Visualizing Crowdfunding

Alexander Chao
UC Berkeley B.A. Statistics 2015
2601 Channing Way Berkeley, Ca 94704
alexchao56@gmail.com

ABSTRACT

With websites such as Kickstarter and Indiegogo rising in popular appeal, the area of crowdfunding has become a modern phenomenon that has allowed people to easily create campaigns and raise support for their various projects. As an area of analysis, crowdfunding has largely featured literature that focused more on predicting the success/failure of campaigns. However, as a field of visualization, the data has relatively been left untapped; most visualizations that exist simply show the accuracy of these prediction algorithms. Because data surrounding crowdfunding is mostly textual, modern text mining techniques can provide a significant tool to aid in the exploration and interpretation of crowdfunding data.

So, using scraped data from Indiegogo web pages and the R language's several text mining and clustering libraries, several data structures like a Term Document Matrix and visualizations like dendrograms, word-frequency matrices, and word clouds were created on campaign descriptions, perk descriptions, campaign titles, as well as user comments in categories such as business/tech, community, the arts, and grassroots campaigns. What was found was that when comparing successful and unsuccessful campaigns across categories, there were both similar and dissimilar words used by both the campaign descriptions and user comments. However, as a whole, more descriptive and unique words were characteristic of successful campaigns. Also sentiment of comments on both successful and unsuccessful campaigns were typically more positive and neutral than negative**.** Both researchers of crowdfunding and group behavior as well as people interested in starting their own campaigns can benefit from such tools as they can utilize these visualizations to make better sense of the data. Because of this emerging domain, the visualizations explored in this paper are just the beginning of what can be an ever-increasing domain of research and analysis for this growing field. The project can be found on https://www.github.com/alexchao56 under the ClusteringCrowdfunding repository.

Author Keywords
Crowdfunding; text mining; visualization; clustering

INTRODUCTION

Crowdfunding is the practice of using small amounts of capital from a large number of individuals to fund a project or venture typically through the Internet. Crowdfunding makes use of the easy accessibility of vast networks of friends, family and colleagues through social media websites like Facebook, Twitter and LinkedIn to get the word out about a new business or campaign and attract investors [1]. Campaigns can range anywhere from technology, business, nonprofit, political, charity, commercial, or financing for a startup. With the rise of online platforms such as Indiegogo and Kickstarter allowing people to easily create campaigns, crowdfunding has emerged as a particular area ripe for research. According to reported numbers from Kickstarter, only 44% of campaigns have reached their funding goals [2]. In the dataset for this paper, the proportion of successful campaigns ranges anywhere from 5% to 10% depending on the category. Depending on the platform, some campaigns may be structured with an "all-or-nothing" funding model meaning that a campaign has to be fully-funded before the project founder receives any of the money. Kickstarter has adopted this framework, emphasizing that it creates less risk for all parties and motivates people to tell others about the campaign they want to see funded. Additionally, Kickstarter focuses on only featuring "creative projects" [3]. Indiegogo, which the following paper is based on, adopts a more flexible range of the types of projects they feature and does not have an "all-or-nothing" funding model allowing for people to raise as much as the crowd is willing to pay. Thus, depending on people's expectations of how much they can raise and the creativity of the project itself, people may opt to go for one platform over the other.

RELATED WORK

In the current literature, several researchers have sought to predict the success/failure of crowdfunding campaigns, trying to identify which features have been characteristic of each. For instance, using predictors that utilize time series of money pledges to classify campaigns as probable success or failure, Etter, Grossglauser, and Thiran's in their "Launch Hard or Go Home!" theorized that users whose campaigns are failing might want to increase visibility through social media [4]. What they found, using k-nearest neighbors, Markov chains, and support vector machine classifiers, was that failed campaigns have much higher goals on average, but also interestingly a longer duration.

In a paper studying the language and word choices of campaigns, Mitra, Gilbert in "The Language that Gets People to Give: Phrases that Predict Success on Kickstarter" study a corpus of 45000 crowdfunded projects, analyzing 9M phrases and 59 other variables commonly present on crowdfunding sites. What they found was that the language used in the project has surprising predictive power—accounting for 58.56% of the variance around successful funding [5].

While these and other papers have provided recommendations and insights into what goes into a good campaign, few visualizations exist that actually explore the actual textual data of the campaigns.

METHOD

Thus in order to tackle the task of visualization, I chose to use the R programming language for its many text mining and plotting libraries. I worked with over 1 TB of public scraped data from Indiegogo campaign pages dating as far back as five years ago. The data was segmented into particular categories like business/tech in one, community in another, small businesses in another, and non-tech/business. Because the scraped data were saved in into a database and outputted as tab-separated files, they all had the same underlying structure. Thus, processing work and ultimately the visualizations that would be done for one of the categories could be used for the others.

