Identification of sustainability-focused campaigns on the kickstarter crowdfunding platform using NLP and ML boosted with swarm intelligence

Data Analysis: part 2

Submitted by: Jossin Antony

Affiliation: THU Ulm

Date: 11.06.2024

Overview

- Introduction
- Extraction of key words
- Attention!
- To Dos

A. Introduction

We continue our analysis with the filtered dataset from part 1. The data set consists of features 'is_environmental' and 'is_social' which are thought to be very essential in the upcoming analyses. However, only 1% of these columns hold values. In this script, we try to populate the rest of the columns with values using NLP analyses.

```
We print 2 random rows of the dataset for preliminary impressions.
```

	campaign_name	blurb	main_category	sub_category	is_environmental	is_social	country
129564	Graen Magazine Winter Issue / Graen rhifyn gaeaf	Adventure culture magazine & website created b...	Journalism	Print	NaN	NaN	GB
117751	Hand knitted designs for Kidz 'n' Cats dolls ~...	An heirloom designer’s knitwear pattern book f...	Crafts	Knitting	NaN	NaN	US

B. Extraction of key words

We try to extract the main keywords which will help to classify the blurbs- description of the project- as environmentally or socially relevant.

Some of the data are manually curated and classified as socially or environmentally relevant. We start with the analysis of this data in the hopes that it might reveal some clues to understand how the data was actually classified, beyond the human notions of what is socially or environmentally relevant.

First we replace all the NaN values with the term 'unspecified'. Next we check how many samples were manually curated.

```
Observation: The Dataset has 1638 rows with an "Yes" or "No" value in "is_environment" and "i
s_social" columns.
Note: Due to manual curation all the selected samples have values in both "is_environmental"
and "is_social" columns.
```

Next we check the proportion of 'yes' and 'No' values.

```
is_environmental
unspecified    137949
No               1602
Yes                36
Name: count, dtype: int64
is_social
unspecified    137949
No               1617
Yes                21
Name: count, dtype: int64
```

Observation: The classes are not well balanced. We see that an overwhelming number of samples are classified 'NO' for social or environmental relevance. Classical classification machine learnings cannot be applied here, because the 'null accuracy' (prediction 'No') is well over 90%.

As an alternate (and easy) approach, we try to find the most important words that appear in the 'blurb' classified as socially/environmentally relevant.

We start with the 'tf-idf' algorithm. The aim is to calculate the mean tf-idf scores of the words that appear in the corpus marked as socially or environmentlly relevant and later attempt to use the appearance of these words to classify uncategorized extracts.

Note:

- We use stemming to find the 'root' form of the words that appear in the corpus. We start the analysis with snowball stemming.
- The stopwords in english (e.g. 'and', 'these') are omitted from the analysis. Similarly, all numbers, symbols etc. are also ignored (e.g: 'covid-19' -> 'covid').
- To increase the amount of training data, the 'campaign_name' is also considered along with 'blurb'.

```
No. of identified top words distinguising environmentally relevant blurbs: 68.

The first column represents the relevant words and the second column gives the mean tf-idf sc
ore
Note: The word are in the stemmed format. e.g "sustain" can mean "sustainability", "sustainin
g", "sustained" etc.

Top words (is_environmental)
-----------------------------
organ         0.104617
sustain       0.095428
friend        0.064725
eco           0.061742
design        0.059673
natur         0.055885
world          0.05404
recycl        0.053087
farm          0.052183
build         0.052123
use            0.04635
produc        0.044987
make          0.043931
provid        0.041628
save          0.041407
compani       0.040932
food          0.039646
small         0.039454
local         0.037731
tea           0.037069
fashion       0.036918
hous          0.036563
healthi       0.036306
mobil         0.035991
communiti     0.035228
men           0.032951
ve            0.032935
befor         0.032935
way           0.032613
tree          0.032129
vegan         0.029253
materi        0.029078
sourc         0.029041
high          0.028883
shirt         0.028846
ethic         0.028003
bring         0.027754
innov         0.026653
brand         0.026165
agricultur    0.024807
better        0.024807
collect       0.024083
work          0.024061
famili        0.022617
europ         0.022384
creat         0.022088
```

```
time          0.022017
project       0.021593
care          0.020975
product       0.020908
anim          0.020666
educ          0.020657
plant         0.020446
america       0.020028
cloth         0.019696
fresh         0.019619
pair          0.019607
qualiti       0.019541
servic        0.019321
year          0.019164
round         0.019164
hand          0.018911
non           0.017863
gmo           0.017863
electr        0.016948
bicycl        0.016948
new           0.016441
environ       0.009889
dtype: Sparse[float64, 0]
```

