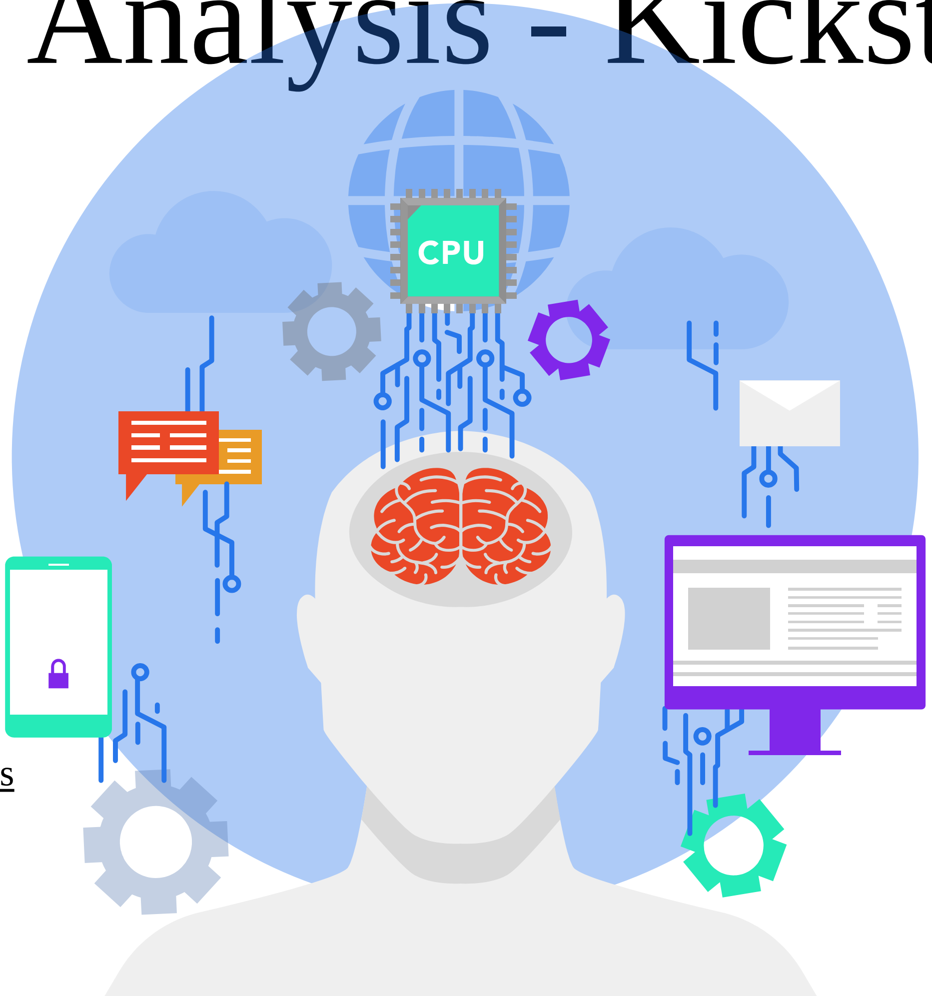
Text Analysis - Kickstarter



Kanishk Gupta

Executive Summary

01 Dataset

Kickstarter dataset with 160,007 rows and 19 features.

02 Features Generated - Dictionary

Dictionary1 – Personal pronoun and words related to the project doer.
Dictionary2 – Pleading words, Example: Help, Need

03 Features Generated - Lexicon

Using AFINN Method for sentiment variable generation

04 Features Generated - Roberta

Roberta is off the shelf library on Hugging face and it is used to classify dataset based on pre-trained dataset of Roberta.

05 Classification

Classifying the dataset into plead or no plead.
Whether the user is asking for money directly for the kickstarter project.

