Colective Behavior and Crowd Funding Campaigns

Dominic Rendero
Computer Science and Engineering
University of Connecticut
Storrs-Mansfield, United States
dominic.rendero@uconn.edu

Matthew Caro
Computer Science and Engineering
Univeristy of Connecticut
Storrs-Mansfield, United States
matthew.caro@uconn.edu

I. Introduction + Backround

Over the past few decades, technology and its integration into people's daily lives has exploded. With the addition of personal computers, laptops, smart phones, and tablets into the average consumer's household, they are now more and more integrated into the online sphere of social media. This social media has become so ingrained into people's live that they are spending hours a day scrolling, posting, and consuming content to both their friends and the public, allowing all different information to spread, and be seen by people from all different walks of life.

Before social media development and the growth of the online space, ideas were spread by word of mouth at shared activities or in the less distant past, the news and radio. Whenever there was some tragedy or someone needed help funding a new venture it fell to the families, friends, and the communities around those people to help. They did this through labor, financial contributions via crowd funding, or both. It could be local religious organizations that would go clean up debris from a natural disaster, a local charity group helping to fund reconstruction of a house fire, or something as simple as loaning/gifting money to a friend so they can get their business up and running.

While this community crowd funding still exists, it has also changed a lot by the introduction of social media and widespread access/ownership of innovative technologies. Where in years past crowd funding was mostly done by geographically and personal relationships in tight-knit communities this is no longer the case. With the new rapid spread of information, this financial generosity is now coming from more communities than ever before, and the amount of help people can get now surpasses anything from the past. People have been seen to gain thousands if not millions of dollars in response to tragedies or for business ideas in the belief that the money is more impactful for others as opposed to themselves.

The current largest crowd funding website for tragedies is none other than GoFundMe, having over 100 million different donors over their time as a company. GoFundMe has led the way to allow people from all over the world to receive help in their time of need, normally from cascading systems in different online and social media communities. Following the George Floyd incident back in 2020, between traditional media and social media spreading the story as well as his families GoFundMe link nearly $40 million was raised for the Floyd family showing how impactful this technology is towards raising money during tough times.

There is also a different form of crowd funding seen on websites such as Kickstarter and Indiegogo. This form of crowd funding is less charitable but is still a big part of the crowd funding market. These websites seek to get donations for some upcoming project or product that one is seeking to work on but do not have the appropriate funds to finish the product or make it on a larger scale. People donate to these projects in the hope that one day the product will be made and available for purchase. While it is less of a charitable venture and more about further advancement it is just as much a part of crowd funding as the contributions brought in at GoFundMe.

Now there have been multiple papers that seek to predict if a crowd funding campaign will be successful and reach the intended goal put forth by those running the campaign. There have also been several papers that use

machine learning to try to predict how collective behavior will be affected by a cascading system such as social media. That said, there have not been those with the goal of bringing these two concepts together in one paper. This project seeks to build on these topics and use machine learning to determine if there will be enough. Specifically, on all these websites a campaign will have a goal amount of money the campaign considers a success. Our original goal is to modify existing studies such that it can accurately predict if these goals will be met or not based on how it has been shared and interacted with on social media. Sadly, the access to the data set containing the shares of a given campaign was taken down and we no longer have access to it. Instead, we now have the goal of making a prediction model for Kickstarter campaign successes based on the category that the campaign falls under. While not the same as social media shares, we feel that this is a suitable replacement given that it still largely shows collective behavior because people in the space that the campaign has to do with would be the ones most likely to donate and help.

Possible expansion for this project that we may look to add includes predicting the ratio of funds given as compared to what the campaign is asking for. This would change the problem from a binary classification problem into a more complex prediction based on other projects. While this is certainly a possible expansion it likely will not be complete prior to deadlines given the time constraint but we may revisit it assuming our models for the base project are up to standard.

II. Related Work and Paper Survey

While there are several studies done related to crowdfunding, and cascading systems, none explore the idea of how social media cascading will affect the success of a crowdfunding campaign. A study published in the *Journal for Consumer Psychology* highlights parts of social media that will benefit or harm the success of a GoFundMe campaign [3]. While the study demonstrates social media influence, it does not do a deep dive into the cascading system area of social media. We can build off ideas that are mentioned in this study to see how the posts of people with a large social media presence will cause others to post, causing a crowdfunding campaign to be successful.

