# Using multiple forms of cues and signals to predict crowdfunding success

Zijian Zhang

Department of Information Systems

City University of Hong Kong

***Abstract*— In this course report, I summarize two types of cues that may have an impact on crowdfunding outcomes: explicit cues and implicit cues. Explicit cues emerged as crowdfunding projects started, such as real-time updates of social media comments, support numbers; Implicit cues are set before crowdfunding begins which include various project description in the form of text, image, audio, and video. In order to comprehensively measure the predictive effect of different cues on crowdfunding outcomes, I initiated two prediction tasks, respectively. In that first task, I use a logistic regression model to measure the contribution of each feature signal pair. It was found that whether the project included a video, the number of backers, and the number of projects supported were the top three high impact signals. In the second task, I used the mainstream deep learning frameworks CNN, RNN, and BERT as language models to compare the prediction performance of text, speech, and video signals under the three models. The results show that the prediction accuracy of text-based signals is generally higher than that of the other two signals, with TextCNN performing best at 76.1% (10.5% improvement over BiRNN and 2.6% improvement over BERT). The experiments for the two tasks were performed on data Kickstarter-3.8 and Kickstarter-1.1, respectively. In addition, in order to explore the interaction among different signals, I also designed a multi-modal model, which is detailed in Appendix D.**

***Keywords—crowdfunding success, language model, logistic regression***

## I. INTRODUCTION

Crowdfunding is growing rapidly. Accurate prediction of crowdfunding outcomes requires high-quality cues and accurate prediction models. There are currently two types of cues available on crowdfunding platforms to predict outcomes: **explicit and implicit cues**. Explicit cues are mostly structured data, which are high-semantic and dynamically updated, such as the use of project entrepreneur information, project social comment information. Each feature(signal) has clear business and social information that can help people make quick investment decisions about a project with some prior knowledge and experience. For example, when a project has a large number of supporters, people will favor it more. However, some explicit information can only be obtained after the project is officially launched. Moreover, the feature information extracted in different stages is different, which leads to lower robustness of the result prediction and is more suitable for post-interpretation. On the contrary, most of the implicit clues are unstructured, which are low-semantic and can be set before the project launched but hard to modify, such as various project description in the form of text, image, speech and video. For implicit clues, readers will spend some time to get the message the entrepreneur wants to convey. Therefore, these information usually are converted into high-semantic information by the language model to complete the prediction.

In this article, I aim to compare the predictive performance of different signals in explicit and implicit cues for crowdfunding outcomes. Specifically, my experimental strategy is that I sample explicit and implicit cues on the mainstream crowdfunding platform kickstarter. In the first task, I use structured data (metadata) to explore the impact of explicit signals including entrepreneur information and project information on crowdfunding with linear regression method. In the second task, I first transcribed video data into textual video descriptions and textual audio descriptions, combined with text descriptions, and placed them separately into deep networks to explore the predictive power of each signal.

The remainder of this paper proceeds as follows: Section II summarizes the related work of using explicit and implicit cues to predict crowdfunding results. Section III describes my approach. Section IV presents experiments and results. Section V draws conclusions and Section VI is future work.

## II. RELATED WORK

A typical process for a startup using reward-based crowdfunding to raise seed fund is: (1) Entrepreneurs launch a fundraising campaign on the site, which may be about a concept product or service. (2) During the fundraising period, which usually ranges from one to three months, typical investors judge whether to invest in the project by reviewing the text and video descriptions of the project or interacting with other investors in the comments section. The amount of funding is determined by the investors themselves. (3) Once the project is successfully financed, investors will receive non-monetary rewards related to the project provided by entrepreneurs based on the amount of investment. In All-Or-Nothing mode, a minimum fund goal will be set by the entrepreneur's own discretion. Investors will only receive rewards if the goal is met by the deadline. Otherwise

**Table I** Summary of studies for prediction in crowdfunding market. While some articles use only explicit cues (numerical features), others use only implicit cues such as images and text, and still others use cues combination.

| Scholar | Data | Sample | Method | Accuracy | Features Form |
|---|---|---|---|---|---|
| Greenberg et al. (2013) | Kickstarter | 13,000 | Random Forest | 67.53% | numerical |
| Ethan Mollick(2014) | Kickstarter | 48,500 | Linear Regression | - | numerical |
| Etter et al. (2013) | Kickstarter | 16,042 | SVM | 76.22% | numerical |
| Zecong Ma (2021) | Kickstarter | 652 | Logistic Regression | - | Image |
| Hui Yuan et al.(2016) | Dreamore&Zhongchou | 23,113 | LDA | - | textual |
| Mitra and Gilbert (2014) | Kickstarter | 45,815 | Logistic Regression | 58.56% | textual |
| Chaoran Cheng (2019) | Kickstarter | 18,511 | Neural Network | 83.26% | Image&textual |
| Simon J. Blanchard (2023) | Indiegogo | 10,487 | Bayesian additive tree | 88.80% | Image&textual |
| Desai et al. (2015) | Kickstarter | 26,000 | Logistic Regression | 74.02% | numerical&textual |
| Jermain Kaminski (2020) | Kickstarter | 20,188 | Neural Network | 72.65% | numerical&textual |

the investor will receive the money back but no return. In Keep-It-All mode,where the entrepreneur keeps the entire amount raised regardless of achieving the goal. The process is showed in **Appendix A**.

*A. Explicit clues predict crowdfunding results*

Scholars have carried many researches to examine the determinants of funding success in various crowdfunding platforms. According to the existing literature, the main factors can be divided into three categories. I summarize these studies in Table I.