06 Models Used

Decision Tree
Random Forest
SVM
KNN

07 Model Score

Decision Tree – 75%
Random Forest – 78%
SVM – 69%
KNN – 63%

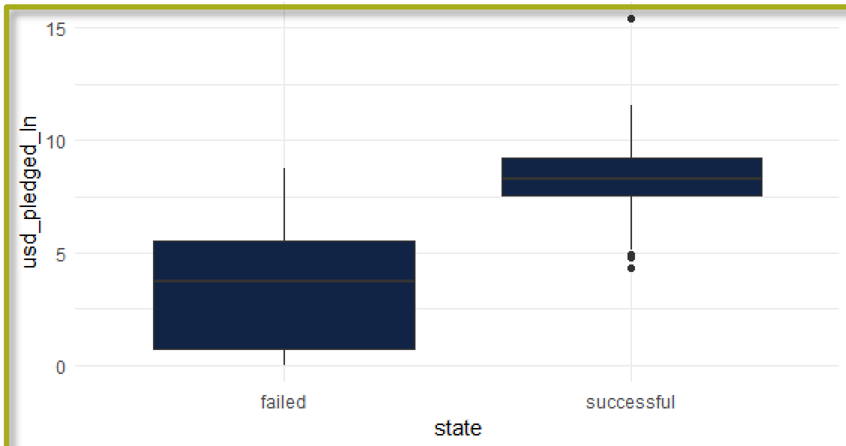
08 Analysis

- Using less pleading words leads to success in campaign
- Using personal pronouns in the description results in low campaign success
- More words in the description, less chances of success in the campaign.
- Less the goal, more chances to achieve campaign success.
- If a campaign has positive description text, USD pledged and backers of the project will be higher.
- Therefore, if the date difference is high, there are more chances of campaign success and getting more money for the project.

Executive Summary - Findings

Classification Prediction	Not using pleading words is highly significant with campaign success. Therefore, using less pleading words leads to success in campaign
Dictionary – Personal Pronouns	Dictionary of personal pronouns and similar words is negatively correlated with campaign success and is highly significant
Dictionary – Pleading Words	Using Pleading words results in less USD Pledged.
Lexicon - AFINN	Lexicon created with off the shelf ‘AFINN Method’ is not significant and is negatively correlated with campaign success.
Word Count	Word count is negatively correlated with campaign success and usd pledged for the campaign. Therefore, more words in the description, less chances of success in the campaign.
Goal	Goal is highly negatively correlated with campaign success. Therefore, less the goal, more chances to achieve campaign success.

Data Exploration



State vs. USD Pledged

In State vs. USD Pledged box plot graph, you can see the proportion of projects managed to raise the desired amount of money.

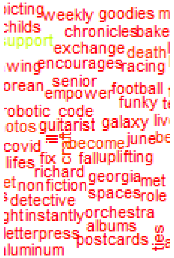
USD Pledged Mean of Successful projects	USD Pledged Mean of Failed projects
~$8 Million	~$4.5 Million

```
> df2 %>% summarize(goal_ln = median(goal_ln))
   goal_ln
1 8.006701
> df2 %>% summarize(usd_pledged_ln = median(usd_pledged_ln))
  usd_pledged_ln
1       7.316636
```

Median Funding Goal and Median of Funds Raised

Median Funding Goal	USD Pledged Mean of Failed projects
$8.006 Million	$7.31 Million

Visualization of the overall distribution of words in the data in a compelling way – Word Cloud*