Now we verify that this approach works! We expect that the words we found as relevant occur multiple times (atleast one time) in the samples manually curated as relevant and do not occur at all if they were manually curated as irrelevant. From this data we calculate the accuracy as the number of correctly classified/ total classified.

We try this approach first on the samples marked as 'environmentally' relevant.

```
Categorization summary
========================
yes_count: is_envt
at least one keyword    33
No keyword               3
Name: count, dtype: int64
accuracy: 0.92
Observation:
Out of the 36 samples available, 33 were classified correctly and 3 incorrectly, giving us an
accuracy of ~0.92. We can also inspect the dataframe in detail, so that we know where the res
ults were false.
```

	campaign_name	blurb	is_environmental	yes_count: is_envt	ranked_words
40	Beluga tent 6-in-1 from Qaou	The first all in one highly eco-friendly tent ...	Yes	3	[eco, friend, recycl]
63	Thé-tis Tea : Plant-based seaweed tea, rich in...	Delicious tea infusion made with seaweed. Heal...	Yes	3	[organ, eco, friend]
108	Chique Addiction	High fashions made from ethical and sustainabl...	Yes	3	[sustain, friend, world]
138	Hearth & Market - Wood Fired Food Truck & Mobi...	A wood fired food truck & mobile farmers marke...	Yes	2	[farm, organ]
193	Rebel Swim - Men's swim shorts, designed with ...	Buy a pair of our beautiful men's swim shorts ...	Yes	1	[design]
235	Ash Apothecary: Small Batch, All-Natural Simpl...	Small-batch simple syrups for bartending, mixo...	Yes	2	[natur, organ]
292	Stitchmill Clothing // The Perfect Henley Shirt	Redefining Henley fashion for women and men. S...	Yes	1	[sustain]
317	Tree Rally	A David and Goliath story about a Sydney commu...	Yes	0	[]
333	Organic agriculture against desertification: t...	A tool for a better farming in semi arid regio...	Yes	2	[organ, farm]
398	Greenr	A company that tracks green behavior and rewar...	Yes	1	[world]
399	Diving Deep	Irrepressible underwater filmmaker Mike deGruy...	Yes	0	[]
641	Join the Blue Revolution	Through education, sustainable agriculture and...	Yes	3	[sustain, build, world]
643	Longwater Community Farm	Growing our own food, caring for animals and t...	Yes	2	[farm, sustain]

	campaign_name	blurb	is_environmental	yes_count: is_envt	ranked_words
738	Stinger Surf	Stinger Surf Co. is an innovative brand that m...	Yes	2	[eco, friend]
752	A NEW EXCITING ELECTRIC BICYCLE BRAND IN NORTH...	Darrvin is an innovative e-bike design and bui...	Yes	3	[design, build, build]
753	Eco Bin the worlds first Eco Friendly Sanitary...	Worlds first Eco friendly Sanitary Bin that is...	Yes	7	[eco, world, eco, friend, world, eco, friend]
885	OS eBike: Open Source Electric Bicycle Design ...	A guide for making a lightweight eBike design ...	Yes	3	[design, build, design]
947	"The Sun Juicer" Ultralight Parabolic Solar C...	A sustainable fuel free, clean energy, 0 emiss...	Yes	1	[sustain]
997	Boulder Denim 3.0: Active jeans for work, play...	Performance denim unlike anything you've worn ...	Yes	1	[sustain]
1061	TEA BAR- Vegan and Eco-Friendly health bar fla...	Organic, non-GMO, Eco-Friendly, Healthy, Super...	Yes	5	[eco, friend, organ, eco, friend]
1204	Sustainable Produce For The Locals, Annually	A self-sustaining farm to feed locals fresh or...	Yes	5	[sustain, sustain, farm, organ, recycl]
1286	Fable: From Farm to Table	Fable aims to be a year-round source for the f...	Yes	1	[farm]
1340	Low Cost Fresh & Finished Organic Food	We are the First Vertical Integrated Organic G...	Yes	2	[organ, organ]
1363	Biobierwinkel	The first online 100% organic beer shop in the...	Yes	2	[organ, organ]
1386	Ellice Ruiz | Girlfriend Tested, Mother Nature...	A sustainable & ethical ready-to-wear collecti...	Yes	2	[natur, sustain]
1389	SJ family farm and ranch organic community garden	The goal of SJ farms is to provide healthy org...	Yes	4	[farm, organ, farm, organ]
1397	Recycle Scrap Paper into Building Material - P...	Everyone loves to recycle, this project	Yes	5	[recycl, build, recycl, recycl, build]