**Campaign Characteristic.** Campaign target, campaign category, campaign duration, as well as campaign backers. In general, the target funding of a project and its funding success shows a negative correlation. High capital demand means attracting more investors to participate in limited time, which is obviously hard in the all-or-nothing model [12,13,14]. The duration of a project may be also inversely related to fundraising success. Because the long term reflects the entrepreneur's lack of confidence in the project, and the investor will also browse other ideas. Over time, the possible result is to finally abandon the initial project. [12,13,15]. Besides, financing success probability of different types is not the same. According to the statistics on kickstarter website [16], by December 31,2022, the success rate of raising funds on the platform was 64.83% for comics, 61.48% for dance and 59.97% for drama, followed by 22.45% for technology, 26.01% for food and 26.91% for handicrafts. On the one hand, this difference is due to statistical bias caused by the number of project released. For example, categories with more post are more likely to represent the true results, while categories with fewer post are more affected by individual differences of projects (Dance:4474, Science and Technology:50666); On the other hand, that is some campaign categories enjoy more popularity and attract more backers and funds. Products of the same kind have similar consumer preferences and market structures, as a result, projects with more supporters will receive more financial support [17]. For example, it's hard for some investors to enter an unknown investment field considering the risks as they have relatively little professional knowledge; Comics and other literary products do not require investors to have professional knowledge background, so investors are relatively more.

**Campaign Understandability.** It refers to visible signals that describe the project itself and distinguish it from other projects such as project functions or service features as well as the form of reward. Signal theory was put forward by Spencer in 1973, the winner of Nobel Prize in Economics in 2001. It describes that under the early market transaction with asymmetric information, the seller can enhance the trust and confidence of the buyer by transmitting high-quality signals, thus realizing the potential transaction profit. In the crowdfunding market, investors can not visit enterprises and talk with employees or entrepreneurs in person to ensure entrepreneurial quality, so the key to making a project stand out and be favored by investors is: Entrepreneurs as sellers with information advantage send a high-quality signal to buyers. For example, the level of concreteness and precision of the textual presentation of the project title and project overview, whether there is a dynamic video show, what are the rewards for return on investment, etc. [13,18,19,20,21].

**Entrepreneur Characteristic.** In addition to the project characteristics, the characteristics of entrepreneurs are also the focus of investors. The credit status, entrepreneurial experience, social network and other information of entrepreneurs are all important references for investors to make investment decisions [22,23,24]. Good credit status means that entrepreneurs will deliver returns on time, rich entrepreneurial experience makes investors more confident that entrepreneurs have the ability to grasp entrepreneurial opportunities, and huge social networks will bring entrepreneurs the possibility of transforming social capital into venture capital [25,26].Implicit cues refer to unstructured information, such as pictures, video frames, sounds, and even words with low-semantic features. I summarize these studies in Table I.

### B. Implicit cues predict crowdfunding results

In the section of text information, Hui Yuan developed the Domain-Constraint Latent Dirichlet Allocation (DC-LDA) topic model for effective extraction of topical features from texts project descriptions on a China crowdfunding platform datasets and achieved good performance[1]. Mitra and Gilbert show phrases have exhibit general persuasion principles by analyzing 9M phrases and 59 other variables commonly present on crowdfunding sites phrases[2]. In another related work[3], Desai include metadata as well as linguistic features for binary classification show that linguistic features like reciprocity, social relationship and emotional appeal play a clear positive role in classification.

In the section on using images and video information, Zecong Ma found that the "workspace" cluster positively linked to crowdfunding projects 'success, while the "event" cluster was negatively related by analyzing 11,264 video frames with the Louvain method of community detection[4]. One of the papers that came closest to my work was[5] that using text, visual, and audio signal analysis followed by feature fusion to achieve better classification results. But this work, by contrast, takes a much better feature extraction technique and classifier. Even though I didn't use feature fusion, the accuracy of predictions using only text cues has improved by 3%, which confirms the potential of deep neural networks in language models.

## III. Approach

I used a simple logistic regression in the first task, which was mainly to visualize the degree of influence of individual signals (features) in the explicit cue. (**Not explained in the section III Approach, only reported results in Experiments**.) The second task is essentially a text classification task, and the analysis process consists feature extraction and classification architecture. For feature extraction, text descriptions can be classified directly after being represented by word embedding. The audio and video features are obtained by transcribing the video into a textual signal using the Google cloud video Intelligence API[6]. Figure 1 shows a processed sample. The frame feature and audio feature here are represented by text instead of RGB image and sound frequency. This is a high-semantic feature extraction. After word embedding, I used the popular deep network architectures of the past decade, CNN, RNN, and the transformer encoder-based BERT model, as the classification architectures. Figure 2 illustrates this idea.

### A. Feature Extracting

**Glove** modified the loss function on the skip-gram model, which carries rich semantic information and has been widely used in feature extraction tasks. As I have decomposed the video signal into audio features and video features represented by text, I no longer need to extract image features through special visual networks such as ResNet. For CNN and RNN models, I use pre-trained Glove[7] for word embedding representation to reduce over-fitting.

**BERT** is a language representation model developed by Google based on the Transformer encoder architecture. The embedding representation of BERT is the sum of the three embedding layers. When the input is a single text sequence, the word embedding layer consists of special tokens '<cls>', text words and separator tokens '<sep>'. The segment embedding layer consists of single parameter Ea. The position embedding layer adopts learnable position coding. Appendix B shows the BERT input embedding architecture.

### B. Model Architecture

**BiRNN** refers to the bi-directional recurrent neural network. It can handle sequence information well as it stores past information and current input by introducing state variables. I used a BiRNN to encode the textual signal after transcription[8]. Specifically, the 2-layer LSTM as the infrastructure of RNN and pairs of hidden states at the initial and final time steps are concatenated at the last layer as a sequential representation of the text. The text representation is then transformed into predictions through a fully connected layer. Figure 2 shows BiRNN architecture.

**TextCNN** is the convolutional neural network for text. Convolutional neural network were originally designed for computer vision, but it is also widely used for natural language processing [9]. For a sequence represented by a d-dimensional vector of length n, let the width of the input tensor and the number of channels be n, d. Define three one-dimensional convolution kernels with widths k1, k2, k3, which have output channel numbers c1, c2, c3, respectively. After the first convolution operation, c1 output channels of width n-k1+1, c2 output channels of width n-k2+1, and c2 output channels of width n-k3+1 are generated, respectively. A maximum time convergence layer is performed on all lanes, concatenating all scalar convergence outputs into vectors, and finally converting the converted vectors into prediction results using a fully connected layer. Figure 2 shows TextCNN architecture.

*(A)Text input*

"Hello Kickstarter Family. We are excited to bring you our story of Do Dah Dolls™We have setup this Kickstarter because we are so close! We have financed 90% of this project ourselves and only need the last 10% from you.(...)"