Each campaign contained a unique campaign ID as well as fields for campaign titles, campaign descriptions, perk titles, perk descriptions, the URL for the page, price information for each perk, how much money they raised, and a campaign’s goal. Using these fields it was straightforward to create a differential column that indicated how far away or how much extra the campaign raised compared to their goal. And depending on whether or not that differential was positive or negative, one could tell whether that campaign was a success or not.

With these features, I was able to build a basic binary classifier to predict success/failure. By taking a small subset of the data and splitting it off into training sets and test sets, I was able to use different machine learning techniques to predict on success and failure. Using an ensemble of methods like support vector machines, general linear models, maximum entropy, boosting, and random forests, I was able to get prediction accuracy ranging from 55% to 60%. Since work had been done already by others in prediction, I did not seek to improve upon this classifier, but it could definitely be done as discussed in the section for future work, especially since it is a matter of optimizing feature selection. However, what I did do was try to use the k-Means algorithm to build rudimentary clusters of the data. The plot for that can be found in Figure 1.

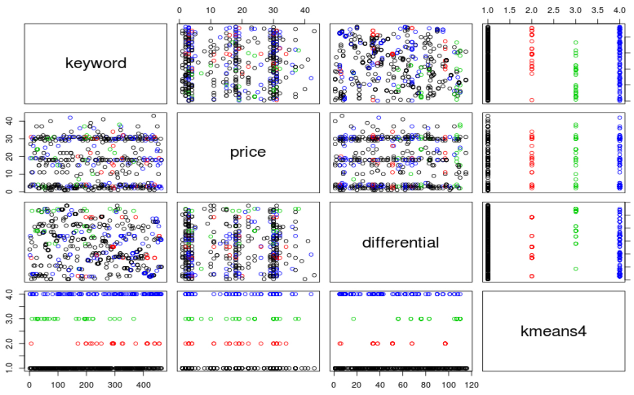


Figure 1: K means plot illustrating the different clusters within the Community campaigns dataset.

In order to select the best values for k, I also created a plot that iterated from 1 to 100 and essentially tested those values of k within the k means algorithm. With these values, I looked at the within-cluster sum of squares. As k increases, the within-cluster sum of squares will decrease, however k being too large loses the effectiveness of creating meaningful clusters. Thus plotting to see where the jagged edges or breakpoints are in the within-cluster sum of squares allows for us to better determine the value of k. This technique is informally known as the elbow method because it seeks to show breakpoints in the cost plot where we should stop adding clusters. One can see an example of this cost plot in Figure 2.

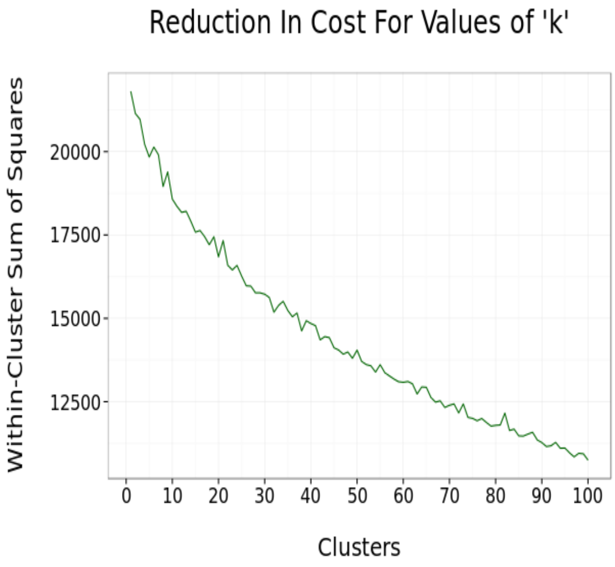
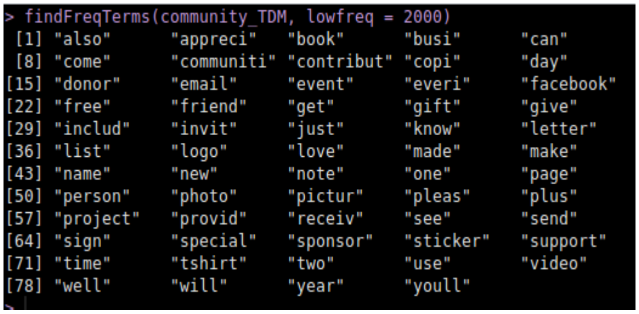


Figure 2: Cost plot for the elbow method to find k for the k-means algorithm to identify clusters.

After some initial cleaning of the data, I began to build the visualizations. Using the ‘tm’ and ‘plyr’ packages in R, I created a stemmed document corpus that removed punctuation, numbers, and stop words from my text of interest. With this corpus, I was able to create a Term Document Matrix data structure that captured each unique word across all campaigns and its frequency. I repeated this process for all permutations of successful and unsuccessful campaign descriptions, titles, and user comments for all four categories of crowdfunding data I possessed.

