Words have meaning: language choice and startup success

Zachary Hayes, Justin Liu, Mike McCormick

2022-04-18

Introduction

Motivation Startups matter. Compared to big companies, startups contribute the majority of the net new jobs every year, and they create new opportunities for the whole society. Large companies tend to invest in incremental technologies, but start-ups often invest in subversive innovation while carrying significantly more uncontrollable risks and unpredictable returns. They extend the productivity frontier of companies and society and reap large rewards in return. A better understanding of the factors that result in startup company success is invaluable to future economic growth.

Background At the same time, startup companies are fraught with risk from having no reliable method for assessing the future success of an unproven company. They tend to lack initial investment before any revenue is generated by a possible product or service that would be offered, and other common issues such as short-term cash flows, high expenses, weak marketing and financial systems. Launching successful startups is a complex endeavor that depends upon many factors.

Y Combinator is a startup accelerator that invests in startups early on and helps them grow and take off. Since 2005, YC has invested in over 3,000 companies that are worth over $400B combined, including Airbnb, Instacart, Coinbase, Dropbox, and Reddit. YC passionately shares their knowledge and experience to entrepreneurs worldwide as a startup educational institution.

Hypothesis Our paper attempts to determine if there is a relationship between the way a company initially describes itself and subsequent success, as measured by funding and successful exit. This paper integrates data from three sources (Y Combinator, Crunchbase, and open source textbooks) and combines textual and regression analysis to determine if such a relationship exists.

We perform our research with the hypothesis that the language a company uses to describe itself in its initial stages is an indicator of later success. Specifically, the degree to which a company describes itself with business language is related to startup funding and successful exit.

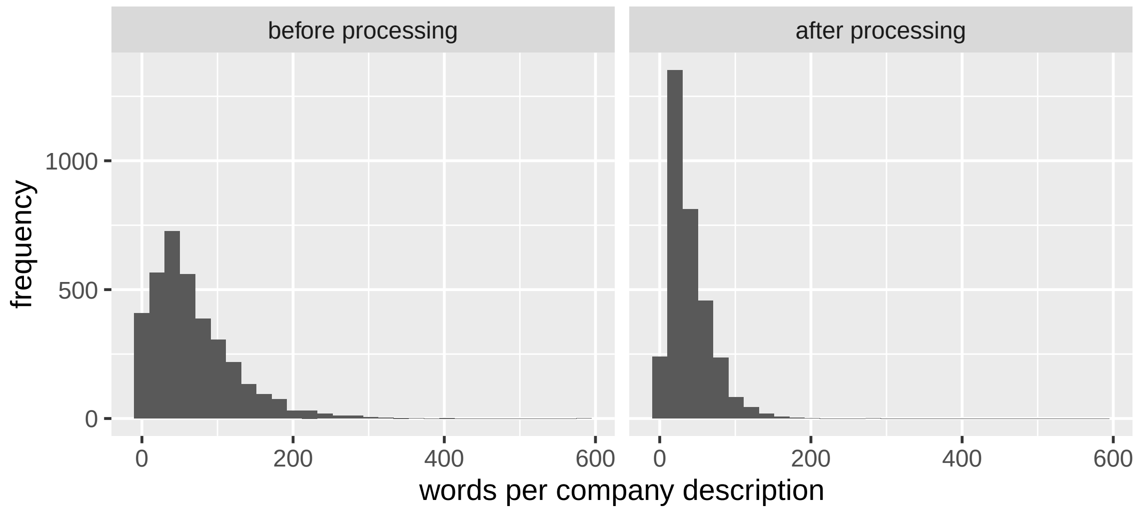
Data

To test our hypothesis, we integrated three different data sets in order to maintain integrity and to have well-rounded data. RStudio was the program deployed to pull the words that we used as the data as well as regex was utilized to clean the less meaningful words out, leaving the actual information needed to test. The three sets used were company descriptions from YCombinator, open-sourced business text books as dictionaries and then funding data from CrunchBase.

Y Combinator startup company descriptions The place we started was with Y Combinator's website. This is where we pulled each specific and unique company description from, that have participated in at least one incubation phase with them. This process utilized a for loop, to run to the site and grab each description systematically one by one placing them in a vector. Followed by cleaning the descriptions with regex which took out filler words, punctuation and symbols that took away clarity with regards to the data. This left a nice clean data set that we were able to use.