*(B)Speech input*

["We will match books from authors to Industry professionals, based on the obvious elements as well as some of the unspoken rules that get most authors rejected. It's not enough anymore(...) "] [0-26.2s] [Confidence: 0.78]

*(C)Video input*

['black and white,0-2.2s','style,0-10.2s','clip art,3.2s-7.8s','graphics,4.7s-20.2s','display device5.2s-22.5s', 'technology,6.2s-33.8s','food,12.3s-27.9s', 'car,18.3s-29.8s', (...)]

***Figure 1***. a processed sample

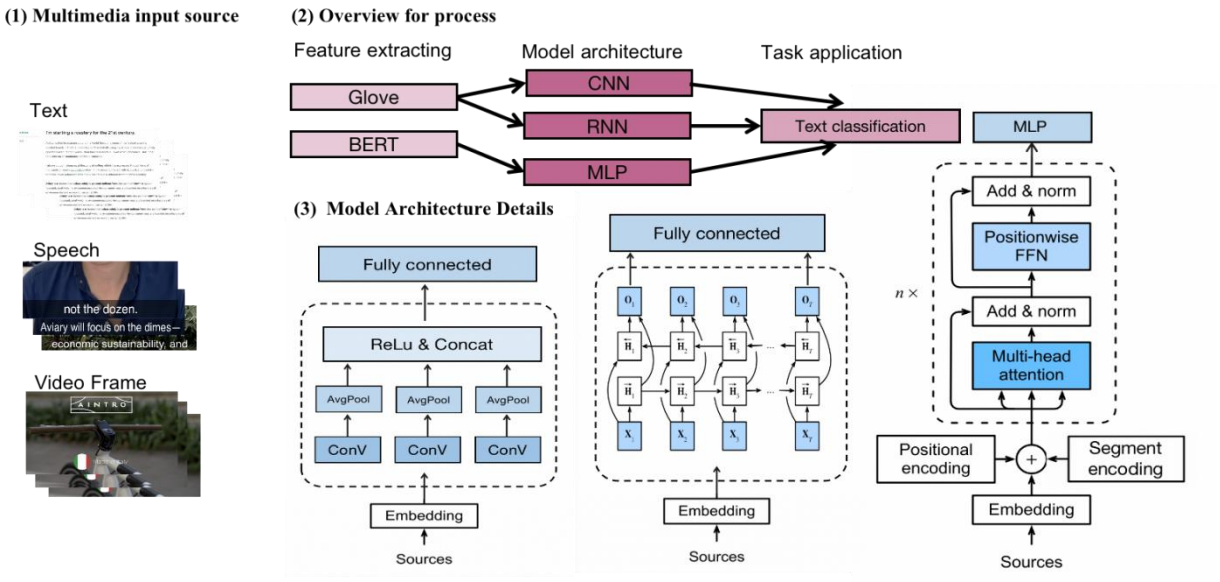



***Figure 2.*** Summary of my approach. The input signals (modalities) are text, voice, and video. Feature extraction is carried out for different model architectures(purple), and the specific model architectures are also shown(blue).

**BERT+MLP** takes the word embedding of BERT and trains them in several transformer encoder blocks. The '<cls> 'BERT representation of the special classification token encodes information over the entire text sequence. It will be fed into a multi layer perceptron with a hidden layer to output that prediction result. The specific design of each Transformer encoder block is shown in Figure 2. The first layer is multi-head self-attention convergence, and the second layer is a position-based feed forward network. The residual connection is used between each layer and the layer normalization is carried out. For each position of the input sequence, the encoder will output a d-dimensional vector representation[10].

## IV. Experiments

In this section, I report the results of the model based on benchmark datasets.

### *A. Task 1: explicit cues*

**Datasets.** Kickstarter-3.8 is a collection of 38,190 project data that I manually collected from Kickstarter. It contains project status, project creator information, and project description information including category, location, etc. but does not contain signals such as video,images and text. The entire datasets has 18 categories distributed across 52 states in the United States. A fuller description of this data set can be found in Appendix C1. The variables used in Task1 are described in Table II.

**Implementation Detail.** In this section, I used linear regression with a sigmoid function to explore what explicit signals can have an impact on crowdfunding success. In the heterogeneity test, I examined the effects of funding goals and project duration. The duration is divided into three groups (0-30)(30,60)(60,-) and the funding goal divided into three groups (0-10,000)(10,000-20,000)(20,000,-). The regression model is from the Statsmodels library[11].

**Performance.** My first analysis was to examine which signals had an impact on success. As you can see, including the video in the project description is 20% more successful than not including the video. This makes sense because in the vast majority of cases, a video presentation can show the features and functionality of a product or service more clearly, which will help investors understand the project. In addition, the number of projects initiated and supported by project entrepreneurs also has a positive impact on financing success. This may be because experienced entrepreneurs are more likely to win the trust of investors. One interesting observation is that project duration is negatively correlated with financing success. I assume as it may signal the entrepreneur's lack of confidence in the project and cause investors to consider other opportunities. Eventually, this could lead to the project being abandoned altogether.

As for project category and origination locations, they are based on Alabama and arts projects, respectively. In addition, I also conducted heterogeneity tests on project duration and financing goal. The specific conclusions are in Appendix C.3. In general, the results of task 1 are consistent with many previous literature studies[13,18,19]. Table II and Figure 3 illustrate the coefficients.

