total. An example of *GoFundMe* project is illustrated in Figure 1. Our proposed dataset distinguishes itself in several key ways:

- **Donation-based crowdfunding**. Our dataset, MDCC, focuses specifically on donation-based crowdfunding, a relatively under-researched area that highlights prosocial and giving behaviors. This dataset aims to spur further research into this domain, enriching our understanding of the determinants of donation-based crowdfunding success, promoting efforts to design novel methodologies and discover new insights that support various medical, educational, and community-oriented crowdfunding campaigns.
- **Multimodal crowdfunding data**. The dataset we proposed consists of multimodal crowdfunding data, including *textual descriptions*, *visual photos*, *time series*, and *project metadata*. Different data modalities have different roles for story-telling, e.g., the language used by fundraisers was showed to have significant predictive power for crowdfunding success [15], while visual signals from campaign photos provide vital references for explaining fundraising causes [2, 8]. Our dataset can facilitate new research studying the interactions and complementations between multimodal crowdfunding data.
- **Crowdfunding dynamics**: The proposed dataset captures the evolution of crowdfunding projects over time, providing key insights into crowdfunding trends and project successes. By integrating dynamic elements like donation times and amounts, donor comments, and fundraiser updates, the dataset enables a deeper understanding of project attractiveness, community feedback, and project progresses, thereby enhancing predictive models and recommendation systems.

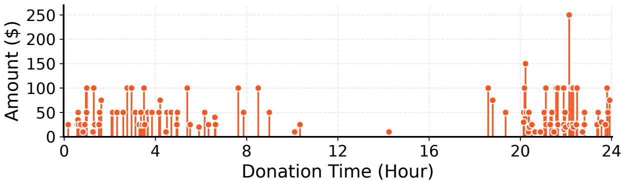


Figure 2: A sequence of donations received by an example crowdfunding project in one day.

Table 1: Data Structure

Feature	Description
`launch_time`	The project launch date and time, from October 16 to November 20, 2022.
`campaign_id`	Unique identifier of the project.
`category`	Each project belongs to one of five categories.
`fundraising_goal`	The amount of money that the fundraiser wish to raise.
`country, city`	The fundraiser’s location, including country and city information.
`description`	The fundraising story provided by the fundraiser.
`cover_photo`	The cover photo of the project.
`main_body_photo`	Main body photos, if any.
`donation`	The sequence of donations received by the project, in the format of pairs of donation time and donation amount.
`comment`	The comments posted by the donors. Only users who have donated can leave a comment.
`update`	The fundraiser can provide updates to inform the visitors the latest information about the project.

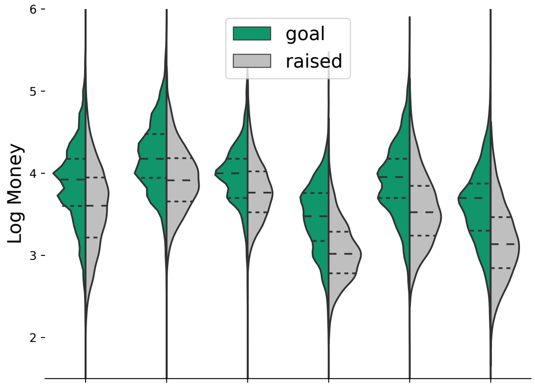
2 DATASET CREATION AND ANALYSIS

This section describes the details of data acquisition and statistics.

Data Acquisition. The construction of the MDCC dataset requires us to track the lifetime of *GoFundMe* crowdfunding projects from the project launch to the end. Since there is no time limit for the project to reach its goal, we track each project’s donations for at least a month. We select the five most popular crowdfunding categories (Medical, Memorial, Emergency, Financial Emergency, and Animals). For each category, we monitor the corresponding webpage (e.g., www.gofundme.com/start/medical-fundraising) to collect the newly launched crowdfunding projects and start to record their donations from donors, updates from the fundraisers, and comments from the donors. For each crowdfunding project, we crawled its textual project description, cover photo and main body photos, and project metadata. Figure 2 depicts a sequence of donations received by an example project. The data structure details are summarized in Table 1.