	campaign_name	blurb	is_environmental	yes_count: is_envt	ranked_words
		is abo...			
1406	@SpeedingDonuts +Donutruck	Bringing healthy and organic eating to donut l...	Yes	1	[organ]
1419	Young Scent - Premium Drinking Vinegar & Vineg...	All natural, handcrafted, organic fruit vinega...	Yes	3	[natur, organ, design]
1434	PIVOT | The Spray Bottle Reinvented.	Effortlessly spray at any angle. Use every dro...	Yes	0	[]
1465	Animal-Friendly Footwear Made Using Apples.	Crafted in Europe using sustainable innovative...	Yes	2	[friend, sustain]
1486	Organic Soap that provides counseling for Fost...	Organic soap that gives back to children effec...	Yes	2	[organ, organ]
1572	DIFFAIR | Eco-Friendly Swiss Designer Fashion	An exclusive collection of affordable and sust...	Yes	5	[eco, friend, design, sustain, design]
1585	Maushaus: Sustainable Desert Microdwelling	ASU Graduate Students must build a functional,...	Yes	3	[sustain, build, sustain]
1634	Tiny House From Recycled Seacontainers	Tiny Houses - Small mobile houses made from re...	Yes	2	[recycl, recycl]
1638	Brooklyn Artists + Nature + Tee Shirts = Brook...	Roni & Dawn Henning, Brooklyn Artists and Prin...	Yes	3	[natur, natur, world]

Observations:

- Some of the terms identified (e.g. row:193->'design') might not be relevant environmentally and may have to be removed from the list of ranked words. (How? -> More on this in sections below.)
- The occurence of more than one different words or the same word multiple times from the list of ranked words in the samples increases the likelihood that the sample is correctly classified as relevant. We will later use this feature to our advantage.
- Sample 398 is correctly classified, but due to the wrong reason! It found the word 'world' among the ranked word-list, but it should have been ideally 'green' which does not appear in our ranked words list. This again emphasizes the importance of more training data. The same can be said of the other

incorrectly classified samples in the dataset. (More on how to circumvent this issue is discussed in later sections.)

We now apply the same approach to samples marked as socially relevant. The results are:

```
No. of identified top words distinguising environmentally relevant blurbs: 26

Top words (is_social)
-----------------------------
support      0.135803
communiti    0.134311
area         0.102205
public       0.092388
covid        0.078722
project      0.076486
hous         0.073711
live         0.070113
shirt        0.065583
build        0.065171
end          0.057024
help         0.057024
solut        0.056928
fight        0.055831
make         0.055679
film         0.055368
save         0.055014
know         0.053547
main         0.046011
app          0.044818
individu     0.042252
children     0.042243
risk         0.041255
creat         0.04123
awar          0.04123
rai                 0
dtype: Sparse[float64, 0]
```

The accuracy on training data is as follows:

```
Categorization summary
========================
yes_count: is_social
at least one keyword    19
No keyword               2
Name: count, dtype: int64
accuracy: 0.90
Observation:
Out of the 21 samples available, 19 were classified correctly and 2 incorrectly, giving us an
accuracy of ~0.90. We can also inspect the dataframe in detail, so that we know where the res
ults were false.
```

We now inspect the data frame in detail.