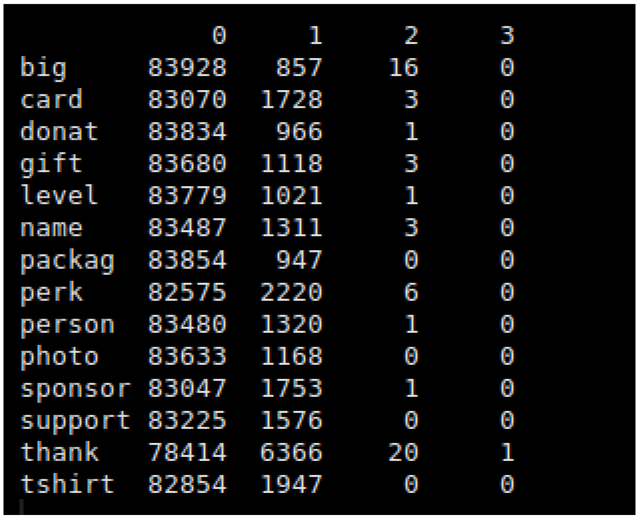
With the Term Document Matrix, I could create specific queries like what were the most frequent terms that appeared in at least 2000 of the campaigns in a category? For instance in Figure 3 one can see the output of this type of query for the ‘community’ dataset.



```
> findFreqTerms(community_TDM, lowfreq = 2000)
 [1] "also"      "appreci"   "book"      "busi"      "can"
 [8] "come"      "communiti" "contribut" "copi"      "day"
[15] "donor"     "email"     "event"     "everi"     "facebook"
[22] "free"      "friend"    "get"       "gift"      "give"
[29] "includ"    "invit"     "just"      "know"      "letter"
[36] "list"      "logo"      "love"      "made"      "make"
[43] "name"      "new"       "note"      "one"       "page"
[50] "person"    "photo"     "pictur"    "pleas"     "plus"
[57] "project"   "provid"    "receiv"    "see"       "send"
[64] "sign"      "special"   "sponsor"   "sticker"   "support"
[71] "time"      "tshirt"    "two"       "use"       "video"
[78] "well"      "will"      "year"      "youll"
```

Figure 3: Most frequent words that show up in at least 2000 campaigns descriptions in the Community category.

For the term document matrix itself, Figure 4 shows the unique words and frequency across all the campaigns in the Community category.



```
             0     1     2     3
big      83928   857    16     0
card     83070  1728     3     0
donat    83834   966     1     0
gift     83680  1118     3     0
level    83779  1021     1     0
name     83487  1311     3     0
packag   83854   947     0     0
perk     82575  2220     6     0
person   83480  1320     1     0
photo    83633  1168     0     0
sponsor  83047  1753     1     0
support  83225  1576     0     0
thank    78414  6366    20     1
tshirt   82854  1947     0     0
```

Figure 4: Table that shows the words and their corresponding frequencies.

Essentially translating this table into a more graphical form, I used the ‘ggplot2’ library to build up the Word Frequency Matrix plot that used three dimensions. The x-axis would be campaigns (so each vertical column or sliver could be interpreted as one campaign), the y-axis would be the unique words, and color would be the third dimension with darker shades indicating a higher frequency and lighter ones, a lower frequency. An example of this plot can be seen in Figure 5.

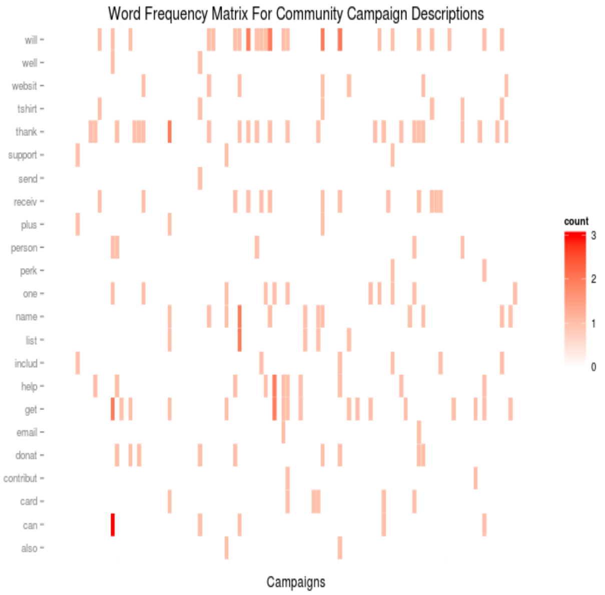


Figure 5: Word Frequency Matrix for the descriptions of campaigns within the Community dataset.

Additionally, using the same Term Document Matrix data structure, I was able to create a dendogram plot (essentially an upside-down tree), of terms that expressed the hierarchical relationship among the words. With this plot, one could see terms that would be grouped together as well as where they fell within the hierarchy. This can be found in Figure 6.

