PREDICTING ONLINE SUCCESS FROM TEXT DATA

TABLE OF CONTENT

KICK STARTER PROJECT

KICKSTARTER

EXECUTIVE SUMMARY

KICKSTARTER PROJECT

01.SITUATION

Kickstarter is a popular crowdfunding platform that is used by people to seek support on a variety of campaigns. We are tasked to leverage unstructured project descriptions from past campaigns to predict funding success

02.HOW

Conduct a structured analysis on 160k past campaigns, exploring Three three Natural Language processing methods (Lexicon, Machine learning, Pretrained model)to derive text features which contributions to the funding success will be evaluated using a logistic and linear regression models

- The lexicon used was based on the usage of personal pronouns(I, me, myself, my)
- Four machine learning models (Decision tree, Random forest, KNN, SVM) have been utilized to generate the joy emotion from the project descriptions
- The off shelf AFFIN dictionary and pretrained Roberta models have been leveraged to output positive sentiments
- The logistic and linear regressions have then been run on the newly generated textual features to determine their impact on the campaign success and amount pledged

03.RESULTS

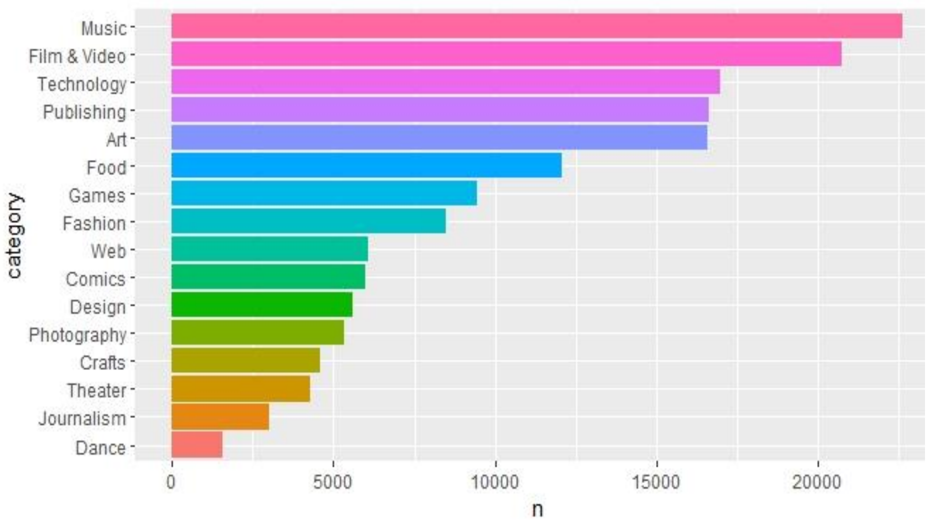
The use of personal pronouns and the joy sentiment derived from the above methods are significantly unfavourable to the success of the campaign

- The positive sentiments derived respectively from AFIN lexicons and Roberta are not significant in explaining their relationship to the campaign success
- The categories are strongly associated with the funding success
- The funding success is relatively significant across the countries

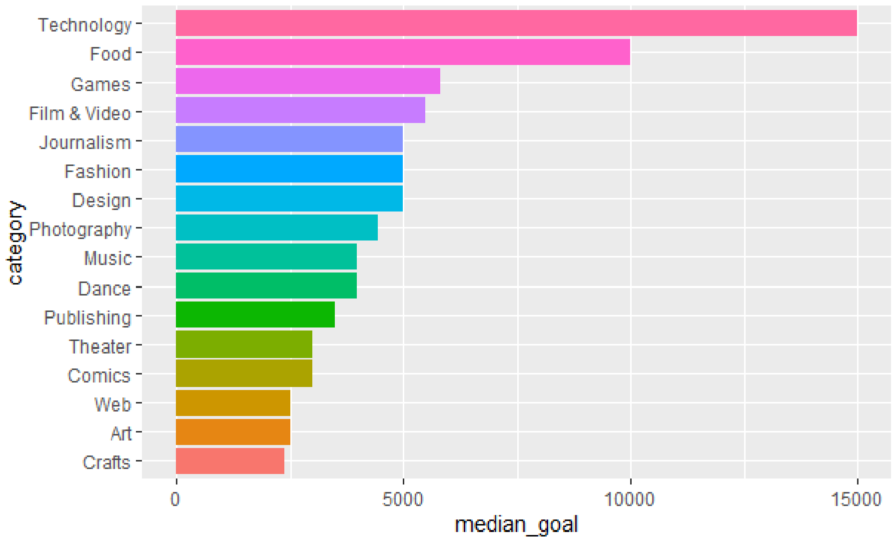
KICKSTARTER

EXPLORATION ANALYSIS

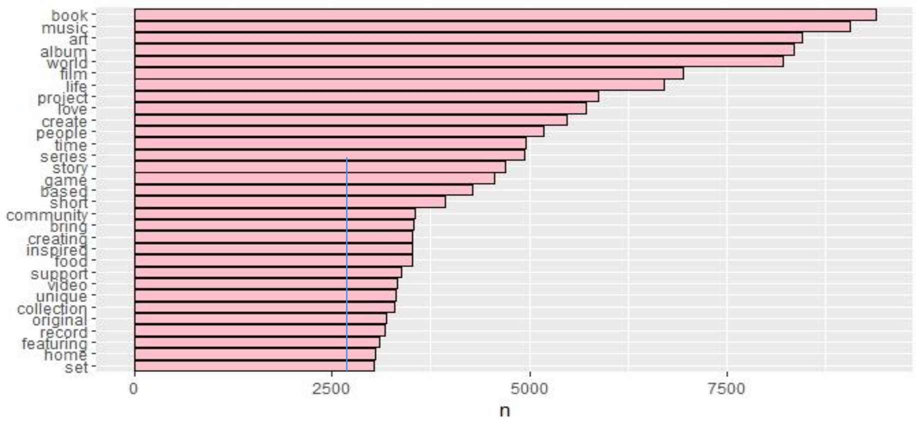
Music. Film&Video, Tech makes the three top categories in number of projects pitched