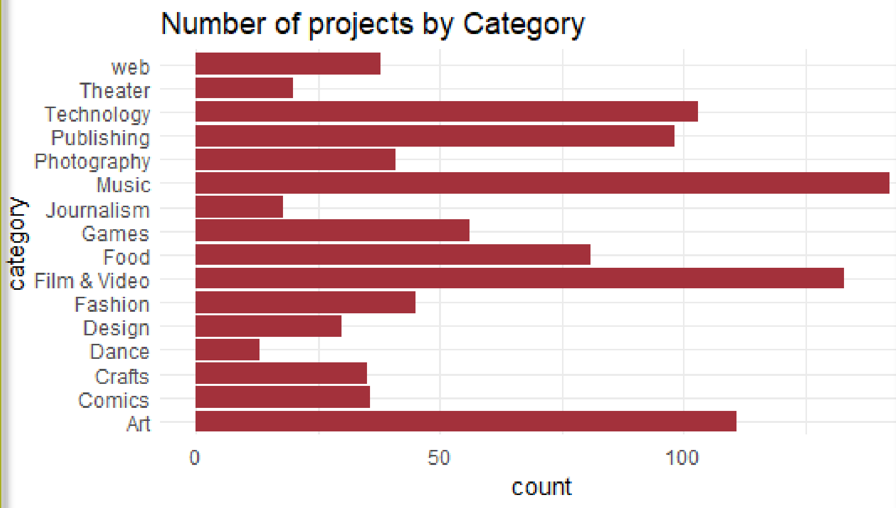


**Due to warnings while running the code, the word cloud couldn't be generated completely*

Data Exploration

```
> length(unique(df2$category))
[1] 16
```

There are 16 different Categories of Projects in the dataset

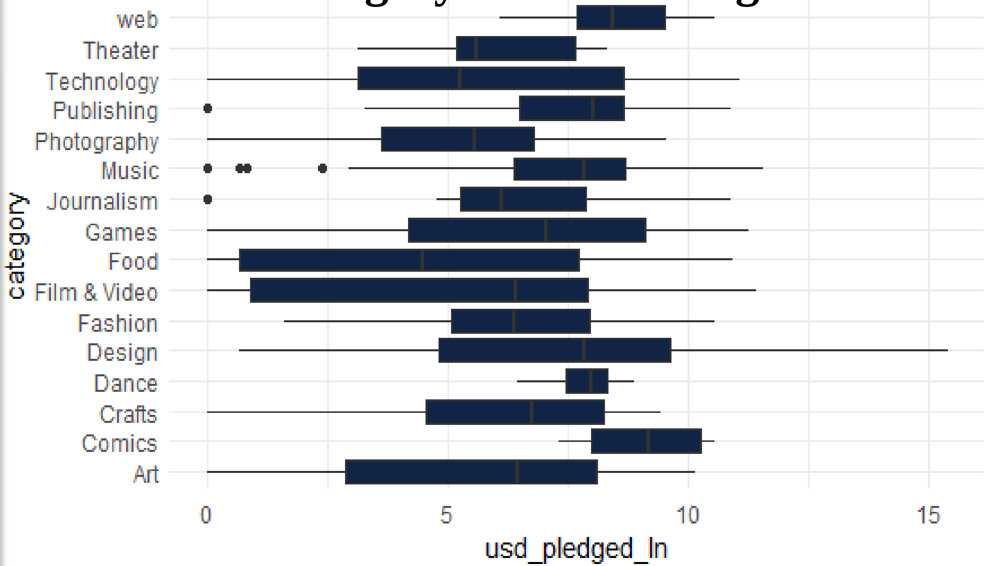


Number of Projects by Category

Music has the highest number of projects listed in the dataset

followed by Film & Video and Art.

Category vs. USD Pledged




```
> df2 %>% group_by(category) %>% summarize(usd_pledged_ln = mean(usd_pledged_ln))
# A tibble: 16 x 2
   category     usd_pledged_ln
   <chr>                 <dbl>
 1 Art                    5.57
 2 Comics                 9.06
 3 Crafts                 6.13
 4 Dance                  7.82
 5 Design                 7.34
 6 Fashion                6.28
 7 Film & Video           5.24
 8 Food                   4.44
 9 Games                  6.59
10 Journalism             6.18
11 Music                  6.93
12 Photography            5.41
13 Publishing             7.16
14 Technology             5.74
15 Theater                6.01
16 web                    8.45
```

Category vs. Mean USD Pledged

Data Exploration

Category vs. Goal Mean

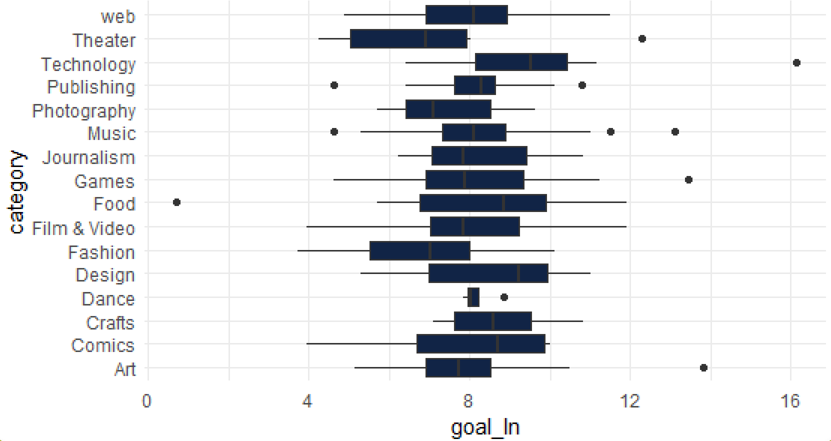
	category	goal_mean
	<chr>	<dbl>
1	Art	7.84
2	Comics	7.84
3	Crafts	8.70
4	Dance	8.17
5	Design	8.49
6	Fashion	6.82
7	Film & Video	8.01
8	Food	8.37
9	Games	8.20
10	Journalism	8.26
11	Music	8.08
12	Photography	7.46
13	Publishing	8.12
14	Technology	9.46
15	Theater	7.17
16	web	8.02

Technology has the highest pledge request in the dataset with a mean of $9.46 Million