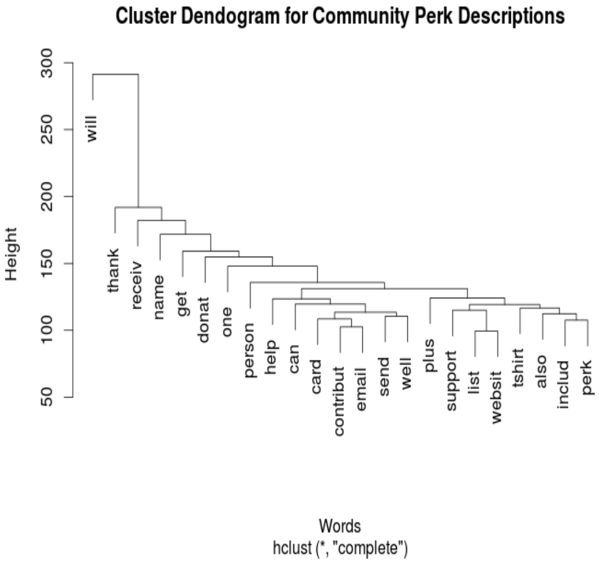


Figure 6: Cluster Dendogram for the descriptions of campaigns within the Community dataset.

Another dataset, I had access to and sought to incorporate was a corpus of user comments for each campaign. Because user comments can signal the status of a particular campaign, it is the ideal text to do sentiment analysis on.

For the comment data set, each row contained the campaign ID, the comment itself, the date the page was scraped, a relative date for when the comment was posted ("posted an hour ago", "almost a year ago", etc.), as well as user information.

In order to calculate sentiment for a particular comment, I wrote my own sentiment scorer that made use of a dictionary of positive and negative words created by Minqing Hu and Bing Liu [6]. My algorithm for scoring was a fairly simple one: sum the counts of all positive words minus the sum of all the negative words in a particular string. Thus words can have a score as negative or as positive as the number of words in the string. Once I was able to determine the sentiment score for a particular string, I then created categories or buckets for scores to fall into. So for scores equal to 0, they received a sentiment value of "neutral." If a score was less than 0 but greater than -4, then it was considered "negative." Anything lower than -4, would be considered "very negative". Scores greater than 0 but less than +4 would receive a sentiment value of "positive" and greater than +4 would be "very positive". With all of this in place, I was able to create a data vector of the sentiment score and its corresponding sentiment category. One could see this binning and the overall distribution in Figure 7.

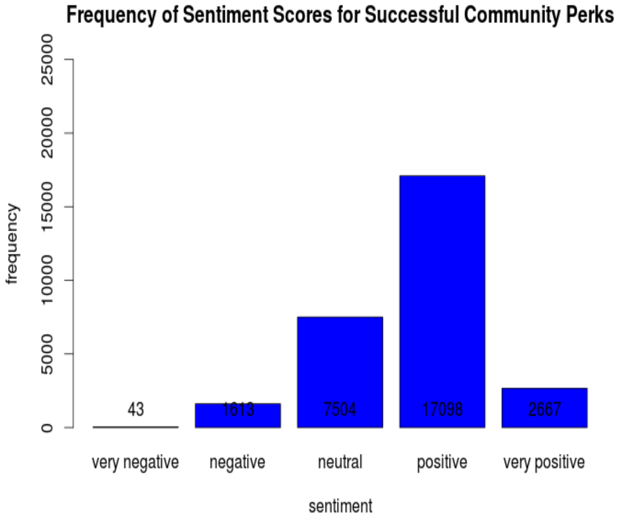


Figure 7: A histogram of the frequency of sentiment scores and the categories they fell into for comments left on successful community campaigns.

Now that I had which sentiment category all the comments classified to, I was able to utilize the timestamp for each comment to create a visualization to show the sentiment of comments over time. One of the tricky aspects of this portion was that the comment date was not in a regular date format. Instead they used relative time such as "posted three days ago" or "posted 12 hours ago" to say when a user entered that comment. In order to create a visualization that showed how the sentiment of the comments evolved over time, I need a way to create a mapping between the text timestamp found on the comment section of campaign pages and an actual comparable time stamp. After some searching, I found that an online tool built in Clojure called Duckling could read natural language about time and create an actual timestamp [7]. Thus with this, I was able to create date strings that could be compared with one another and plotted against. The final output of this was a mosaic plot that captured the number of campaigns that had a particular sentiment value during a particular time frame. An example of this plot can be found in Figure 8.

