CS230

Prediction of Surgical Crowdfunding Campaign Success using NLP

Advait Patil
Department of Computer Science
Stanford University
advaitp@stanford.edu

Hanna Lee
Department of Computer Science
Stanford University
hylee719@stanford.edu

Vinita Cheepurupalli
Department of Computer Science
Stanford University
vinitac@stanford.edu

Abstract

Online crowdfunding platforms have emerged as an increasingly popular funding modality to cover healthcare costs, especially for vulnerable populations such as those that are uninsured or underinsured. However, outcomes for crowdfunding campaigns for surgical care remain poorly understood. Here, we scraped and curated a new dataset of 66,514 surgery-related campaigns from 2010-2020, requesting a combined total of 1 billion dollars sought and $354,849,732 raised. We will apply deep learning to extract word embeddings from campaign text, and then develop a binary classification LSTM network using these word embeddings to predict whether a campaign is able to raise its requested funds. Our final model had an accuracy of 0.6042 and F1 score of 0.3766. Given the sheer difficulty of this prediction task, this is a reasonable improvement over the manual baseline in the literature. To our knowledge, this is the first study to examine surgical campaigns and the first to apply deep learning to predict campaign success across all types of medical crowdfunding campaigns. As a whole, our work comprises important paths forward for application of deep learning to crowdfunding data.

1 Introduction

Nearly 41 million Americans lack adequate access to care or ability to pay for care [1]. Online crowdfunding platforms have emerged as an increasingly popular funding modality to cover healthcare costs, especially for vulnerable populations such as those that are uninsured or underinsured, and are directly linked to health disparities and gaps in social safety-net systems. However, outcomes for crowdfunding campaigns for surgical care remain poorly understood.

In our study, we aim to develop a deep learning framework to predict success (as defined as 100% or more of donations requested) of an online fundraiser for surgical conditions. We will use a dataset of sentences representing campaign titles/descriptions, along with the associated funds requested as well as funds actually obtained. We will then use Bert in pytorch to extract word embeddings, which are used as input to naive bayes, an SVM, and an LSTM model that will then predict whether a campaign was able to obtain its requested funds. This is the first study aiming to examine surgical crowdfunding campaigns and to predict campaign success.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

Characteristics of online crowdfunding campaigns for medical conditions remain understudied. To our knowledge, this is the first study to specifically examine surgical campaigns. In addition, this is the first study to use word embeddings extracted from campaigns to predict success of the crowdfunding campaign.

Prior research by Angraal et al. provides a broad overview of the contribution of cancer, cardiovascular conditions, neurological conditions, and trauma/injury to the total pool of campaigns [2]. The work of Zhang et al. and Zhu et al. provide past approaches for clinical concept extraction as well as biomedical word embeddings, including BERT, which we will lean on for our research [7, 8]. These methods have been demonstrated to be powerful on EHR data, but have not been applied to crowdfunding data.

We aim to use BERT as a word embedding method for processing the text associated with each fundraiser [9]. Similar embedding methods have been used in research involving the "Evaluation of Internet-Based Crowdsourced Fundraising to Cover Health Care Costs in the United States" [2]. Howerver, these word embeddings were only used to cluster campaigns into four broad categories. In addition, this study did not examine surgical crowdfunding campaigns and used a limited dataset that may not be fully representative. We aim to expand on this area by using an almost fully-comprehensive dataset of campaigns.

To our knowledge, no studies to date have attempted to predict success of medical crowdfunding campaigns with computational methods. Aleksina et al. performed the task by hand for medical research campaigns on a dataset of 109 campaigns [4]. While important work, this area would greatly benefit from data-driven approaches that leverage the vast publicly available data, as all crowdfunding campaigns by definition are available on the internet. Given that the 'state of-the-art' is a manually-curated analysis, our work will help pave an avenue for a better solution to the problem at hand.