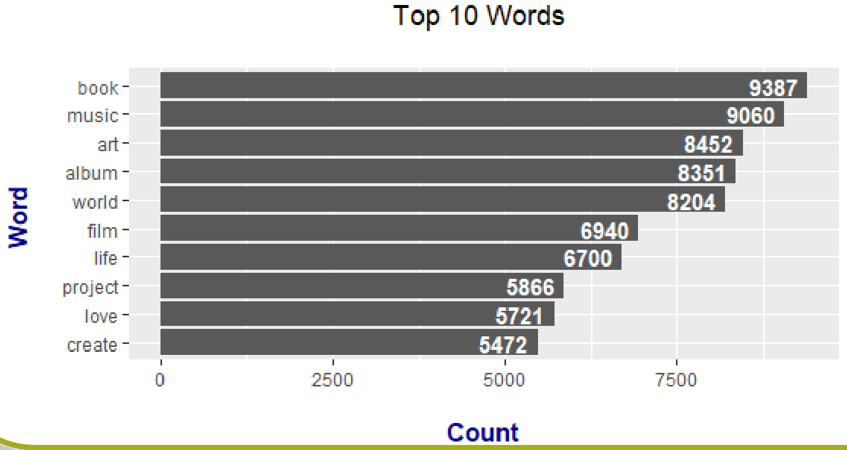
```
> df2 %>% summarize(WC = mean(WC))
      WC
1 19.4303
```

Average Text Contains 19 words

Boxplot of Category vs. Goal



Following are the top 10 most frequent words



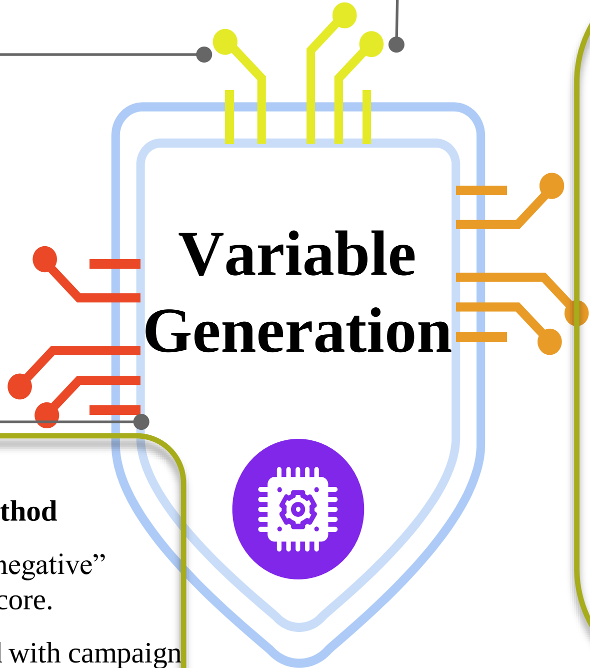
01

Dictionary 1

Words: i, me, myself, my

14.1% of the 160,007 observations contains words from this dictionary

20.9% of the 335 sampled observations contains words from this dictionary



02

Dictionary 2

Words: money, fund, help, love, need, seek, join us, support, appreciate, reinvestment, fundraise, fundraising, donation, be part of community

18% of the 160,007 observations contains words from this dictionary

35.2% of the 335 sampled observations contains words from this dictionary

Reflecting on possible misclassifications by providing examples and suggesting improvement ideas.

Word 'need' can be used in many context:
Example: 1. Creating a vaccine for need of the hour disease
2. Creating a Li-ion battery. This battery is needed in electronic products.
To prevent 'need' from being misclassified, need has to be paired with words such as money, support, help
Example: 'need help', 'need support', 'need money'

03

Lexicon

Off-the-Shelf Lexicon is created using AFINN method

Lexicon classified the dataset into "positive" and "negative" sentiment and assigned each text a sentiment score.

Lexicon is not significant and is negatively correlated with campaign success, whereas it is highly significant with USD Pledged.

Lexicon Example

```
> head(df2$Lexicon)
                                                                                                      example_sentence sentiment_score polarity
1                                                                     We are raising money to fund production of our feature film.               0 negative
2       Help me get this unique photography book that combines my celebrity portraits with CAUSES shared by the CELEBS printed and seen.               3 positive
3            Mark Wallace and I FINALLY have material to record our (EP) original songs album of 6 - 7 songs. As constantly asked of us!               0 negative
4    Help start the first tabletop and gaming bar in California's capital city! We want to build a place where gamers game like grownups.               5 positive
5           This will be the third annual Adam Pehl Photography wall calendar. These make wonderful gifts for the upcoming holiday season.               4 positive
6 I am creating the future of sports entertainment. I have dreamt about being a pro wrestler, now I need YOUR help to make it a reality.               2 positive
```