Data Preprocessing. We have three kinds of textual data extracted from project description, update and comment. We first use `BeautifulSoup` to parse the corresponding HTML sections. Tags and unwanted characters are removed to obtain the clean texts. For visual modality, each project has at least one cover photo. Fundraisers can optionally add main body photos in the project description. For dynamic crowdfunding data including project donations, updates and comments, they are arranged as irregularly

Table 2: Data statistics for different categories



Stat	Total	Medical	Memor.	Animal	Emerg.	Financ.
# projects	14,961	6,580	4,632	2,879	2,526	1,344
avg. goal	81,965	307,303	14,102	4,698	14,463	8,004
avg. raised	7,645	13,109	9,270	1,654	6,403	2,653
avg. photos	1.406	1.408	1.231	1.738	1.469	1.170
avg. updates	1.139	1.799	0.586	1.385	1.076	0.875
avg. comments	3.689	5.666	4.756	1.416	2.849	1.196
success rate	22.09%	22.03%	28.64%	15.10%	20.98%	16.74%

time series. For example, the donation sequence is composed by (`donation_time`, `donation_amount`) pairs, the sequence of updates is composed by (`update_time`, `content`) pairs, the sequence of comments is composed by (`donation_time`, `comment_time`, `comment`) pairs. We note that fundraisers may occasionally post updates directly in the project description. In order to preserve the original project description as much as possible, we search key words like "updat" and manually remove the updated parts.

Data Statistics. In total, the MDCC dataset consists of 14,961 crowdfunding projects, 21,029 photos, 1,219,667 donations, 17,083 updates, and 55,194 comments. During our observation, all projects have a total goal of 1.2 billion dollars, raised 114 million dollars, and 22.1% of the projects are successfully reached their goals. Detailed statistics for projects of different categories are shown in Table 2. Medical projects typically demonstrate significantly larger fundraising goals compared to other project types, they also have the highest numbers of updates and comments. Memorial projects are generally more successful, while Animal projects and Financial Emergency projects underperformed than others. Same trends can also be observed in the left of Figure 3, where Memorial projects have the fastest fundraising speed. Furthermore, there is a variation in the distribution of donor numbers across categories. For instance, as demonstrated in the right of Figure 3, Medical projects attract the most donors. This is likely attributed to their higher funding goals and popularity within the *GoFundMe* platform.

Data Availability. Our original data is available at https://github.com/Jiayang-L1/mdcc. We also provide all the necessary scripts and baseline codes to reproduce the results reported in the paper.

Ethical Consideration and Data Impact. The dataset is collected from public crowdfunding project and should be used only for academic purposes. To protect privacy, we have removed personal information in project metadata, including names of the fundraisers and donors. We also blurred human faces in all photos by using [14]. The dataset can be used to design prediction models and improve user experience. For example, improving the policy and recommendation algorithm of the platform, and helping fundraisers to optimize their projects.

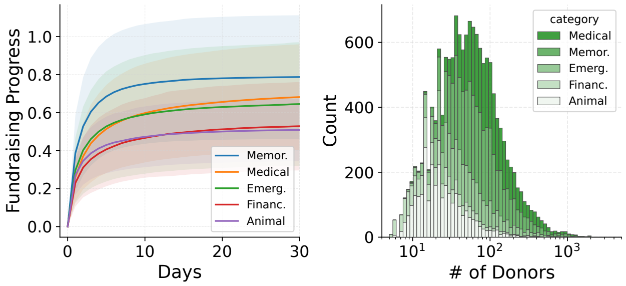


Figure 3: Differences between five crowdfunding categories. Left: Fundraising progress (raised divided by the goal). Right: Distribution of number of donors.