**Table II.** Summary of variables and coefficients. The bold coefficients are significant at the 5% level.

| Variable | description | mean | Coef | Std Error | T-value |
|---|---|---|---|---|---|
| Success | the results of project | 0.727 | | | |
| Goal | project target amount | 10013 | **-0.0001** | 0.0001 | -12.258 |
| Duration | project time from start to finish | 32.704 | **-0.0023** | 0.0001 | -7.49 |
| Backers | No.people supporting the project | 618.673 | **0.0001** | 0.0001 | 1.644 |
| CommentNum | No.comments about the project | 317.843 | -0.001 | 0.0002 | -0.666 |
| PictureNum | No.pictures about in the overview | 11.269 | **0.0004** | 0.0001 | 3.009 |
| Video | whether a video in the promotion | 0.623 | **0.2023** | 0.0008 | 25.012 |
| State | project located | - | - | - | - |
| Category | project type defined of kickstarter | - | - | - | - |
| BackerNum | No.projects creators supported | 57.446 | **0.001** | -0.001 | 2.01 |
| CreatNum | No.projects creators created | 3.759 | **0.0015** | 0 | 4.468 |
| Intercept | | | **0.9033** | 0.014 | 66.589 |

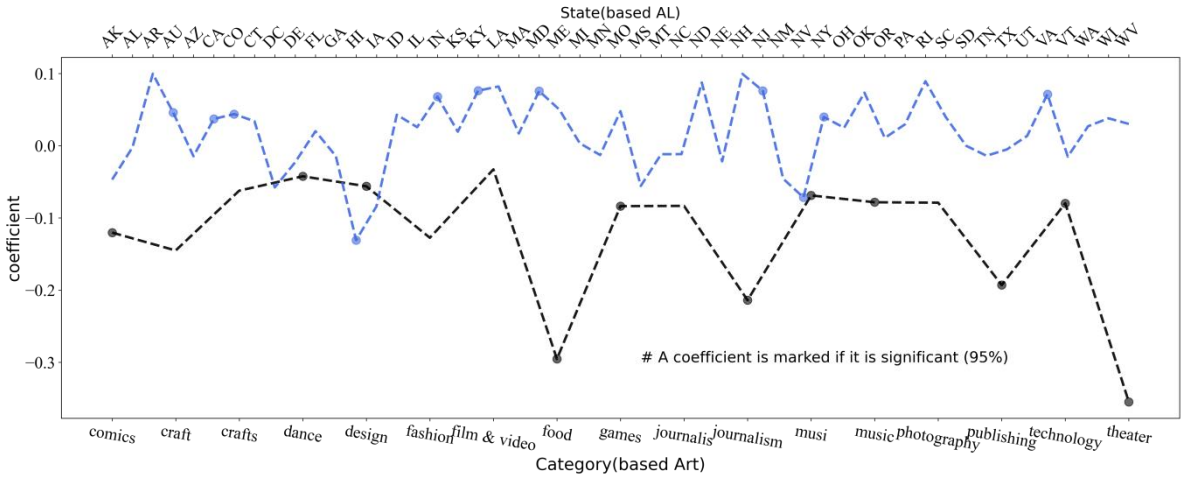



***Figure 3.*** Summary of Coefficients of State and Category. The State variable is based on Alabama, and category is based on Art. Food and Theater are the two categories of projects with a lower success rate.

**Table III** Descriptive statistics of preprocessed-sample

| Source | Mean token | Vocab | Seq length |
|---|---|---|---|
| text | 734.52 | 56246 | **1024** |
| speech | 205.46 | 6002 | 512 |
| video | 100.34 | 1326 | 256 |

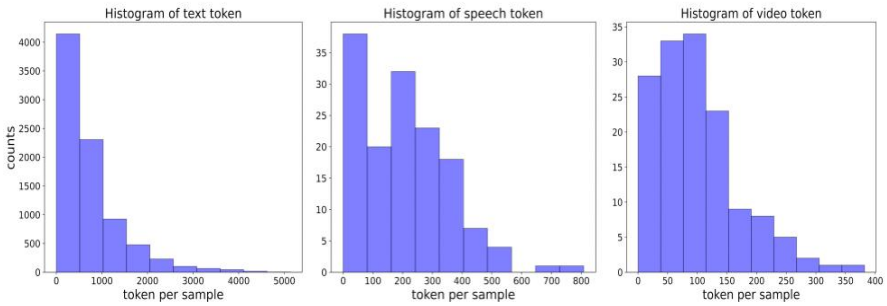



***Figure 4.***Hisograms of three signals token

**Table IV** Data process

*(1) Tokenize('word')*

"['Hello','Kickstarter','Family','We','are','excited','to','bring','you','our','story','Do','Dah','Dolls™We','have','setup','this',Kickstarter',(...)"

*(2)Build a vocab(filter<5,stopwords)*

[3071, 70, 42877, 38, 17, 499, 4, 165, 11, 21, 167, 5, 1262, 0, 0, 24, 3460, 22, 70, 178, 18, 17, 49,(...)]

*(3)padding&truncate()*

[[ 3071, 70, 42877, 38, 17, 499, 4, 165, 11, 21, 167, 5, 1262, 0, 0, 24, 3460, 22, 70, 178, 18, 17, 49, ..., 1, 1, 1]]

*B. Task2:Implicit cues*

**Datasets** Kickstarter-1.1 is a collection of 11,064 project data that I manually collected from Kickstarter. It contains video information and text description information for each project. The entire datasets contains 27,660 minutes of video and about 6M words. The datasets is mainly focus on technology category, and the project is initiated in various regions of the world. A fuller description of this data set can be found in Appendix C2.

For all text, I removed stop words, filtered lemmas with less than 5 frequency, and filled and truncated each sample sequence according to distribution and model requirements after lemmatization. Finally, I set the sequence length to 1024 for the text description (512 for BERT), 512 for the audio description, and 256 for the video description. Table III,Table IV and Figure 4 illustrate the process and data statistics.

**Implementation Detail** The BiRNN model uses two hidden layers, each with 100 hidden units, and the word embedding layer uses a pre-trained Glove of 100 dimensions. The TextCNN model uses three convolution kernels, each with a size of 3, 4, and 5, and the number of channels is 100. For the hyper-parameters of the BERT+MLP architecture, we used the base version of BERT with 12 headers and 12 layers of encoding blocks, the number of hidden units is 768, and the number of hidden units in the middle of the MLP is 3072. I implemented my network on PyTorch and trained the TextCNN network using Metal Performance Shaders, and the BiRNN and BERT+MLP architecture using NVIDIA A100 Tensor Core GPUs. The learning rates are 1e-3, 1e-2, 1e-4. Batch sizes are 32, 16, 256.