Variable Generation

Following are the Machine Learning based classifier models that were used to classify text – Pleading (0) and Not Pleading (1)

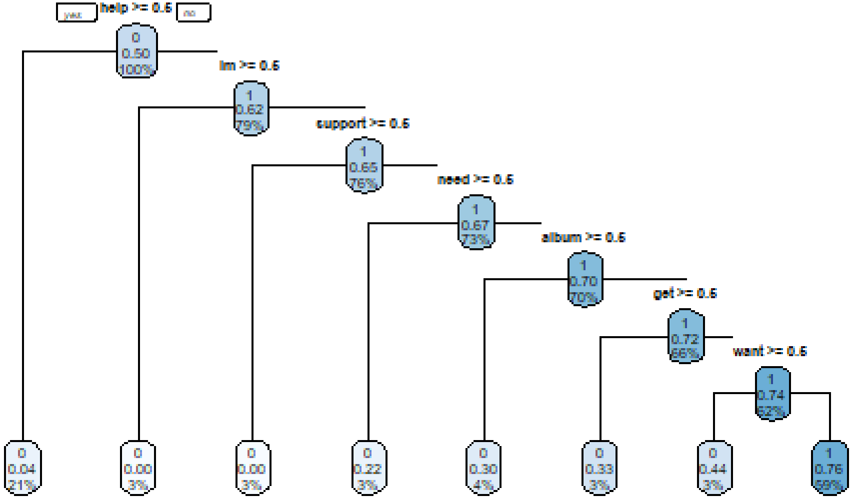
Model Name	Score
Decision Tree	75.7%
Random Forest	76%
SVM	69%
KNN	63%

Predicting on the whole dataset using Decision Tree Model DTM – Unigram and Bigram Weights

Words associated with construct of interest are represented by the decision tree: help, Im, support, need, album, get, want

***All 160,007 observations are evaluated*

Decision Tree



Hugging Face - Roberta

	text	pred	label
0	This project is designed to help protect the e...	1	POSITIVE
1	Help us built a sustainable studio & eliminate...	1	POSITIVE
2	"If I paint something, I don't want to have to...	1	POSITIVE
3	Our free app will allow you pool reservations ...	1	POSITIVE
4	Prohibition themed Gastro Pub and After Dark S...	1	POSITIVE
5	Sean is a naturally talented trumpet player. R...	1	POSITIVE
6	What if we combine food technology, enology ex...	1	POSITIVE
7	A cafe where we can help people reach their he...	1	POSITIVE
8	This project will allow most anyone to view a ...	1	POSITIVE
9	To bring the fantasy and sci-fi world of steam...	1	POSITIVE
10	Delicious Belgian fries and typical Belgian fr...	1	POSITIVE

off-the-shelf language model from Hugging Face – Roberta

Text got classified into "Positive" and "Negative" labels according to the pretrained weights of Roberta

Findings

- Classification column of Plead/No Plead is significant. Therefore, if the campaign did not use pleading words, that campaign has more chances of being successful and raising more money.
- Using dictionary of personal pronouns is negatively correlated. Therefore, for a successful campaign, avoid using personal pronouns.
- The time (number of days) between the start and end dates of each campaign is highly significant. Therefore, if the date difference is high, there are more chances of campaign success and getting more money for the project.
- Goal is negatively significant with campaign success. Therefore, if the goal is higher, less chances of success.
- If a campaign has positive description text, USD pledged and backers of the project will be higher.
- Word Count is negatively correlated. Therefore more words in the description results in less successful campaign and less money pledged.

Refer Figure. 1 in Appendix

Textual analysis of stock market prediction using breaking financial news: The AZFin text system

Authors: Robert P. Schumaker, Hsinchun Chen

The research examines a predictive machine learning approach for financial news articles analysis using several different textual representations: bag of words, noun phrases, and named entities. Through this approach, the authors investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. They applied their analysis to estimate a discrete stock price twenty minutes after a news article was released. Using a support vector machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables, they show that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price (MSE 0.04261), the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return). They further investigated the different textual representations and found that a Proper Noun scheme performs better than the de facto standard of Bag of Words in all three metrics.