3 Dataset and Features

We used data extracted from the largest crowdfunding platform, GoFundMe, from its inception in May 2010 through December 2020 using a webscraper (code provided by Silver et al) [3]. To account for missing campaigns, InternetArchive.org's Wayback Machine was also scraped to access cached versions of completed (but since-deleted) campaigns. This dataset contains 1.8 million scraped URLs of all campaigns in the last decade, including information on campaign title, creation date, location, description, amount of money raised, and the goal amount.

We excluded duplicate campaigns, any campaigns with a non-US location, and those that used a non-US currency. Next, we curated a surgery-specific dataset of surgery campaigns by filtering by filtering to campaigns that contained the substring "surg" anywhere in the title or description. We excluded any campaigns with missing information. This resulted in a final dataset of 66,514 surgery-related campaigns from 2010-2020, requesting a combined total of 1 billion dollars sought and $354,849,732 raised. Table 1 shows an example campaign from the dataset.

title	story	raised_ amount	goal_ amount	percent _raised
The road to surgery!	It's been almost a full year living with a hernia in my belly button. it limits me from being active, even moderately. i can't continue to be a waiter (my primary source of income over the last 5 years) or even a busser, the activity really affects my hernia and makes things painful. i cannot dance, carry things, spin fire, or any other physical activity without aggravating it. so this year i'm determined to get it stitched up! all proceeds go towards the surgery. a big tight hug and sincere thank you to all who help support my surgery. your support means the world to me! <3	1662	2400	69.25

We used a 90/10 training/test split, with 59,863 in our training set and 6651 in our test set. We then extracted word embeddings using BERT, which were then used as features along with numerical data such as description length and time duration of campaign when training our naive bayes, SVM, and LSTM models for prediction of campaign success.

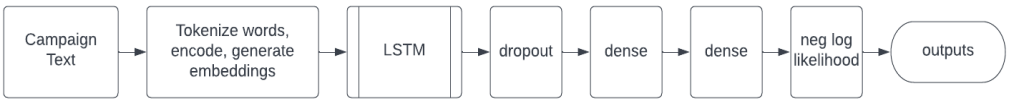
4 Methods

We tried several different approaches for our binary text classification task (predicting whether a medical crowdfunding campaign will raise 100 percent of its goal or not): logistic regression, naive bayes, a support vector machine (SVM), and an LSTM network, a type of recurrent neural network (RNN), using the Keras and TensorFlow packages [5, 6]. As a baseline model, we used logistic regression with the outcome as whether the crowdfunding campaign raised their goal amount. We employed Lasso Regression (L1) and 5-fold cross-validation to tune the hyperparameter lambda.

Some different embedding methods we tried include ELMo, BERT, and a token/frequency word vectorization method. Ultimately, we opted to use a tokenizer method for the text representation of each "story" associated with each campaign. We wanted to emphasize the importance of key words, such as "surgery" or "cancer"; also, we decided to pursue a sequential LSTM network as our main model.

LSTM was chosen because is it known for having good memory efficiency, due to the use of cell memory, and for our problem of text classification we thought it would be important to emphasize the memorizing key information. Additionally, this helps address the vanishing gradient problem that can arise in other RNNs. The descriptions for each fundraiser can be quite lengthy, and thus an LSTM is a reasonable choice for attempting to extract string meaning. Another one of our goals was to backpropagate to update weights during training, since many successful and unsuccessful campaign descriptions could be prone to having similar text- our LSTM architecture allows us to do this.

Our model layers are as follows:



For our loss function for the LSTM, we chose negative log likelihood loss, which is known to be useful for binary classification tasks:

$$\min -l(\theta)$$
$$w.r.t\ \theta$$

$$\min \left(-\sum_{i=1}^{m} \left(y^{(i)} \times \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \times \log\left(1 - h_\theta\left(x^{(i)}\right)\right) \right) \right)$$
$$w.r.t\ \theta$$

Originally, we wanted to use the campaign story text in addition to numerical features (story length, duration of campaign, and U.S. state location (represented as a one-hot vector)) as a way to improve our model, but found that there was no real improvement with this layer of meta data being concatenated with the textual data, so opted for a model that only used the story text.