GoFundMe is a personal crowd funding website where people can donate without a means of financial compensation to assist others. In equity crowdfunding, investors are given a stake of the company they are investing in. A study published in the Entrepreneurship Theory and Practice journal researched the effects early investors had in crowdfunding [4]. Early investors create an information cascade, which is a type of cascading system. When initial investors commit funds to a project, it indicates confidence, encouraging others to follow suit even without detailed personal evaluation. This is similar to social media cascading for GoFundMe campaigns where influencers supporting a cause will lead to other people supporting a cause. People will use other people's judgement to donate to a GoFundMe page. This is especially true when that person is already on their social media network. In addition to using other studies as inspiration for our project, it is important to fully understand cascading behavior in social media.

Another research paper that explains and demonstrates cascading behavior well is "Cascading behavior in complex socio-technical networks." This paper looks at cascading behavior, specifically in socio-technical networks [7]. These are networks where human behavior and technological infrastructure interact. This can become a consideration for our prediction model because socio-technical networks such as school settings or workplaces lead to changes in crowdfunding success. The combination of word of mouth and social media influence can have a large impact on the amount of people that donate, and who donates. This paper examines social contagion which is like the idea of crowdfunding within a social media network. The diffusion of information has similar characteristics as social media post or influencer endorsement.

Some specific forms of social media cascading systems have been explored. A research paper published by the Society for Industrial and Applied Mathematics focused on blog posts, and how re-posts, discussions, and comments are triggered by an initial event [5]. Although blogs are still a form of social media, there are fundamental differences between blog posts and the current more popular social media like X formally known as Twitter, Facebook, Instagram, Ect. Modern social media is more focused on short content, where blog posts are typically lengthier. There are common trends between the two types of social media, such as the larger follower count one has the larger cascade events are. The biggest change is the rate information spreads. Due to the shorter content, it is easier to read and share. This also creates trends that can spread information

extremely quickly. There is also a bigger outreach to modern social media networks. This paper is a good reference for how cascading systems work. In our project, we will keep this information in mind when creating our algorithm and analyzing data.

For this project, it is important to fully comprehend information cascades from a broader point of view. A study published in *Journal of Complex Networks* studies what factors influence information cascades [6]. The networks studied in this are Random networks, Scale-Free Networks, and small-world networks. Each of these networks are defined based on how each node is connected to other nodes. Different types of social media conform to each of these networks. A scale-free network having a few highly connected nodes, and many nodes with fewer connections. There are many lower connected nodes connected to a few highly connected nodes. This type of network is the most prominent in social media. For example, celebrities or social media influencers are connected to many different people. These nodes/people are crucial in the spread of information and start of information cascades. Popular social media platforms that follow this network are Twitter, Instagram, and YouTube. These platforms have many people with few followers/subscribers, and a few people with immense follower/subscriber count. There are other social media platforms that follow different complex social networks as well. A small-world network is defined as networks where there is high local clustering, with short average path lengths. This makes clusters of nodes that are interconnected with each other. Social media websites that follow this pattern are Facebook and LinkedIn. Information cascades in these networks are rapid within the clusters but may get stuck at in the cluster. There must be a bridge node to another cluster for the information to cascade to and be influential enough to continue throughout that cluster. The last network types are random networks. The area where nodes have a certain probability of being connected to another node. There is a uniform distribution of connections. Social media that follows this trend, although not exactly, are websites like Reddit, 4chan, or other places where animosity is favored. Users are less likely to have connection to other users. Understanding these social networks and how they relate to social media is important in our project to predict the success of crowdfunding campaigns. Certain networks will behave differently. For example, in a small-scale network a single person has the ability to influence the campaign immensely, whereas in a random network 1 person will not have as much as an effect.

III. Proposed Solutions

A. Data Preprocessing

The first thing we must do without datasets is preprocess the information and ensure there is no useless information that is kept in the dataset. Datasets can often be noisy, incomplete, or instructed. It is also important to normalize the data within each column, such as turning everything to lowercase, ensuring numbers are in the same format, etc. It is also important to match information if using multiple datasets making sure the same information is in the collum for both. The strategy our group understands the best for data manipulation is Python utilizing NumPy. We could also use NLP tools such as NLTK, spaCy, or TextBlob. Create Cascade detection

This is where we will create a model to find information cascades regarding a specific crowdfunding cause. Here we will have to quantify metrics for how large of a cascade happens. There are several different metrics we can use for this. The first is Reach, which is the total number of users exposed. Another is Depth, which is the levels of cascade it goes through from the original post. There is also Breadth, which is the number of users that interact with the post at every level of the cascade. Speed is also a metric that can be used. Each of these metrics has positive reasons why they should be used. For reach, the more people that see a post about the crowdfunding the more likely someone will assist in the funding. The positive depth is to see how many types of people saw the post. Breadth we believe is the best option, because people who interact with a post means it was significant enough for them to have an action upon. This could be combined with the speed of the cascade, which is important as shown in the research before where crowdfunding campaigns that do not reach the goal within 10 days are much less likely to. We then must build algorithms to identify cascaded crowdfunding posts.