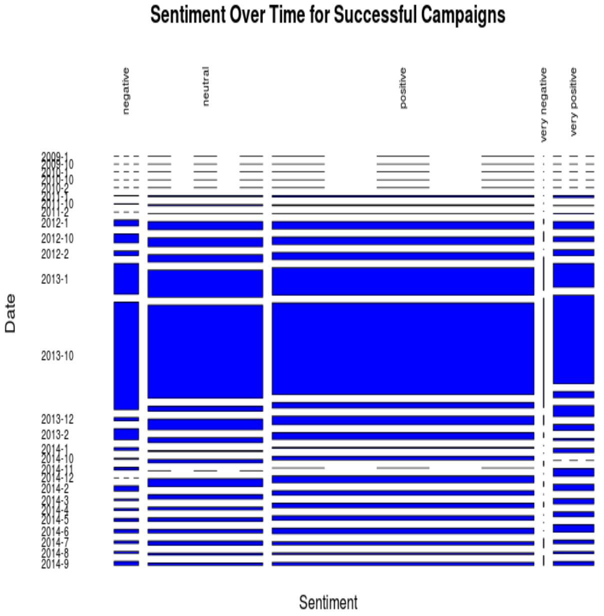


Figure 8: Mosaic plot of sentiment values over time.

Finally, to examine the actual text of comments, I made use of the 'wordcloud' package in R to take in a corpus document (campaigns and their associated comments) and created a word cloud that captured the counts of the most frequent words using redundant encodings of size and color. I also utilized this visualization method to graphically view the titles and descriptions in addition to the comments. One can see an example of a word cloud for unsuccessful comments left on community campaigns in Figure 9.

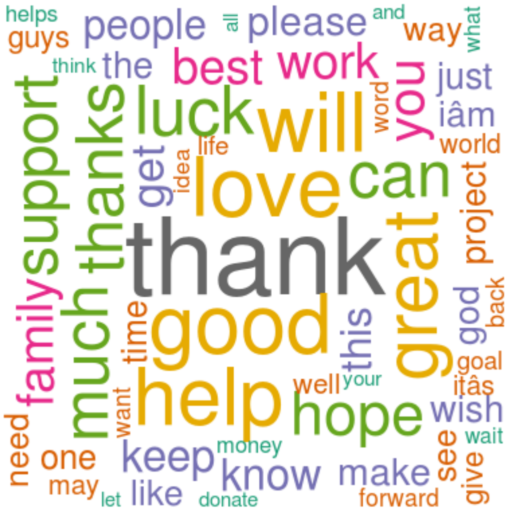


Figure 9: Word cloud for unsuccessful comments left on Community campaigns.

RESULTS

One of the immediate things to note when creating these visualizations is that the size of the data can become a major stumbling block. There are almost one million points of data for a particular category (some categories having more than others) so creating data structures like the Term Document Matrix and even performing operations to munge the data can take a significant amount of time. One way to handle this that I used was to subset the data into reasonable but still sizeable chunks. This way plots can be created in a reasonable time (about one minute for each plot) and more time can be spent actually exploring what the data is telling the reader. Because I made use of several R libraries whose implementations are hidden to me, there may be some ways to optimize the creation of some of the pertinent data structures however to avoid re-inventing the wheel and spend more time visualizing the data, libraries were a better choice.

In order to begin to analyze the results, one must first pick a category of interest. While the nonTech/Business and small business campaigns do have some insights to be gained, because of the increased presence of NA values within these two datasets, it is not as informative to compare between those two. However, where there are much more complete datasets is in the business/technology campaigns as well as the community focused campaigns. Thus comparing within those two categories would be the most informative. Since some of the plots for Community campaigns have been shown, the following will mostly look at business and technology.

For business/tech categories, one can first look at the dendogram for successful campaign descriptions vs the dendogram for unsuccessful campaign descriptions.

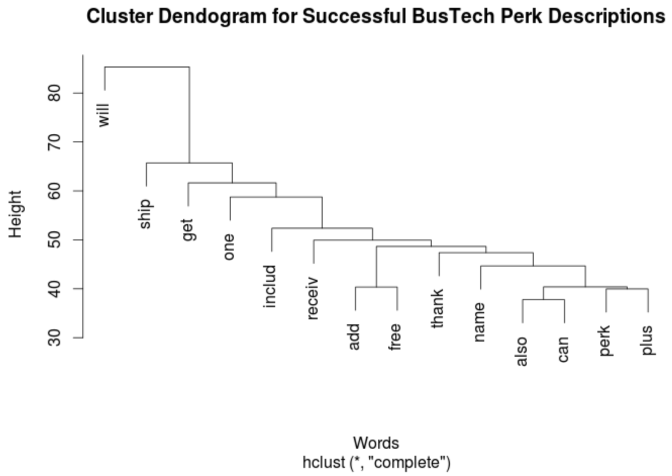
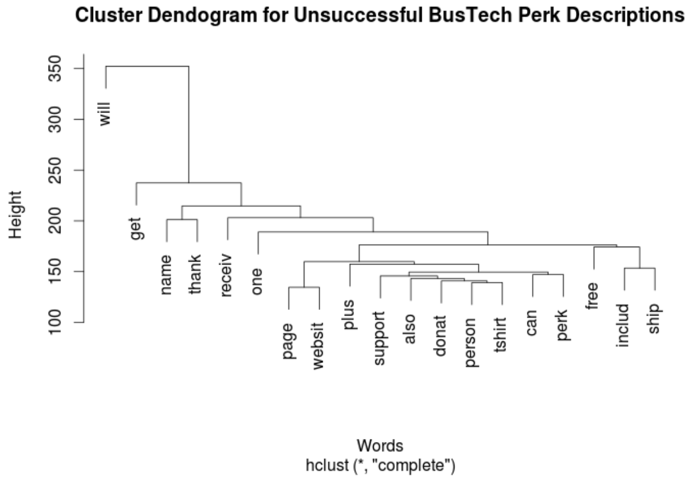