Findings

Top Recommendations

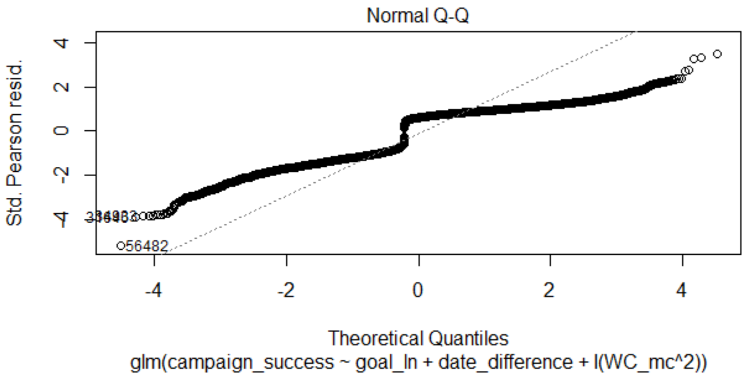
- Do not ask or beg for help in the product description.
- Avoid using personal pronouns in the description.
- Be concise about the description. Do not write big descriptions
- Time allocated to for the campaign online should be sufficiently large to allow more backers to participate
- Goal of the campaign should be reasonable. It should not be too high to scare of potential backers.

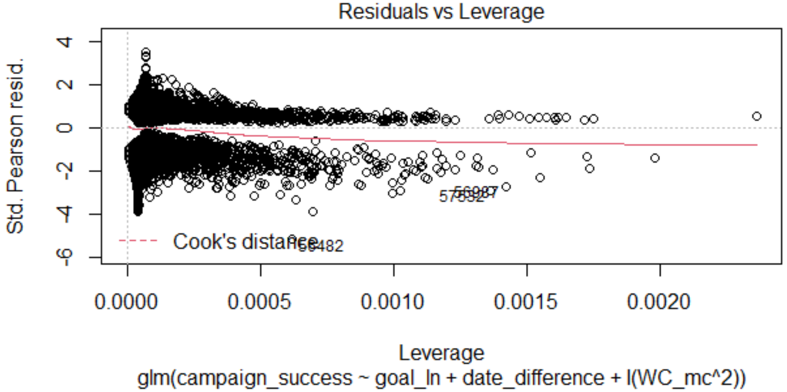
Limitations

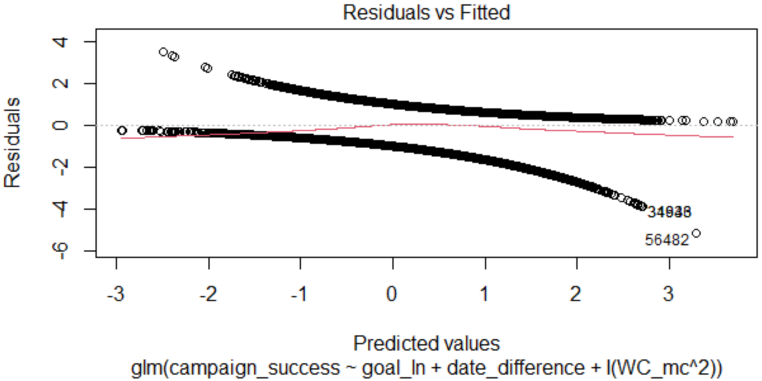
- More computing power is needed to run the word cloud and non-linear regression models
- There was class imbalance. More data is needed to avoid class imbalance. After manual classification, only 165 rows of each class were selected to train the model
- There were absurd data in the text
  Eg: “Wearable Billboards for Goodness' Sakeâ„¢. Worn to unify people while addressing societyâ€™s issues”