B. Create predictive model

Next, we must use these cascading algorithms to analyze and predict crowdfunding campaigns that will succeed or fail. We can test several types of machine learning to see which gives the best results. Logistic regression, random forest, and neural networks can all possibly give good results. We have several different

options to create and train this model including Scikit-learn or TensorFlow. We will evaluate the models on accuracy, precision, recall, and F1 score.

C. Visualize data.

The last process of our project will be visualizing the data once we have different models trained and tested on different datasets. We will create charts and graphs to analyze and compare the data. A tool to do this is Matplotlib. We will create charts for the precision and accuracy of each model trained on different datasets and evaluated on different test sets to provide accurate results.

IV. Results

A. Dataset

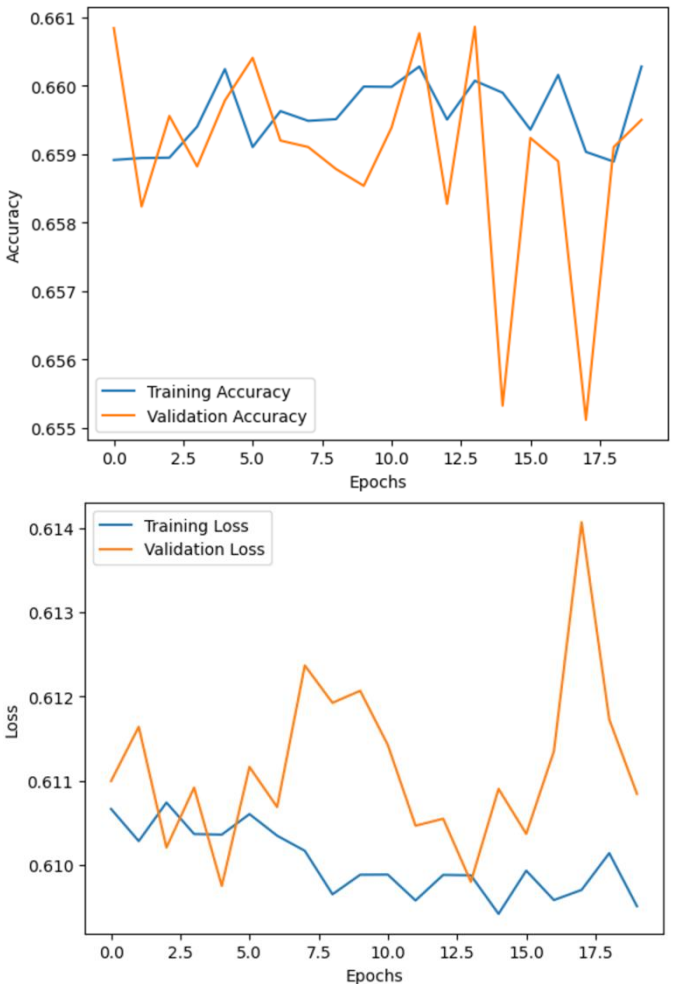
As stated in earlier parts of this paper, the original intent was to use a data set based on the website GoFundMe and the link between shares and the campaign succeeding. Unfortunately, that data set it no longer available for us to use and hence we have changed our project slightly. We now are using a data set found on Kaggle detailing Kickstarter projects and many different stats to go along with them.

B. Data Preprocessing

The first thing that we did when approaching our project was to identify what data we believed that we may need for our project. When opening the CSV file for our dataset we were met with 14 different columns of data, each showing a different metric that was recorded for each crowd funding campaign. While many of the columns could be useful others were not and so we simplified the data set down to the main category, the goal amount in USD, and the length of time that the campaign was live for. We felt like these would be the most important parts of the data to allow the most commonality when training. We chose first felt that it was most important to have the categories given that that is the main part of this project. Then we felt that timeline likely was a major proponent for if a campaign was successful or not. Next, we wanted felt a large contributor was likely towards the success was the total amount asked for, we feel a higher goal is less likely to succeed so we felt it important to train on. We also wanted to standardize this, so we made sure to have all the data in USD rather than the original currency associated with the campaign. Lastly on the data processing we needed to check if our data set was balanced or imbalanced, so we used Excel to find that approximately 35% of our data was deemed successful and hence it was balanced enough for us to start developing models.
We have experimented with different data preprocessing techniques in an attempt to increase the accuracy of our models. A couple of these attempts include using a ratio of goal:days to complete the goal, removing the goal amount altogether, and adding weights to certain parameters.