while Tech, Food and Games are the most goal oriented categories



Top 30 most popular words includes book, music, art, album, love, life



Word cloud associated with the campaign success

Additional insights

- 58.5% of projects managed to reach their target
- The median fund raised is $1,645
- The median goal is $5,000
- The correlation between goal and the pledged amount is very weak (0.139)
- **The word cloud displays words with the joy connotation. Therefore, it might be interesting to hypothesize this pattern as potentially explaining the fund meeting. Moreover, we will also looking at how relating the pitch to oneself impact on the success of the campaign.**

FEATURES GENERATION

Dataset 160 K observations and 17 features

Variable Types Categorical, Numerical , and Text variables

Task type: Classification
Target: Generate potential textual features based on hypothesized linguitics patterns and model based sentiments

01 Custom Lexicon

- Personal pronouns is used 14% of the time
- The median amount pledged when using personal pronouns is $242, 8.5 times less than that when not using personal pronouns
- The accuracy of the lexicon is 39%, less than the random chance accuracy.

02 AFFIN Lexicon

Generates the positivity attribute in the text,
Predict 50% of the time the campaign success. This dictionary performs better the custom lexicon

03 Bag of words Models

We sampled 600 observations , labelled them, split and train a decision tree, random Forest, KNN and SVM models to generate joy sentiment features, then select the best model (KNN) to predict across the entire set

04 Roberta

We've generated a positivity based sentiment across the data set.

KICKSTARTER

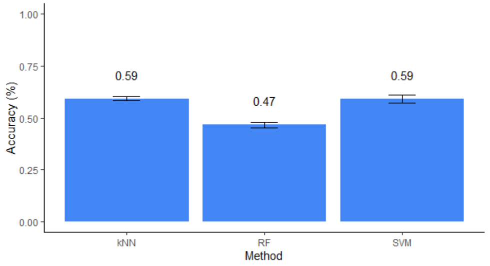
FINDINGS

Objective: Evaluate the significance of the textual features.

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)
WC_mc	-.008*** (.001)	-.015*** (.002)	-105.166 (54.064)	-.013*** (.001)	-.008*** (.00004)
i	-.910*** (.017)	-1.524*** (.022)	-3,266.436*** (779.232)	-.832*** (.013)	-.839*** (.001)
polarity	-.0005 (.012)	.048** (.015)	1,387.433** (531.108)	.031*** (.009)	.063*** (.0004)
Sentiment1	.110* (.046)	.133* (.060)	-1,094.151 (2,068.232)	.111** (.035)	-.060*** (.002)
labelPOSITIVE	.049 (.029)	.032 (.038)	1,027.911 (1,316.047)	.025 (.022)	.041*** (.001)
goal_ln	-.310*** (.004)	.257*** (.005)	7,851.998*** (159.888)	.154*** (.003)	.419*** (.0001)
date_difference	.001*** (.00004)	.001*** (.0001)	2.147 (1.866)	.001*** (.00003)	.0002*** (0.00000)
Constant	2.315*** (.132)	3.951*** (.174)	-59,905.380*** (6,035.731)	1.390*** (.103)	.274*** (.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		.128	.033	.147	
Adjusted R2		.128	.033	.147	
Log Likelihood	-92,977.380				-32,452,947.000
Residual Std. Error (df = 159960)		2.949	102,501.300	1.741	
F Statistic (df = 46; 159960)		509.453***	118.794***	600.633***	
Note:	*p<0.05; **p<0.01; ***p<0.001				

Features	Model/lexicon
i	Custom lexicon
Polarity	AFFIN
Sentiment	KNN
Label	Roberta



1. The "word count" and the emotions derived from the machine learning (ML)models and the custom based lexicon are the only textual features which, we can make inferences with campaign success. In particular, longer word counts and the use of personal pronouns reduce the chance of success on average while the joy emotion plays the opposite.

2. The longer the time to reach the goal, the greater the probability to succeed.

3. The goal is also statistically related to the campaign success, which suggests that those having conservative (affordable) goals have higher chance to meet them.

4. The sentiment derived from Roberta are not statistically significant to be connected to the campaign success or the amount of amount pledged

5. The R2 of the regression models are very low and therefore, the textual features barely explain 3 to 14% of the variability of the capital pledged or backer count

KICKSTARTER

Recommendations

01.Pitch briefly

Talking less allow to structure the message in a very efficient way to cater it to audience so it can quickly grasp it and eventually back up the project

02.Avoid talking about oneself

Drawing attention on oneself gets the audience loose the project goals.

03.Setup realistic expectations

This is about having a realistic budget and a reasonable timeline to raise funds. Keeping in mind the median fund pledged from past data ($1645)

LIMITATIONS

01
Data Labeling
Manually labelling the data by one person introduces a certain amount of bias. It would be more relevant to include diverse views

02
Semantic meaning
The features extracted from the lexicons did not account for context and therefore would be prone to misclassification.

03
Computational Power
Predicting the sentiments using Roberta was computationally demanding

04
Pretrained model
In our case, the feature generated did not have any significant impact on the goal and one may attribute this to the context within which the pretrained model was built

Thank You!

KICKSTARTER