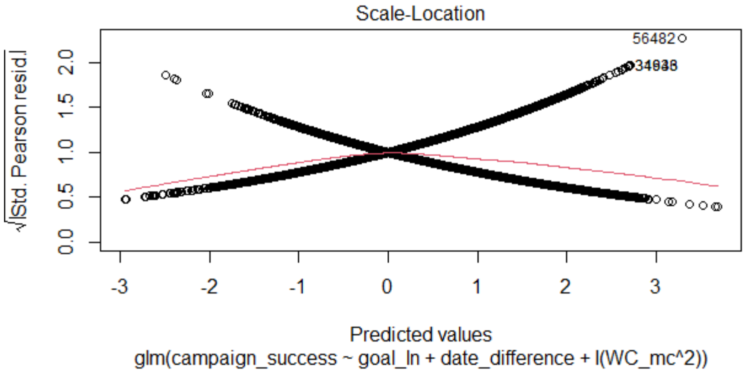
Findings

Non-Linear Effects









Appendix

Regression Results

	Dependent variable:				
	campaign_success logistic (1)	usd_pledged_ln OLS (2)	usd_pledged OLS (3)	backers_count_ln OLS (4)	backers_count Poisson (5)
pleadornoplead1	.094***	.157***	1,545.485	.069***	
	(.018)	(.024)	(838.166)	(.014)	
sentiment_score_full_dataset	-.001	.010***	359.912***	.006***	
	(.002)	(.003)	(100.562)	(.002)	
pred	.320***	.499***	2,275.776	.253***	
	(.029)	(.039)	(1,342.756)	(.023)	
dictionary	.082***	.086***	-1,706.956	.025	-.102***
	(.019)	(.025)	(875.209)	(.015)	(.001)
WC_mc	-.008***	-.015***	-77.558	-.013***	-.007***
	(.001)	(.002)	(54.382)	(.001)	(.00004)
i	-.902***	-1.505***	-2,863.358***	-.822***	-.828***
	(.017)	(.023)	(785.902)	(.013)	(.001)
goal_ln	-.312***	.253***	7,817.049***	.152***	.419***
	(.004)	(.005)	(160.132)	(.003)	(.0001)
date_difference	.001***	.001***	2.088	.001***	.0002***
	(.00004)	(.0001)	(1.865)	(.00003)	(0.00000)
Constant	1.983***	3.392***	-61,832.080***	1.129***	.368***
	(.133)	(.174)	(6,068.159)	(.103)	(.005)
Country Fixed Effects	Yes	Yes	Yes	Yes	Yes
Category Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	160,007	160,007	160,007	160,007	160,007
R2		.129	.033	.148	
Adjusted R2		.129	.033	.148	
Log Likelihood	-92,907.450				-32,453,750.000
Residual Std. Error (df = 159959)		2.947	102,495.300	1.741	
F Statistic (df = 47; 159959)		504.007***	116.703***	591.565***	
Note:	*p<0.05; **p<0.01; ***p<0.001				