**Performance** The main results are shown in Table V. First, when the signal source is text description, the overall performance of the model is better than audio and video. But it is obviously that the comparison is unfair because the amount of text is much higher than the other two modalities. The specific reason is that I don't have the budget to transcribe video content using the Video Intelligence API. (**$0.1/min**) . I try to propose a more economical feature extraction method in the **future work section**. Specifically, when the signal source is a text description, the accuracy of the baseline BiRNN is 0.656, and I achieved 7.9% and 10.5% improvements compared to the baseline. When the signal source is an audio description, the accuracy of the baseline BiRNN is 0.532, and I achieved 1.3% and 5.6% improvements in TextCNN and BERT compared to the baseline. When the signal source is a video description, the accuracy of the baseline TextCNN is 0.521, and I achieved 0.9% and 5.2% improvements in BiRNN and BERT compared to the baseline. Figure 5 shows the results of each training round when the signal source is a text description.

**Table V**. Task2 results.

| Inputs | Classifier | Accuracy | support |
|---|---|---|---|
| **Text** | TextCNN | **0.761** | 10147 |
| | BiRNN | 0.656 | 10147 |
| | BERT+MLP | 0.735 | 10147 |
| **Speech** | TextCNN | 0.545 | 160 |
| | BiRNN | 0.532 | 160 |
| | BERT+MLP | **0.588** | 160 |
| **Video** | TextCNN | 0.521 | 160 |
| | BiRNN | **0.577** | 160 |
| | BERT+MLP | 0.534 | 160 |

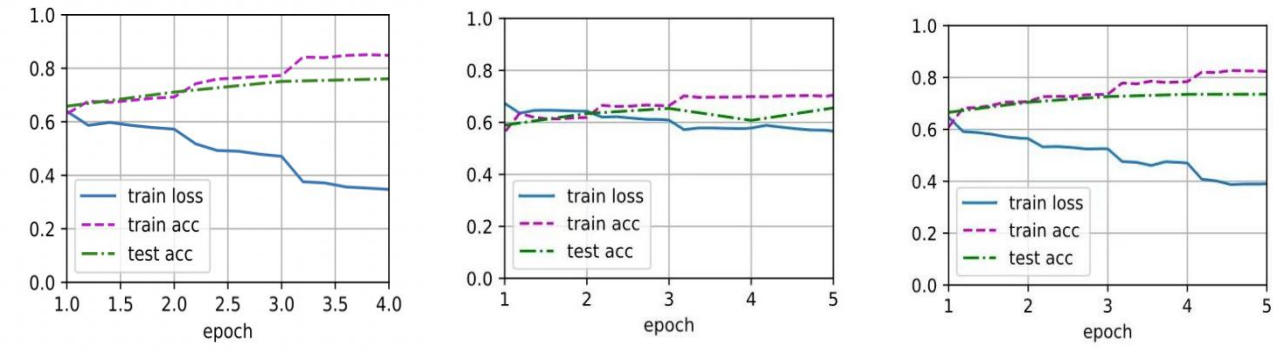



***Figure 5.***Training detail for text source.

## V. Conclusion and Discussion

In this work, I study the prediction for crowdfunding success with two type of cues and various signals. For explicit cues, I used regression analysis to explore the predictive power of project characteristics and creator information, and performed heterogeneity tests based on duration and goals. For implicit cues, I developed three mainstream network to learn the features of text,speech and video respectively.I implement my experiments on two large-scale datasets manually collected from Kickstarter.

It is evident that there is a significant cost associated with the process of extracting high-semantic text information directly from video and audio, as well as an irreversible loss of information such as color, brightness, pitch, and frequency. To avoid using expensive third-party APIs, I considered using deep neural networks to extract video and audio frame features. I utilized ResNet-152 to extract video frame features and the librosa library to extract Mel-frequency cepstral coefficients of audio, preserving the original signal's information while reducing cost. To explore multi-modal fusion's potential, I combined these two representations with the previous text's BERT representation, utilizing a MLP for late fusion. This design achieved a prediction accuracy of 72.5%, close to our best model's performance. (**more details in Appendix D**) For future work, I will focus on the following three aspects:

1. Improving the speed and accuracy of modal feature extraction methods, such as using i3d networks for video features to capture temporal information.
2. Continuing to explore the interaction between different modalities, not only implicit signals but also explicit signals.
3. Collecting standardized and extensive data sets from different crowdfunding platforms to facilitate better training for the model.

## REFERENCE


[1] Yuan, Hui, Raymond YK Lau, and Wei Xu. "The determinants of crowdfunding success: A semantic text analytics approach." Decision Support Systems 91 (2016): 67-76.

[2] Mitra, Tanushree, and Eric Gilbert. "The language that gets people to give: Phrases that predict success on kickstarter." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. 2014.

[3] Desai, Sheena, et al. "Use of crowdfunding for expenses related to medical hair loss." Journal of the American Academy of Dermatology 86.5 (2022): 1109-1110.

[4] Ma, Zecong, and Sergio Palacios. "Image-mining: exploring the impact of video content on the success of crowdfunding." Journal of Marketing Analytics 9 (2021): 265-285.

[5] Kaminski, Jermain C., and Christian Hopp. "Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals." Small Business Economics 55 (2020): 627-649.

[6] Google cloud video Intelligence : https://cloud.google.com/video-intelligence?hl=zh-cn

[7] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[8] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.

[9] Chen, Yahui. Convolutional neural network for sentence classification. MS thesis. University of Waterloo, 2015.

[10] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[11] Statasmodels : https://www.statsmodels.org/stable/index.html

[12] Cumming, DJ, Leboeuf, G, Schwienbacher, A. "Crowdfunding models: Keep-It-All vs. All-Or-Nothing. Financial Management," 2020,49: 331-360

[13] E. Mollick, "The dynamics of crowdfunding: an exploratory study, " *Journal of Business Venturing,2014.* 29 ,1-16.

[14] H. Zheng, D. Li, J. Wu, Y. Xu. "The role of multidimensional social capital in crowdfunding: a comparative study in China and US," *Information Management* 51 (2014) 488–496.

[15] J. Härkönen, "Crowdfunding and its Utilization for Startup Finance in Finland: Factors of a Successful Campaign(Master's thesis) ",2014.

[16] kickstarter.com

[17] Ho-Dac, N. N., S. J. Carson, and W. L. Moore. 2013. "The Effects of Positive and Negative Online Customer Reviews: Do Brand Strength and Category Maturity Matter." *Journal of Marketing* 77 (6): 37–53.

[18] P. Belleflflamme, T. Lambert, A. Schwienbacher, Individual crowdfunding practices, *Venture Capital* ,2013.15 (4) 313–333.