Figure 10: On the top, the cluster dendogram for successful business/tech campaigns. On the bottom, the dendogram for unsuccessful business/tech campaigns.



One can see from this that both of these dendograms at the highest node have the same word “will” appearing as the most frequent word. From there we begin to see differences, ultimately seeing more of a diversity in unsuccessful campaigns with larger families or clusters than in successful campaigns.

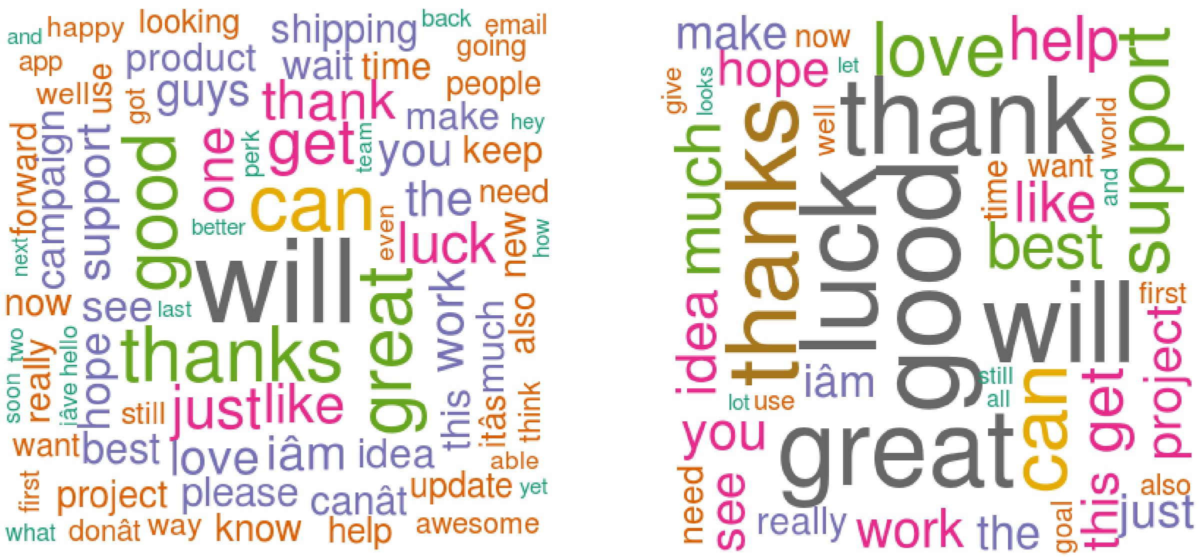


Figure 11: To the left, comments left on successful campaigns. To the right, comments left on unsuccessful campaigns.