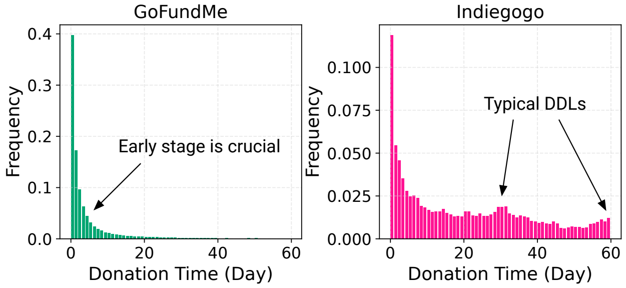


Figure 4: A donation time comparison between *GoFundMe* projects and *Indiegogo* projects [13].

A Comparison with Reward-based Crowdfunding. The *GoFundMe* crowdfunding projects in our dataset do not have time limits for reaching their goals, while reward-based crowdfunding platforms such as *Indiegogo* and *Kickstarter* often require the fundraisers to set a fixed time. We compare the donation times of our dataset with Indiegogo [13] in Figure 4. We can observe that *GoFundMe* projects' donations are mostly happened in the early stage of the campaign (following a power-law distribution). The donation distribution of *Indiegogo* projects is relatively flat and has several rises at typical end times, such as 30 days or 60 days. This phenomenon highlights the importance of *GoFundMe* projects paying special attention to initial project quality and promotion.

3 EXPERIMENT

We evaluate the MDCC dataset using various learning methods such as pretrained language and vision models, point process-based models, and recurrent neural networks. We focus on two downstream tasks: crowdfunding success and outcome prediction. Achieving satisfactory performance on this dataset is challenging due to its multimodal and dynamic nature; effective learning strategies are required to extract signals from different data modalities and fuse them for predictions.

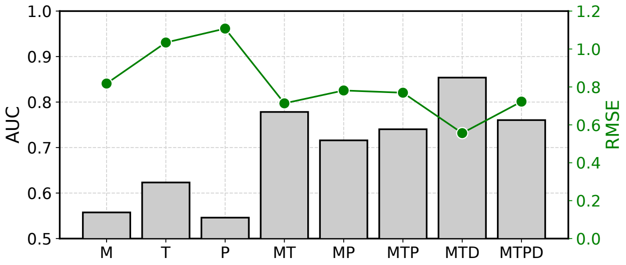


Figure 5: Experimental results for the effects of multimodal crowdfunding data. Abbreviations: Metadata (M), Text (T), Photo (P), Donations (D).

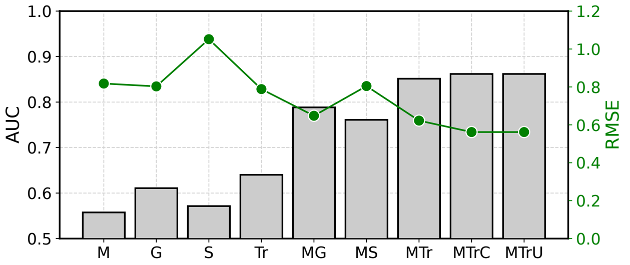


Figure 6: Experimental results for the effects of dynamic crowdfunding data. Abbreviations: Metadata (M), Comment (C), Update (U). For donations, they are encoded by GRU (G), Self-exciting point process (S), or Transformer (Tr).

Experimental Settings. The crowdfunding success prediction task is an imbalanced binary classification, predicting whether the project is succeeded or not within 30 days. We use area under the ROC curve (AUC) as the evaluation metric. The crowdfunding outcome prediction task aims to predict the money raised 30 days after the project launch, in log scale. We use root mean squared error (RMSE) as the evaluation metric. For each of the prediction task, two branches of learning methods are adopted, i.e., multimodal-based methods and dynamic-based methods.