	campaign_name	blurb	is_social	yes_count: is_social	ranked_words
5	Surviving the Unknown	A family struggles to survive off the grid in ...	Yes	0	[]
19	The Call - a voice to the voiceless	This is a project, which aims to save lives of...	Yes	2	[project, live]
47	Et al. Creatives	A collaborative employment, resource, and comm...	Yes	1	[communiti]
55	the breast express	pumpspotting is going cross-country to support...	Yes	1	[support]
85	MIRZ PLAYING CARDS : 2ND EDITION (feat. Hope F...	Change lives. End Slavery.	Yes	1	[live]
112	Seattle Streets to Main Street: End Child Traf...	Help me build the social impact of my award wi...	Yes	1	[build]
175	Aegis	Aegis- A turnkey security solution that scans ...	Yes	3	[area, public, covid]
317	Tree Rally	A David and Goliath story about a Sydney commu...	Yes	1	[communiti]
326	The French Quarter Parklet	We're building a public parklet on 21st at Mai...	Yes	4	[build, public, project, communiti]
328	The Veterans Daily Journal	We would like to raise public awareness of ind...	Yes	2	[public, communiti]
329	Lights in the Clouds	College students numb themselves with drugs, s...	Yes	0	[]
466	LinQupp	I am developing an app dedicated to those on t...	Yes	2	[support, covid]
487	FREE! Fitness for all	A public area for everyone to keep fit and hea...	Yes	2	[public, area]
652	Bring Know Orchestra to Boston Area Kids	With your support, Know Orchestra is seeking t...	Yes	4	[area, support, live, communiti]
850	ComfPort: Clothing With A Cause - Making Cance...	Fashion forward clothing that is designed for ...	Yes	2	[support, shirt]
961	Covid-19 Helper	An app that explains everything about Covid-19...	Yes	2	[covid, covid]
1603	Chicagoland Soccer	Support high school boys soccer\ncoverage in t...	Yes	2	[support, area]
1624	Make a Home for SweetRoot Farm	House the farmers at SweetRoot in a cozy yurt	Yes	3	[hous, build, communiti]

	campaign_name	blurb	is_social	yes_count: is_social	ranked_words
		...			
1632	(Pet-A-Tree), "where every pet deserves a pedi...	Project (Pet-A-Tree) is a humanitarian based f...	Yes	1	[project]
1634	Tiny House From Recycled Seacontainers	Tiny Houses - Small mobile houses made from re...	Yes	3	[hous, hous, hous]
1638	Brooklyn Artists + Nature + Tee Shirts = Brook...	Roni & Dawn Henning, Brooklyn Artists and Prin...	Yes	2	[shirt, shirt]

Observations

- The observations we made in the case of environmentally relevant samples are more or less valid in the case of socially relevant samples also.
- Row: 85 is interesting. It is correctly classified, but is is questionable that the words ('card') are really relevant (Emphasis on more traing data!). It is also questionabl if the manual curation is also correct in this case, since´the project is about playing cards.

Finally, we also try to get the most ranked words list taking the data frame as a whole. The least overlap in this list of words with other ranked word lists will confirm that the top ranked words list corresponding to each topic is indeed distinct and represents the particular topic it is assigned to.

```
"\nblurb_all = df[['campaign_name', 'blurb']].agg(' '.join, axis=1).tolist()\nblurb_all = ste
m([text for text in blurb_all])\nranked_words= get_ranked_words(vocabulary= blurb_all\n
,text_extracts=blurb_all, \n                              stop_words=STOP_WORDS, min_df= .05, tok
en_pattern=TOKEN_PATTERN)\n\nprint(f'No. of identified top words in all blurbs: {len(ranked_w
ords)}')\nprint('Top words (all)')\nprint('----------------------------')\nprint(ranked_word
s)\n"
```

As expected, the top ranked words in the entire dataset is different from the other words list we derived.

```
set()
```

C. Attention!

Some of the important parameters in the tf-idf algotithm relevant to our analysis are:

1. **min-df:**

   This is the minimum number of documents in which the word should appear, in order for it to be considered relevant. To ideally represent a topic, the min-df should be large. However, we have only very little training data (~50 for environmentally relevant and ~25 for socially relevant) and we run to the risk of losing information with a higher min_df value. We initially set it at 0.05%- this means, an term that appears in fewer than 5% of the documents (~2 documents) will be ignored and not considered for analysis. This emphasizes the importance of having more training data.