[19] Bi, S., Z. Liu, and K. Usman. 2017. "The Influence of Online Information on Investing Decisions of Reward-Based Crowdfunding." *Journal of Business Research* 71 (1): 10–18.

[20] Chan, C. S. R., H. D. Park, and P. Patel. 2018. "The Effect of Company Name Fluency on Venture Investment Decisions and IPO Underpricing." *Venture Capital,*20 (1): 1–26.

[21] Xu, B., H. Zheng, Y. Xu, and T. Wang, "Configurational Paths to Sponsor Satisfaction in Crowdfunding," 2016. *Journal of Business Research* 69 (2): 915–927.

[22] Shane, S. 2001. "Technological Opportunities and New Firm Creation." *Management Science* 47 (2):205–220.

[23] Zhao, H., and S. E. Seibert. 2006. "The Big Five Personality Dimensions and Entrepreneurial Status: A Meta-Analytical Review." *Journal of Applied Psychology* 91 (2): 259–271.

[24] Marvel, M. R., and G. T. Lumpkin. 2007. "Technology Entrepreneurs' Human Capital and Its Effects on Innovation Radicalness." *Entrepreneurship Theory and Practice* 31 (6):807–828.

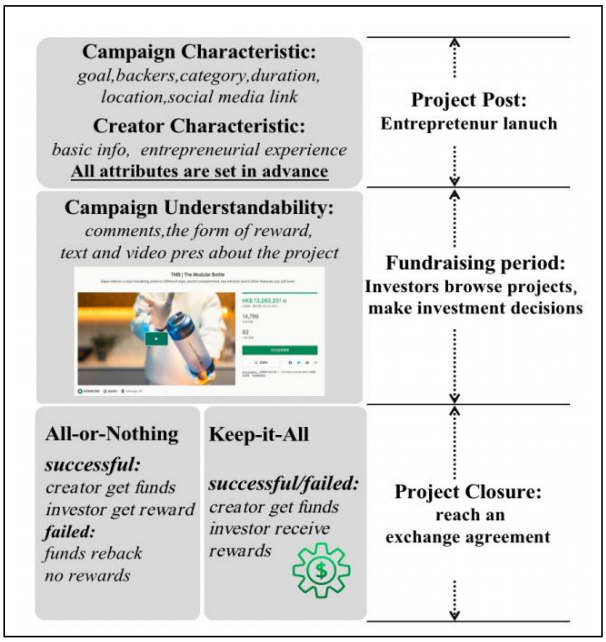
[25] Ahlers, G. K., D. Cumming, C. Günther, and D. Schweizer. 2015. "Signaling in Equity Crowdfunding." *Entrepreneurship Theory and Practice* 39 (4): 955–980.

[26] Colombo, M. G., C. Franzoni, and C. Rossi-Lamastra. 2015. "Internal Social Capital and the Attraction of Early Contributions in Crowdfunding." *Entrepreneurship Theory and Practice* 39 (1): 7100.

APPENDIX

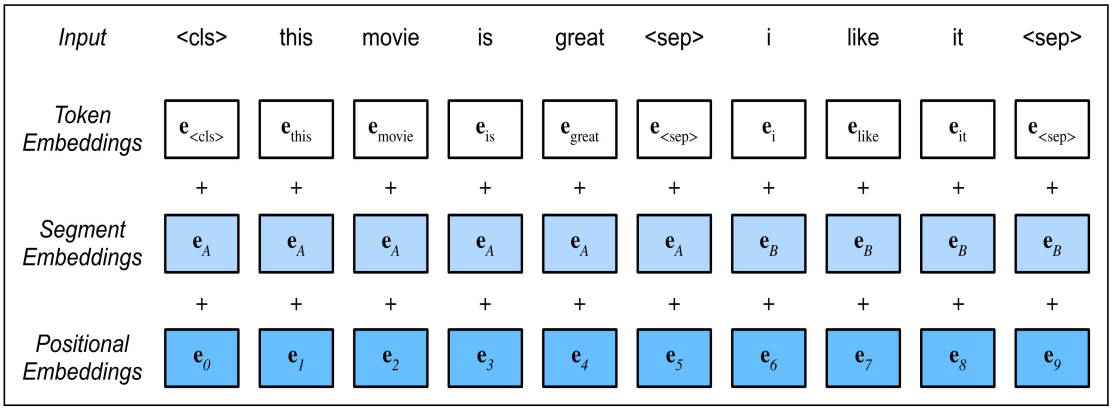
## A Related Work:A Typical Fundraising In Reward-crowdfunding

A typical process for a startup using reward-based crowdfunding to raise seed fund is: (1) Entrepreneurs launch a fundraising campaign on the site, which may be about a concept product or service. (2) During the fundraising period, which usually ranges from one to three months, typical investors judge whether to invest in the project by reviewing the text and video descriptions of the project or interacting with other investors in the comments section. The amount of funding is determined by the investors themselves. (3) Once the project is successfully financed, investors will receive non-monetary rewards related to the project provided by entrepreneurs based on the amount of investment. In All-Or-Nothing mode, a minimum fund goal will be set by the entrepreneur's own discretion. Investors will only receive rewards if the goal is met by the deadline. Otherwise, the investor will receive the money back but no return. In Keep-It-All mode,where the entrepreneur keeps the entire amount raised regardless of achieving the goal. The process is showed below.