Multimodal. For multimodal-based methods, we use the data modalities available at the beginning of project launch, including project metadata, project description, and campaign cover photo. For metadata, `category` and `country` are categorical features encoded as one-hot embeddings, other features (e.g., `num_words`) are treated as numerical values. For textual modality, we extract features via pretrained BERT model [6], resulting in a 768-dimension embedding. For visual modality, we use ResNet-152 [7] to extract photo feature as a 2048-dimension embedding. Following existing multimodal works [1, 3, 18, 20, 24], we concatenate all features and make final predictions via fully-connected layers.

Dynamic. For the branch of crowdfunding dynamics, we treat the sequences of donations, updates and comments as discrete time series. We use the first 24 hours of the time series as the model input. Rather than simply fusing them with multimodal features, we adopt several dynamic data learning methods to capture long-range temporal dependencies of the time series. Specifically, we use the following three methods to encode crowdfunding time series: (*i*) GRU [4], a variant of recurrent neural networks. The hidden size of the GRU is set to 256; (*ii*) Self-exciting point process [23], which is a statistical model that captures the rise and fall patterns of the time series [21]. We choose *Infectiousness* as an intermediate variable to extract temporal features. After fitting the donation time series, the features of point process are represented as a 1441-dimension vector ($24 * 60 + 1$ minutes); (*iii*) Transformer [19] is a powerful learning model that achieved great success in various language and vision tasks. Other settings can be found in the running scripts.

Results. The outcomes of the two distinct experimental branches, namely, crowdfunding multimodal learning and dynamic learning, are presented in Figure 5 and Figure 6, respectively. Evaluating the results from the multimodal branch, it becomes evident that the `Metadata` and `Photo` modalities serve as comparatively weak predictors. Combining the `Metadata` and `Text` modalities results in a substantial improvement in performance. Counterintuitively, the integration of the `Photo` modality did not enhance the prediction performance; instead, it led to a degradation in performance. This phenomenon underscores the complexities and challenges associated with leveraging visual patterns to augment crowdfunding prediction performance. In total, the `Metadata+Text+Donations` combination yielded the highest performance, achieving an AUC of 0.854 and an RMSE of 0.556.

Upon evaluating the findings derived from the dynamic branch, it becomes apparent that the `Transformer` model demonstrates superior efficacy in encoding the donation time series. The self-exciting point process underperforms when compared with learning-based neural networks. Notably, integrating the `Metadata` modality with donation sequences results in a significant enhancement in the model's predictive ability. This observation implies that neither Metadata nor donation sequences, when used in isolation, can deliver robust predictive performance. Additionally, `Update` and `Comment` data both slightly improved the performance.

Overall, our findings highlight the importance of the project description (`Text`) and donation sequence in predicting the success of donation-based crowdfunding projects. In terms of encoding the donation sequence, the `Transformer` model proves more effective in extracting time series patterns.

4 CONCLUSION

We introduce a multimodal dynamic dataset for donation-based crowdfunding campaigns, encompassing textual, visual, and dynamic modalities. This dataset facilitates enhanced predictive modeling of crowdfunding success. Future directions encompass: (*i*) developing methodologies to derive semantics from the dataset for improved predictions, and (*ii*) analyzing the evolution of donation-based crowdfunding to offer informed decision-making insights.

ACKNOWLEDGMENTS

The authors would also like to thank the anonymous referees and track chairs for their valuable comments and helpful suggestions. This work was supported in part by the National Science Foundation under Grant SWIFT No. 2030249; in part by the National Natural Science Foundation of China under Grant No. 62072077 and Grant No. 62176043; in part by the Natural Science Foundation of Sichuan Province, China under Grant No. 2022NSFSC0505.

REFERENCES

[1] Liqian Bao, Zongxi Liu, and Huimin Zhao. 2022. Reward-based Crowdfunding Success Prediction with Multimodal Data. *AMCIS* (2022).