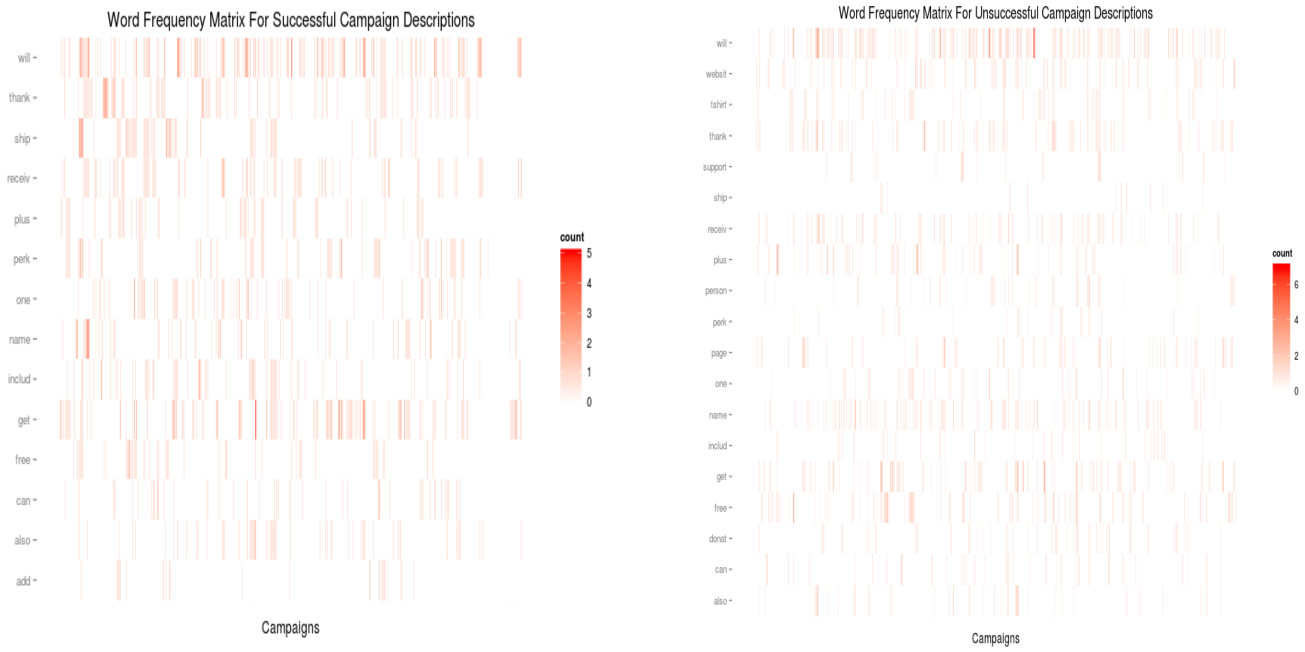


Figure 12: Word Frequency Matrix for Successful Campaigns on the left and unsuccessful campaigns on the right

In Figure 11, one can see that both successful and unsuccessful campaigns in the business/tech category share similar words. It appears as a whole that people use positive words in their comments on campaigns.

And in Figure 12, one can see that, there are some words that appear quite a lot in both successful and unsuccessful campaigns. One thing that could be improved upon in this graph would be to see some of the next most frequent terms because they would be more unique to their campaigns. Also one of the pitfalls with this graph is that the slivers that should represent a single campaign are stacked and thus could be seen as having a higher frequency count. In order to avoid this, one would have to look at fewer campaigns so that the slivers would be more easily defined.

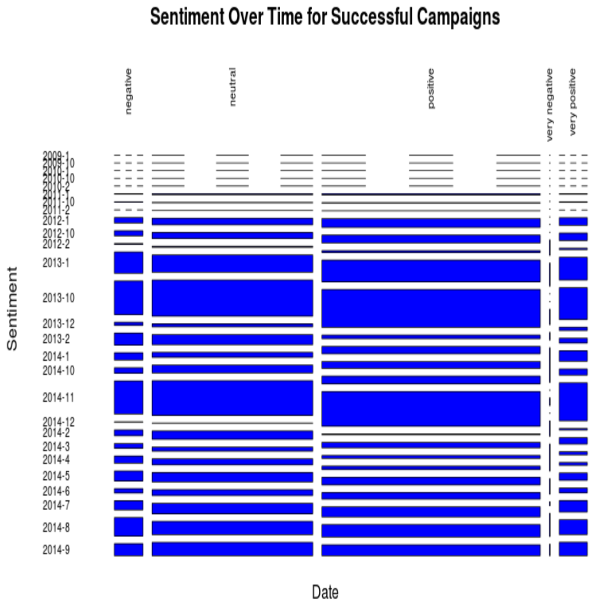
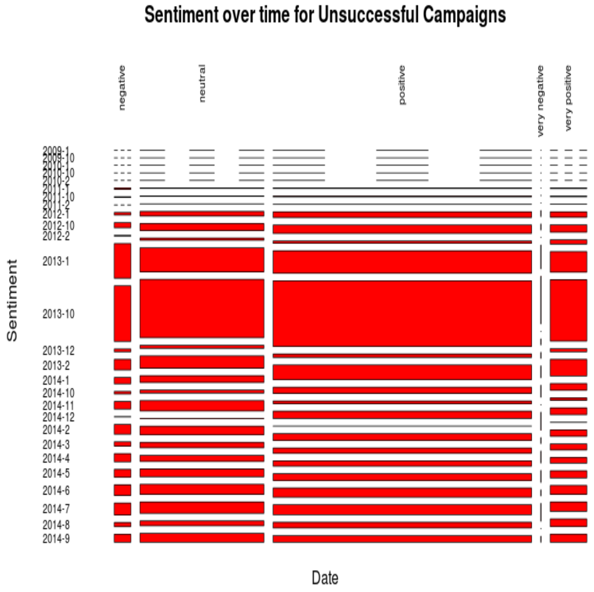


Figure 13: On the top, sentiment over time for successful campaigns. On the bottom, sentiment over time for unsuccessful campaigns.