C. Models

So far, we have trained a total of 3 models all with aspirations of getting over 80% accuracy but unfortunately, we have repeatedly come up short on that front. The first model that we trained was a logistic regression model. Originally, we felt that this would likely be a decent model given the problem that we were trying to solve was essentially a binary classification, 1 if the campaign succeeded and 0 if it failed. The models and configurations that we have attempted all receive a similar accuracy to each other. They all fall between 60-70%, with some models performing better than others. The highest accuracy we have received from a model was logistic regression, while only considering the category as a parameter. This model achieves 66% accuracy, but the recall for successes was 11%. This recall can be improved in several different ways. One way we tried to increase the recall was to create synthetic oversampling using the SMOTE library. This led to the recall being increased to 67%, at the cost of dropping the accuracy to 60%. The attempts to train a neural network model led to the same trend. We have an accuracy of around 66%, and a very low recall. This is a graph of our base neural network model.



In addition to the neural network we attempted a random forest model. Unfortunately, neither of these have surpassed the logistic regression model to this point and hence we are trying

to better calibrate the logistic regression model. We are also considering developing an SVM model that we hope will be able to beat our current models but that still is not complete. To this point our best model is simply the logistic regression model that we are continuing to work on.

V. Conclusion

This project will look to experiment with a dataset with the hope to train a machine learning model that can predict the success of a given crowd funding campaign based on collective behavior and category it belongs to. The data set that we plan to use are based on stats of the crowd funding website Kickstarter that includes the campaign name, dates posted, days passed, category, goal amount in USC, and if the campaign was a success or failure.

We feel that each these categories are all impactful for if a given campaign is going to succeed or not. To ensure consistency we also made things as standard as we could be translating the prices all into USC and checking to make sure that the data was indeed balanced.

So far we have made a model that has a decent starting accuracy at 66% but we plan to continue to tweak it as well as work on more models such that they will be capable of hitting higher accuracies with less loss. By the end of this project we hope to accurately predict if a campaign will succeed over with over 80% accuracy and present graphs to show its growth throughout training.

The preliminary timeline for the project is listed below:

Week	Goal
13	BREAK Optimize current models and Develop an SVM
14	Prepare Presentation
15	Final Presentation

This timeline allows for the project to be complete in a reasonable time with time left over to prepare a presentation. If all goes well and the timeline is not too tight, more may be incorporated into the project, but it is more important for the time to ensure that the base goals are met for the end of the semester.

Responsible Contents

Throughout this project both Matthew Caro and Dominic Redero have done equal contributions. They have regularly held meetings where they have strategized, researched, and reviewed materials pertaining to these projects during these meetings. They also worked together to preprocess data and develop multiple different models to test which works best. Dominic was responsible for the first presentation while Matthew will be responsible for the final presentation. Over the break they intend to continue to communicate and work on the project both on their own and during meetings.

References

[1] Gvsa123. (2020). *GoFundMe* [Source Code]. GitHub. https://github.com/gvsa123/GoFundMe

[2] Kabure. (2019). *Kickstarter projects EDA, statistical tests, pipeline* [Notebook]. Kaggle. https://www.kaggle.com/code/kabure/kickstarter-projects-eda-stat-tests-pipeline

[3] Dehdashti, Y., Namin, A., Ratchford, B. T., & Chonko, L. B. (2022). The Unanticipated Dynamics of Promoting Crowdfunding Donation Campaigns on Social Media. Journal of Interactive Marketing, 57(1), 1-17. https://doi.org/10.1177/10949968221074726

[4] Vismara, S. (2018). Information Cascades among Investors in Equity Crowdfunding. Entrepreneurship Theory and Practice, 42(3), 467-497. https://doi.org/10.1111/etap.12261

[5] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst, Proceedings of the 2007 SIAM International Conference on Data Mining (SDM). 2007, 551-556

[6] Mahdi Jalili, Matjaž Perc, Information cascades in complex networks, Journal of Complex Networks, Volume 5, Issue 5, October 2017, Pages 665–693, https://doi.org/10.1093/comnet/cnx019

[7] J. Borge-Holthoefer, R. A. Baños, S. González-Bailón and Y. Moreno, "Cascading behaviour in complex socio-technical networks," in Journal of Complex Networks, vol. 1, no. 1, pp. 3-24, June 2013, doi: 10.1093/comnet/cnt006. keywords: {contagion;diffusion;social influence;computational social science;big data},