   Please see the <a. illustartion: min_df= 0.1> in the section below.

2. **mean tf-idf score:** It is possible to set a minimum threshold score value so that words with scores below the threshold in the ranked word list are not considered for analysis. In the analyses above, the default value is 0.05.

   please see the <b. Illustration: tf_df_threshold= 0.06> in the section below.

3. **words_num_threshold:** We saw from the previous sections that the more nummber of times different words appear in the extract, the stronger the categorisation is. In order for us to do this, we need to increase the confidence in the selected list of words.

Other methods to improve confidence:

- **Stemmimng and Lemmatization:**

  We saw stemming in a previous section. We could also experiment with 'lemmatization' and various combinations of both to try to improve the performance.

- **manual pruning of words:**

  We saw in the previous sections that (due to insufficient training data) some top ranked words might not be relevant in the domain of investigation, afterall. (e.g.'card' in socially relevant topics.) It is worth a try to manually prune the ranked words list and remove irrelevant words.

- **Manual inclusion of words:**

  Similarly, it is also recommended to include words which might be relevant to the topic. For example, words such as 'tree', 'endangered', 'eBike' etc. might be relevant to environmental projects.

a. Illustration: min_df= 0.1

```
min_df= 0.1
No. of identified top words distinguising environmentally relevant blurbs: 16.

The first column represents the relevant words and the second column gives the mean tf-idf sc
ore
We see stronger tf-idf scores, but lower number of terms which will have an effect on categor
ization.This means that there are fewer, but surer terms which indicate if the sample is rele
vant or not.

Top words (is_environmental)
-----------------------------
sustain    0.188206
organ      0.168073
friend      0.11566
design     0.100777
eco        0.093922
farm       0.089674
natur      0.086659
build      0.081072
recycl     0.079075
world      0.075608
produc     0.070755
healthi    0.069428
make       0.068772
food       0.058554
provid     0.053881
compani    0.052514
dtype: Sparse[float64, 0]
```

b. Illustration: tf_df_threshold= 0.07

```
Categorization summary
========================
yes_count: is_envt
at least one keyword    22
No keyword              14
Name: count, dtype: int64
accuracy: 0.6111
```

We see that the accuracy has dropped, because understandably there are only a lower number of terms now available for categorization. But this is not necessarily bad! it is quite possible that we were overfitting on the training data and the model might not work quite as expected on data it has not seen before. Therefore, it is again good to have more training data, so that we can increase the threshold confidemntly.