Figure 1: Regression Results

Appendix

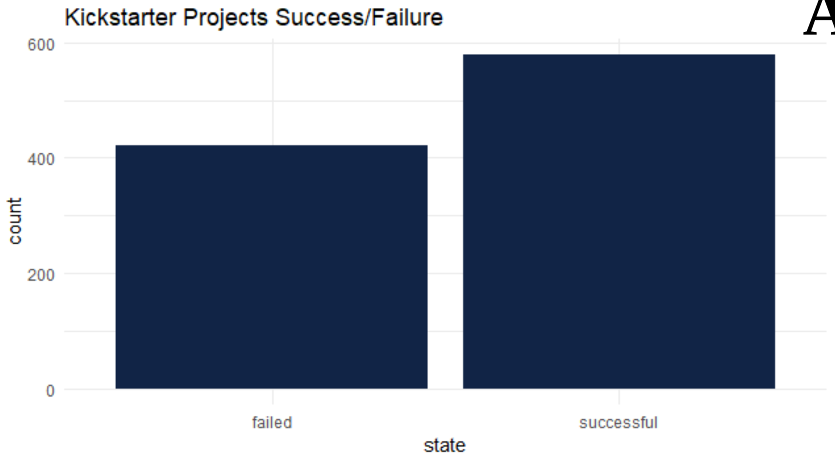


Figure 2: Number of Failed and Successful Campaigns

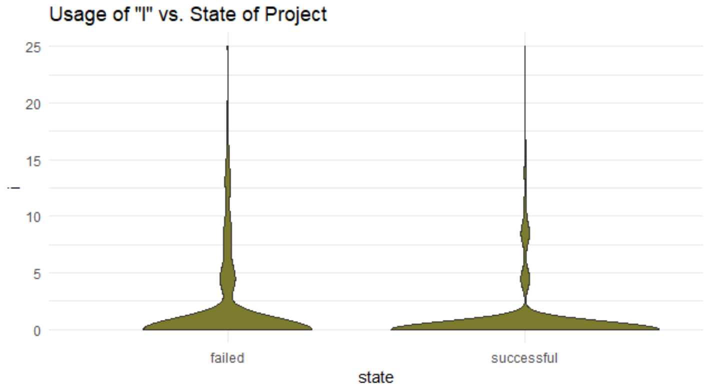


Figure 4: Usage of I and the corresponding campaign outcome

Figure 3: Mean Pledge in each country

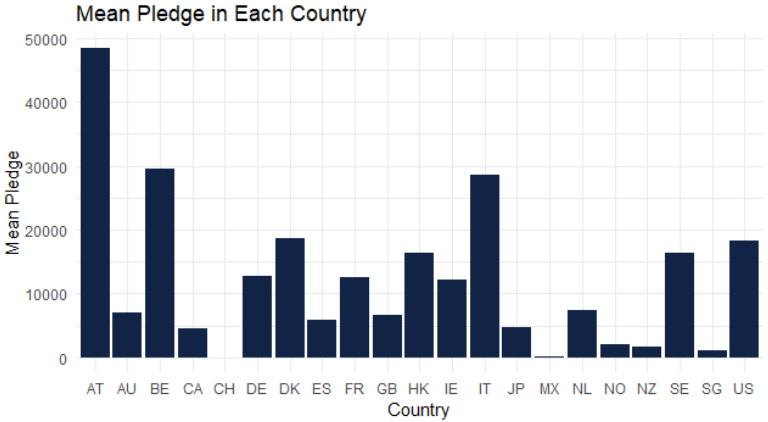
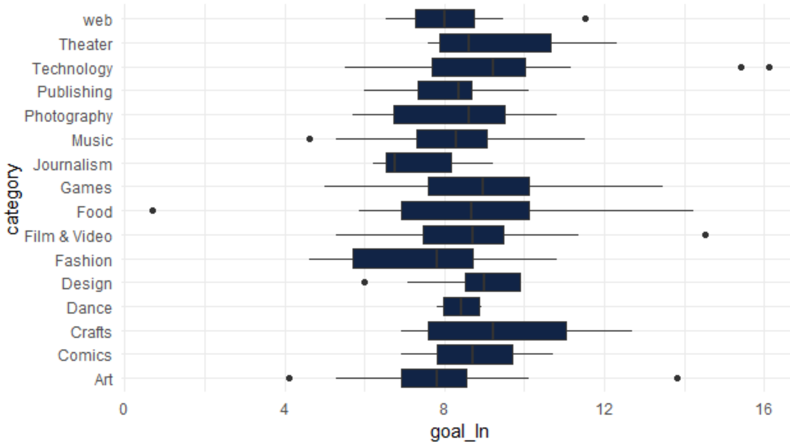


Figure 5: Goal of funding in each category



Appendix

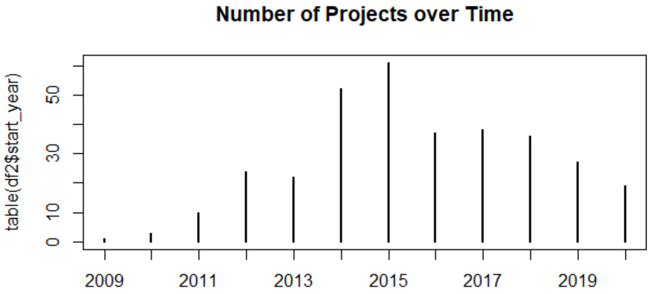
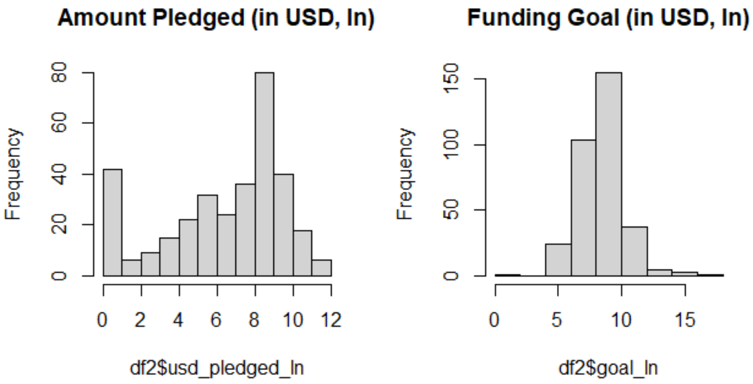


Figure 6: Number of projects over time

Figure 7: Amount Pledged and Funding Goal



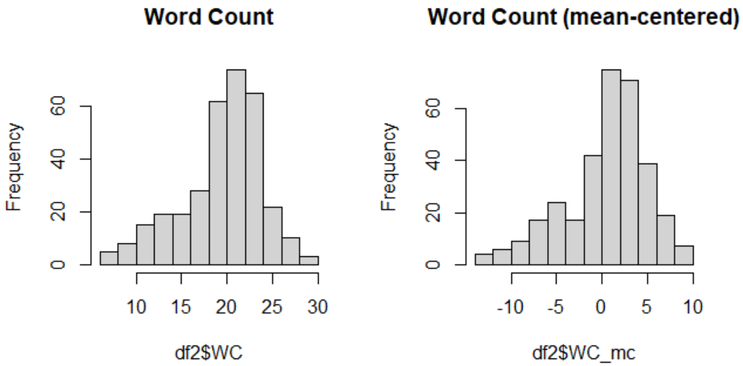


Figure 8: Word Count and Word Count (Mean Centered)