One of the other interesting visualizations to look at and compare against is how sentiment of comments evolved over time as seen in Figure 13. A particularly fascinating and unexpected finding is that for unsuccessful campaigns, there were more positive sentiment comments for unsuccessful campaigns in a given time frame than for successful campaigns in the same time frame. A result that is not as surprising is that there are more neutral and positive comments on both successful and unsuccessful campaigns than negative ones. In fact, there are very few negative comments at all across all categories. This is probably because that very few people will go to a crowdfunding campaign and actively post negative comment. Most will try to be supportive of someone's attempt to raise funds and leave comments accordingly.

One of the pitfalls of this plot is that some of the times are not perfectly aligned which can cause confusion. A way to handle this would be to color code each of the times so that there would be consistent way to visually see an entire campaign across all the categories of sentiment. Additionally it might make sense to swap the axis as the labels would indicate because people usually interpret time going from left to right. There are several ways that a plot showcasing sentiment over time can be made; this being one of them.

DISUCSSION

Crowdfunding is a field that is only going to continue to grow with time and the ability to analyze the data and make actionable solutions does not just benefit sites like Indiegogo or Kickstarter but also to researchers studying group dynamics, pricing, and fundraisers as well as people looking to start campaigns themselves.

From the visualizations presented and the accompanying analysis that can be made, one can begin to see that there are many actionable steps that can be taken in light of all this data and their plots. For one, people who are interested in creating their own campaigns can look at the words that show up frequently in successful campaigns vs the words that show up in unsuccessful ones and make sure to prioritize using those terms that are in the successful campaigns but not in the unsuccessful ones. By doing this type of analysis across all the campaigns and across all categories, one can build a dictionary of terms that are characteristic of campaigns that have met their funding goal. And since there are costs both monetary and emotional that go into crowdfunding, one should be able to have the tools necessary to create a campaign that has the highest probability for success.

In the area of visualization, one aspect that the work in this paper highlighted was how difficult it is to visualize really large datasets as well as really complex ones with many features. Not only is runtime affected by it and memory constraints can prevent a plot from even loading, but also there involves so much cleaning of the noise before one can actually do some meaningful analysis of the data.

FUTURE WORK

As mentioned previously, one of the ways that this work can improve is in the classifiers that are looking to predict the success and failures of campaigns. By incorporating some of the calculated values like sentiment score, sentiment category, and possibly even the relative date, one could build a more accurate classifier and then from there better cluster campaigns based on their features.

One of the things that Professor Andreea Gorbattai, the provider of this dataset, suggested that she is particularly interested in is how gender affects the language of campaigns. Her hypothesis is that women are in general better at raising funds than men. This phenomenon may be reflected in their language in how they word perks, the description, and even titles of campaigns. In order for this to be done, there would need to be another column for each of the datasets that signals the gender of the person posting or at the very least the first name of the organizer. By then crossing that name against a database from the Social Security Administration that has a list of first names and their corresponding genders, one can build a classifier to predict given some text whether or not it is a campaign organized by a male or female. From there interesting visualizations can come about.

And lastly if I had more time myself, I would have attempted to create a more interactive tool that will allow a user to easily identify a particular campaign given a visualization. Supporting selection is important especially for people interested in starting their own campaigns because they would like to see what particular and specific campaigns were plotted in some of the several plots I displayed. Also an interactive k-means algorithm in which users could select the particular features that they would like to cluster around would be a great tool as well for more of the exploratory data analysis.

With a dataset such as this, there really are so many different ways to further extend this research. From the work presented in this paper, I hope to have done some work to advance the area of visualization for this particular data domain

ACKNOWLEDGMENTS

Thank you to Maneesh Agrawala and Jessica Hullman in the CS department for teaching CS294-10 Visualization and providing the opportunity to pursue this project. The techniques, methodologies, best practices, and philosophy about data visualization exposed me to different ways of thinking about how to tell a story with the data. They will certainly carry with me for the rest of my career. Also, thank you to Andreea Gorbattai at the Haas School of Business for taking me on as a researcher for this project and providing the dataset and mentorship in overseeing my work. Without you, this project would not have been possible.

REFERENCES

1. Crowdfunding
http://www.investopedia.com/terms/c/crowdfunding.asp

2.Kickstarter Basics - Frequently Asked Questions
https://www.kickstarter.com/help/faq/kickstarter%20basics

3.Kickstarter vs Indiegogo - Choosing your Crowdfunding Platform
http://crowdfundingdojo.com/articles/kickstarter-vs-indiegogo-choosing-your-crowdfunding-platform

4. Etter, Grossglauser, Thiran. Launch Hard or Go Home! *École Polytechnique Fédérale de Lausanne.*

5.Mitra, Gilbert.“The Language that Gets People to Give: Phrases that Predict Success on Kickstarter”.

6. Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

7. Duckling

http://duckling-lib.org/