Also, we found that the distribution in our webscraped crowdfunding data was heavily skewed; a vast majority of the campaigns had not reached their goal amount. Because of this, in the early stages of our models, we observed that relatively high accuracies were achieved while predicting that every campaign would be unsuccessful. To combat this, we performed a mixture of over sampling and under sampling for our final LSTM model. Successful campaigns were in the minority, so we oversampled, and we under sampled unsuccessful campaigns, resulting in a training ratio of about 2:3 (successful: unsuccessful).

In addition to the other model adjustments made during the process, we experimented with hyper-parameter tuning and modified our batch size, number of epochs, and learning rate. We also tried improving our pre-processing methods by trying different ways of tokenization and filtration (e.g. removing stop words, lemmatization).

We tested different evaluation metrics throughout this project, including accuracy, precision, recall, F1 score, and loss. With the understanding that both precision and recall are important in the context of our problem (false positives and false positives both have relevant financial consequences to healthcare cost coverage), we decided upon using F1 scores as our primary evaluation metric.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

5 Experiments/Results/Discussion

The main hyperparameters we chose to manipulate when training our LSTM are as follows:

a) batch size: We started with a batch size of around 40, but found that increasing the batch size to around 64 improved our results and yielded slightly better data representation.

b) epochs: After experimenting with different numbers of epochs, we found that 5 yielded the best results out of the ones we tried.

c) patience: This parameter was responsible for the early stopping condition and we found that changing it did not have a huge impact on our results. Our value was 3.

d) learning rate: While we started with a smaller standard learning rate of 3e-4(for AdamW Optimizer), we found that increasing it to 3e-3 yielded slightly better accuracy and F1 scores for our model.

Results from our baseline models as well as LSTM network are shown below.

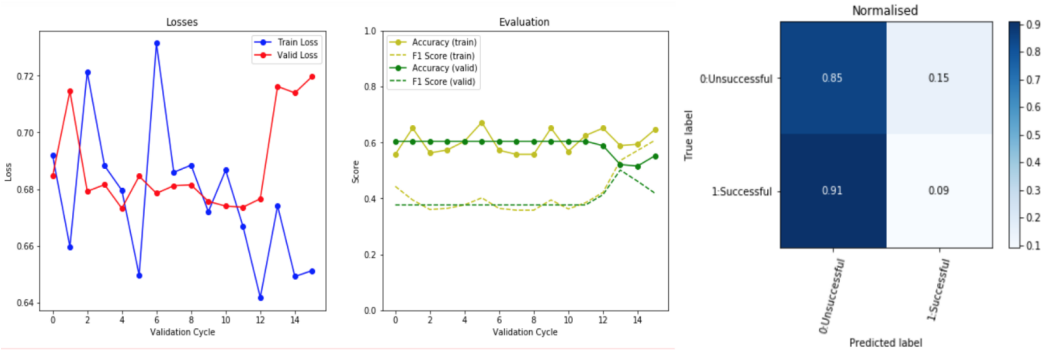
	Logistic Regression	Naive Bayes	SVM	LSTM
Accuracy	0.5193	0.8585	0.8585	0.6042
F1	0.2172	0.4619	0.4619	0.3766

As mentioned above in our methodology, our early models achieved a high accuracy by predicting that none of the test campaigns would be successful, due to the nature of our data. In our Naive Bayes and SVM models, we did not perform under sampling and over sampling, and the identical high accuracies from both models indicate this problem.

The logistic regression model did not perform very well, obtaining an accuracy close to random guessing. Our LSTM model performed slightly better, achieving an accuracy of about 60 percent. However, this also does not appear to be a highly effective model.

Pictured below is a confusion matrix, which illustrates the model's failure to identify true positives. This could partially be due to the unequal distribution between successful and unsuccessful campaigns, or it could be a result of the model architecture. We found that there was consistent success in identifying unsuccessful campaigns.

As we can see in the figure below, the F1 score remained low relative to the accuracy, but saw gradual improvement. It is also promising that the accuracies between the train and validation cycles stayed somewhat similar.