D. Categorize the dataframe

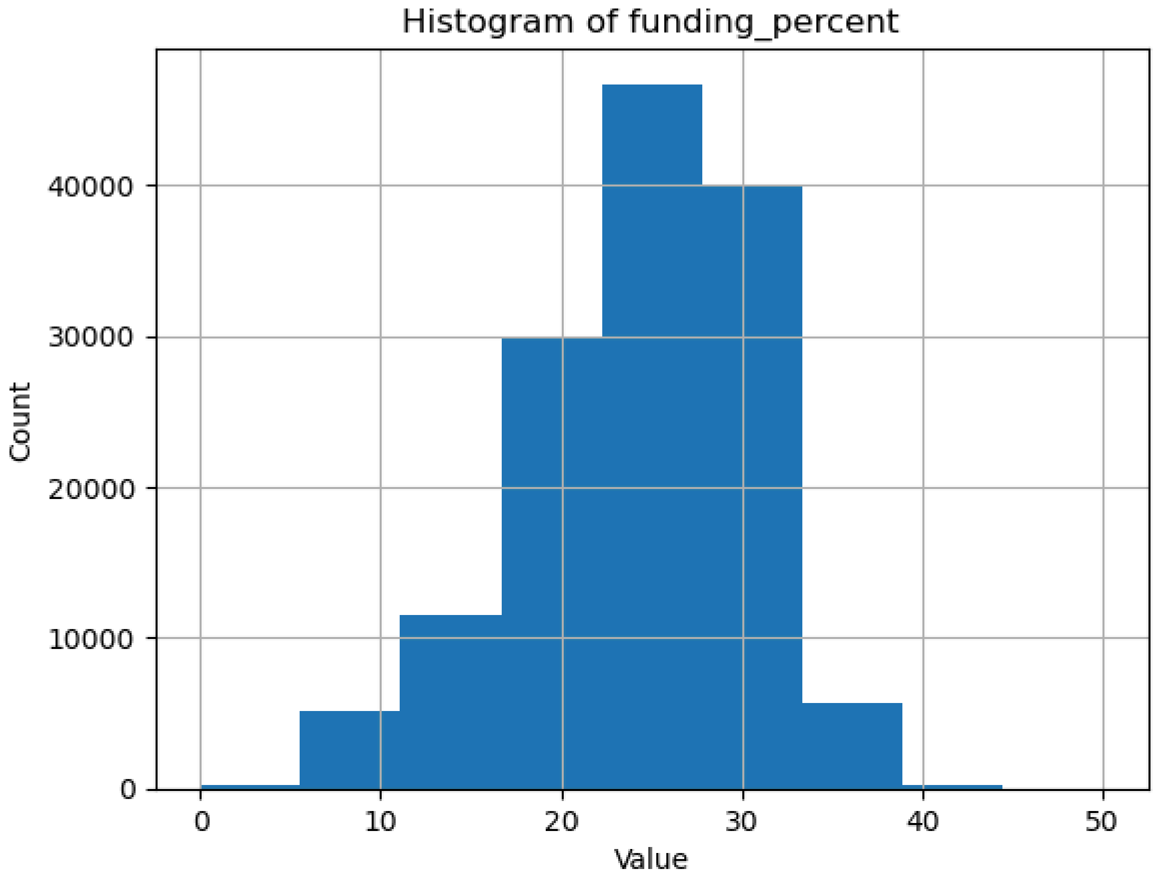
We now use the selected words list to categorize the whole data frame as socially or environmentally relevant. We use the following parameters:

- min_document_frequency = 0.05 i.e, the words should appear in atleast 10% of the corpus it sees.

- threshold_ranked_words = 0.05, we select only the high ranking words
- threshold_words_frequency = 2, Atleast 2 words should exist in an extract in order to be classified into the category we test for.

Note: The manually curated samples are NOT overwritten, irrespective of the resulting relevant word counts for these samples.

But before doing this processing, we also ensure that a minimum word count is available for successful classification. We take a look at the histogram of the campaign descriptions.



It can be observed that most samples have word count grater than 20. We set 5 as a cut-off and consider only those samples whose word count is greater than 5.

We curate the ranked_words_envt and ranked_words_social. We remove some of the words which are not representative of the categories and add manually some words (manual pruning and curation, as explained in the section above), which we think are relevant to the category.

```
ranked_words_envt:
 organ      0.104617
green           0.1
fresh           0.1
e-bike          0.1
environ         0.1
sustain    0.095428
friend     0.064725
eco        0.061742
natur      0.055885
recycl     0.053087
farm       0.052183
dtype: Sparse[float64, 0]

ranked_words_social:
 support      0.135803
communiti    0.134311
senior            0.1
farm              0.1
women             0.1
family            0.1
social            0.1
aware             0.1
children          0.1
educate           0.1
covid        0.078722
solut        0.056928
fight        0.055831
dtype: Sparse[float64, 0]
```

We apply the findings from above and categorize the whole dataframe.

After saving the data, we look at some metrics.

```
Number of samples marked as environmentally relevant: 3216; i.e, 2.305 % of total samples
Number of samples marked as socially relevant: 3638; i.e, 2.608 % of total samples
Number of samples marked as success: 76642; i.e, 54.943 % of total samples
Number of environmentally successful samples: 1247; i.e, 0.894 % of total samples
Number of environmentally successful samples: 1707; i.e, 1.224 % of total samples
```

Please also note that the data can of course contain false positives and false negatives. These can be reduced by suitably adjusting the parameters mentioned in the previous sections.

E. To Dos

- Critically analyze the findings of this notebook. Try different combinations of the suggested parameters and evaluate results.
- Critically analyze the categorized dataframe
- Manually enrich the training data as suggested in the previous sections and see if it brings out better results.