## B Approach:BERT input embedding architecture

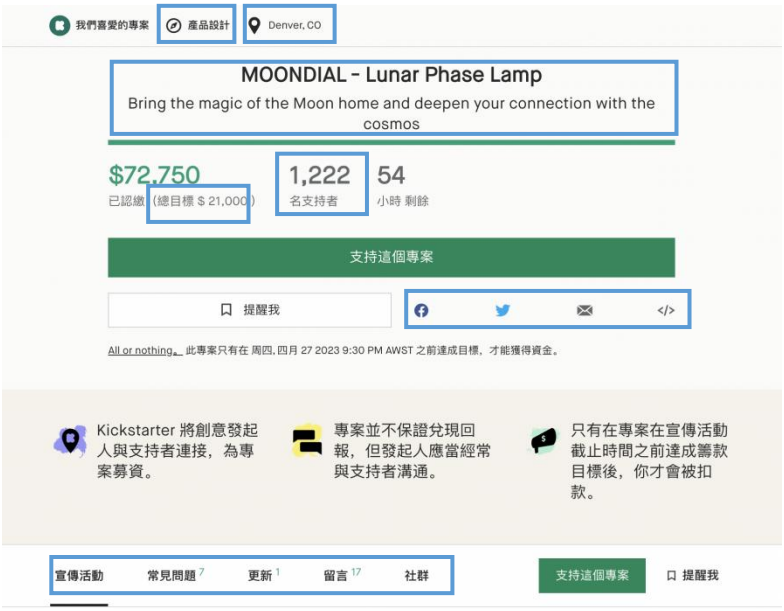
The embedding representation of BERT is the sum of the three embedding layers. When the input is a single text sequence, the word embedding layer consists of special tokens '<cls>', text words and separator tokens '<sep>'. The segment embedding layer consists of single parameter Ea. The position embedding layer adopts learnable position coding.

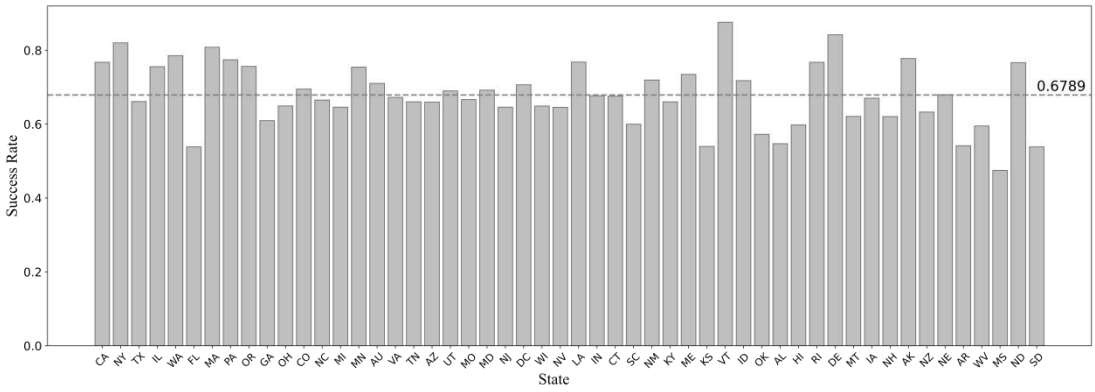


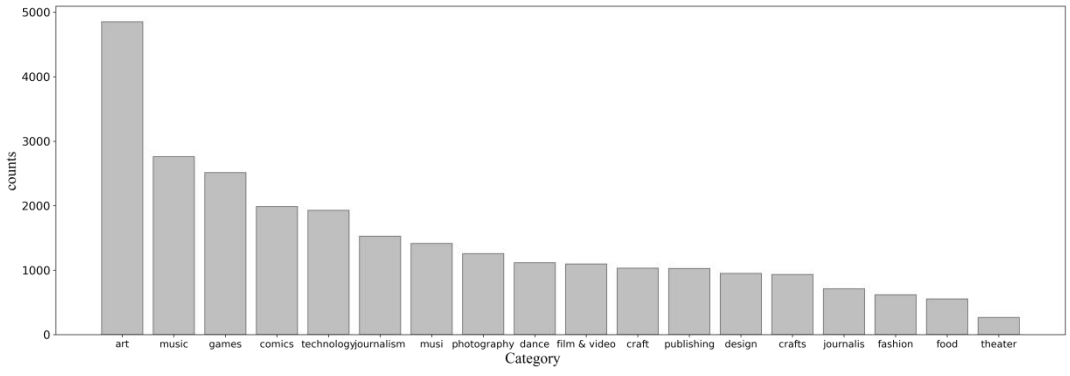

## C Experiments

C.1 Kickstarter-3.8

Kickstarter-3.8 is a collection of **38,190 project** data that I manually collected from Kickstarter. It contains project status, project creator information, and project description information including category, location, etc. but does not contain signals such as video,images and text. The entire datasets has **18 categories** distributed across **52 states** in the **United States**.