Based on our training loss trending downward and validation loss trending upward (pictured above), we conclude that our model was overfitting to our training set. However, this is an improvement from an early iteration of our model, which did not have under sampled and over sampled data to create a more equal distribution of successful and unsuccessful campaigns. We also attempted to rectify the overfitting by including early stopping, although we still never reached an optimal performance level.

Overall, we believed that due to the similarity in content of medical crowdfunding campaigns, both successful and unsuccessful, the nature of our task using text embeddings proved quite difficult, despite our attempts to ameliorate our model through pre-processing methods and adding meta data.

6 Conclusion/Future Work

Surgical crowdfunding campaigns remain completely unstudied in the medical or computer science literature, and represent a unique breakdown of the healthcare system for medically marginalized and underinsured populations. We tested logistic regression, naive bayes, SVM, and LSTM models and while performance was limited due to the difficulty of the task (accuracy = 0.6042, F1 = 0.3766), this comprises important steps forward for the literature as this is an unexplored area.

Some interesting ideas for future work include:

1) Expanding our neural networks to further medical conditions beyond those requiring surgical intervention. This would involve expanding the filtered dataset as well as considering new analysis methods to consider (such as clustering by medical condition subtype).

2) Exploring further techniques to deal with class imbalance. Successful completion of a crowdfunding campaign's goals is a relatively rare event within the dataset, leading to challenges of class imbalance. Synthetic Minority Oversampling Technique (SMOTE) is one good avenue to generate synthetic data for the minority class.

3) Investigating different methods of text embedding/feature combination with more crowdfunding data to better identify factors that are conducive to predicting the success of a medical crowdfunding campaign.

7 Contributions

All authors contributed equally to this project.

Advait: Project conceptualization, used webscraper and obtained dataset, literature search, data processing and filtering, regression model, and write-up.

Hanna: Explored several different text embeddings methods (ELMo, BERT, tokenization/vectorization), tested several models (SVM, Naive Bayes, LSTM), contributed to methodology and results discussion, experimented with different text pre-processing and nn layer concatenation methods.

Vinita: Explored embedding methods (ELMo, BERT), explored LSTM model, organized GitHub repository

Additionally, we would like to thank our project mentor, Sarthak Consul, for his useful guidance and advice throughout the quarter!

References

[1] Galvani AP, Parpia AS, Foster EM, Singer BH, Fitzpatrick MC. Improving the prognosis of health care in the USA. Lancet. 2020 Feb 15;395(10223):524-533. doi: 10.1016/S0140-6736(19)33019-3. PMID: 32061298; PMCID: PMC8572548.

[2] Angraal S, Zachariah AG, Raaisa R, et al. Evaluation of Internet-Based Crowdsourced Fundraising to Cover Health Care Costs in the United States. JAMA Netw Open. 2021;4(1):e2033157. doi:10.1001/jamanetworkopen.2020.33157

[3] Silver ER, Truong HQ, Ostvar S, Hur C, Tatonetti NP. Association of Neighborhood Deprivation Index With Success in Cancer Care Crowdfunding. JAMA Network Open. 2020;3(12):e2026946. doi:10.1001/jamanetworkopen.2020.26946

[4] Aleksina A, Akulenka S, Lublóy Á. Success factors of crowdfunding campaigns in medical research: perceptions and reality. Drug Discov Today. 2019;24(7):1413-1420. doi:10.1016/j.drudis.2019.05.012

[5] Chollet F. Keras: Theano-based Deep Learning library. Code: https://github. com/fchollet. Documentation: http://keras.io 2015

[6] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv Preprint arXiv:1603.04467, 2016.

[7] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data. 2019;6(1):52. doi:10.1038/s41597-019-0055-0

[8] Zhu H, Paschalidis IC, Tahmasebi A. Clinical concept extraction with contextual word embedding. arXiv. Preprint posted 2018. Accessed November 9, 2022. https://arxiv.org/abs/1810.10566

[9] Devlin, J. Chang, MW. Lee, K. Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Preprint posted 2018. Accessed December 5, 2022. arXiv:1810.04805