[2] Simon J Blanchard, Theodore J Noseworthy, Ethan Pancer, and Maxwell Poole. 2022. Extracting Image Characteristics to Predict Crowdfunding Success. *arXiv:2203.14806* (2022).

[3] Chaoran Cheng, Fei Tan, Xiurui Hou, and Zhi Wei. 2019. Success Prediction on Crowdfunding with Multimodal Deep Learning.. In *IJCAI*. 2158–2164.

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555* (2014).

[5] Lingfei Deng, Qiang Ye, DaPeng Xu, Wenjun Sun, and Guangxin Jiang. 2022. A literature review and integrated framework for the determinants of crowdfunding success. *Financial Innovation* 8, 1 (2022), 41.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* (2018).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[8] Jian-Ren Hou, Jie Zhang, and Kunpeng Zhang. 2019. Can title images predict the emotions and the performance of crowdfunding projects?. In *HICSS*.

[9] Junjie Huang, Huawei Shen, Qi Cao, Li Cai, and Xueqi Cheng. 2021. How Medical Crowdfunding Helps People? A Large-scale Case Study on the Waterdrop Fundraising. In *ICWSM*, Vol. 15. 220–229.

[10] Mark Igra, Nora Kenworthy, Cadence Luchsinger, and Jin-Kyu Jung. 2021. Crowdfunding as a response to COVID-19: Increasing inequities at a time of crisis. *Social Science & Medicine* 282 (2021), 114105.

[11] Kickstarter. 2023. Kickstarter Stats. https://www.kickstarter.com/help/stats Accessed: 2023-06-15.

[12] Bruce Y. Lee. 2022. Most GoFundMe Campaigns For Medical Bills Fail, Less Than 12% Reach Goals. *Forbes* (5 Feb 2022). https://www.forbes.com/sites/brucelee/2022/02/05/most-gofundme-campaigns-for-medical-bills-fail-less-than-12-reach-goals

[13] Qi Liu, Guifeng Wang, Hongke Zhao, Chuanren Liu, Tong Xu, and Enhong Chen. 2017. Enhancing Campaign Design in Crowdfunding: A Product Supply Optimization Perspective.. In *IJCAI*. 695–702.

[14] Asmaa Mirkhan. 2020. BlurryFaces: A tool to blur faces or other regions in images and videos. https://github.com/asmaamirkhan/BlurryFaces. GitHub repository.

[15] Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: Phrases that predict success on Kickstarter. In *CSCW*. 49–61.

[16] Ethan Mollick. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing* 29, 1 (2014), 1–16.

[17] Xiaoying Ren, Linli Xu, Tianxiang Zhao, Chen Zhu, Junliang Guo, and Enhong Chen. 2018. Tracking and forecasting dynamics in crowdfunding: A basis-synthesis approach. In *ICDM*. 1212–1217.

[18] Zhe Tang, Yi Yang, Wen Li, Defu Lian, and Lixin Duan. 2022. Deep cross-attention network for crowdfunding success prediction. *IEEE Transactions on Multimedia* 25 (2022), 1306–1319.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS* 30 (2017).

[20] Xovee Xu, Ting Zhong, Ce Li, Goce Trajcevski, and Fan Zhou. 2022. Heterogeneous dynamical academic network for learning scientific impact propagation. *Knowledge-Based Systems* 238 (2022), 107839.

[21] Liu Yu, Xovee Xu, Goce Trajcevski, and Fan Zhou. 2022. Transformer-enhanced Hawkes process with decoupling training for information cascade prediction. *Knowledge-Based Systems* 255 (2022), 109740.

[22] Hongke Zhao, Hefu Zhang, Yong Ge, Qi Liu, Enhong Chen, Huayu Li, and Le Wu. 2017. Tracking the dynamics in crowdfunding. In *SIGKDD*. 625–634.

[23] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *SIGKDD*. 1513–1522.

[24] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *Comput. Surveys* 54, 2 (2021), 1–36.