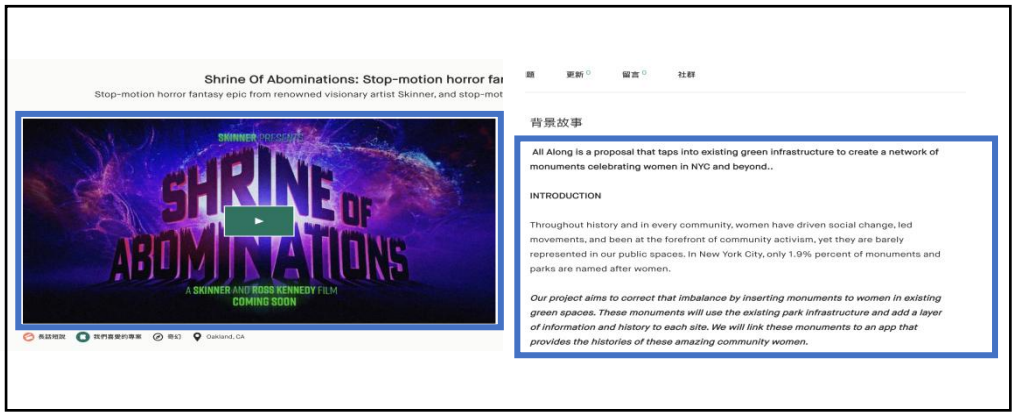


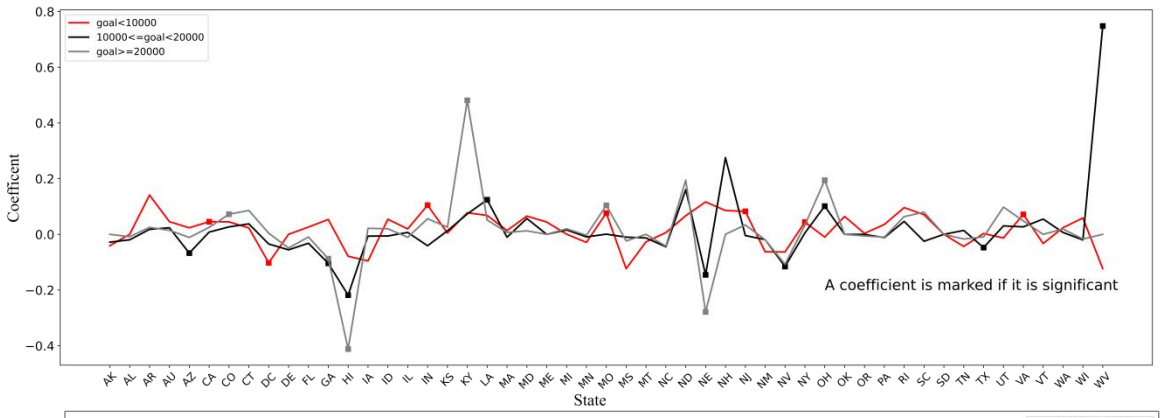

C.2 Kickstarter-1.1

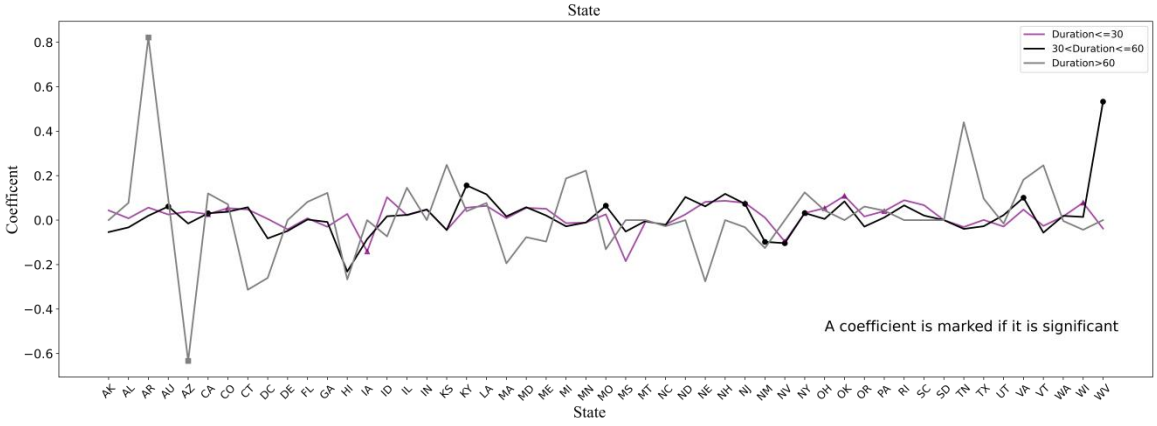
Kickstarter-1.1 is a collection of **11,064 project** data that I manually collected from Kickstarter. It contains video information and text description information for each project. The entire datasets contains **27,660 minutes** of video and about **6M** words. The datasets is mainly focus on technology category, and the project is initiated in various regions of the world.




| Source | Count | Positive Share |
|---|---|---|
| text | 11064 | 55.18% |
| speech | 11064 | 55.18% |
| video | 11064 | 55.18% |
| #categories:4 | | |
| #text length (mean):694 words | | |

C.3Heterogeneity tests



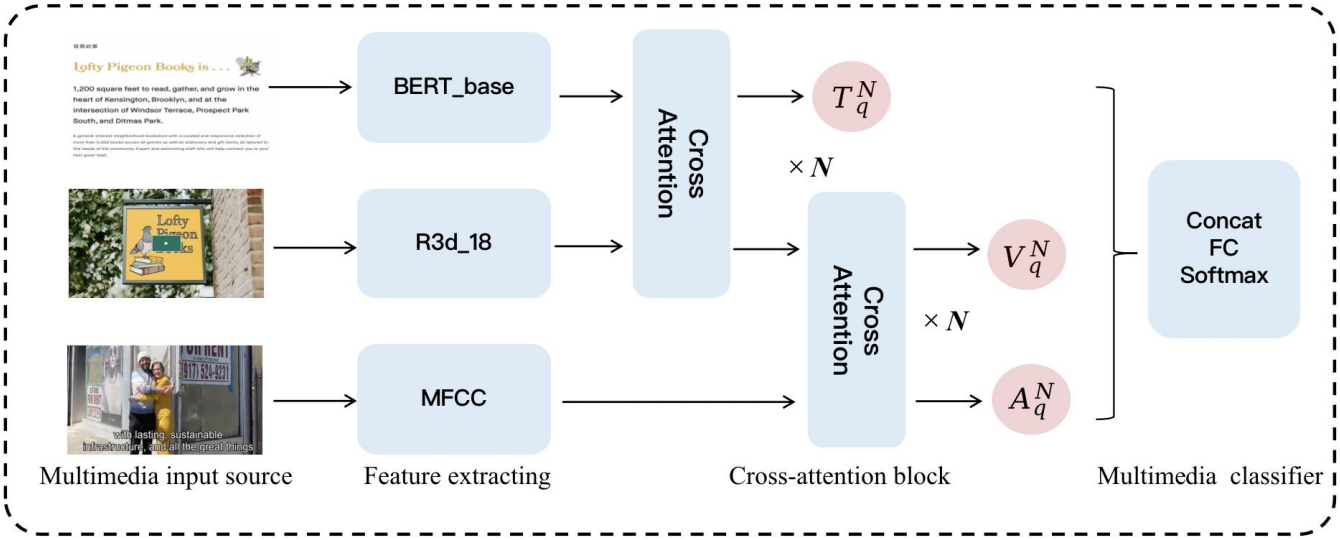




## D Feature Work:Multi-modal Analysis

#**Pseudocode for model fusion.**
#text_encoder-BERT_base
#speech_encoder-MFCC
#video_encoder-ResNet-152
#T[n,1,512]-minibatch of texts
#S[n,1]-minibatch of speeches
#V[n,h,w,c,f]-minibatch of videos

#extract feature representations of each signals
T_M=text_encoder(T) #[n,1,768]
S_M=speech_encoder(S) #[n,1,20]
V_M=video_encoder(V) #[n,1,400]

#signals fusion
F_M=concat(T_M,S_M,V_M,axis=1)
R_M=MLP(F_M)




## E IS6912 Materials Submitted Screenshots