Gathering the different descriptions, it is clear there are some strengths and limitations that should be taken into consideration. A strength for this data set is that it is very large in the context of time. It is for all of the companies that went through an incubation phase and it dates all the way back to 2005. This shows growth and consistency. The time frame also shows data from different economic conditions like the 2008 crash to better conditions like after 2014 giving a broader picture to the data. A few limitations are that this data set focuses mainly on start-ups as well as the majority of the companies founded in California. This could lead the data to have a skew due to these limitations since it is from a specific time in a business's life and the companies were mainly located in a specific area. With all of that the company descriptions from Y Combinator are good to use. The below figure breaks down the distribution of words per company description, before and after textual analysis processing.

Figure 1 − histogram of words per company description



In addition to textual data, we were able to extract founding location data for the majority of Y Combinator companies, depicted in the below two figures. This data set is heavily skewed towards companies founded in the United States, specifically California.

Figure 2 − top five locations where companies are founded

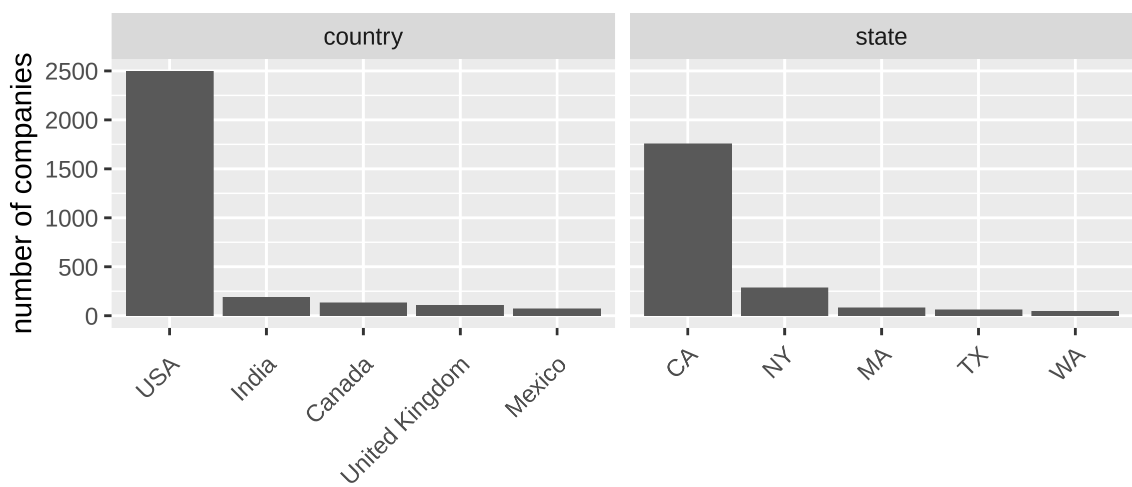
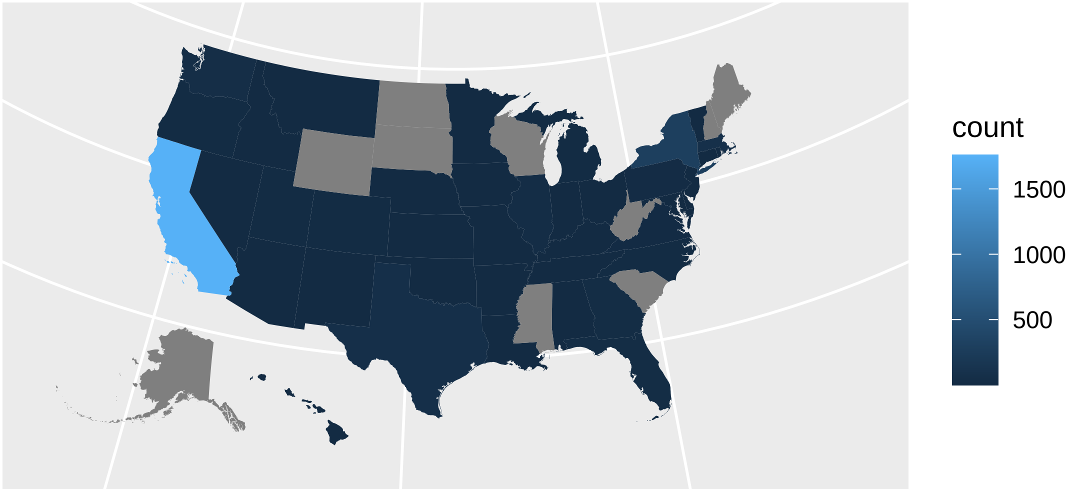


Figure 3 − geographic distribution of Y Combinator companies



Open source textbook extracts The second data set used were 5 different open source textbooks that focused on specific business topics like finance and management. These were chosen because they would allow us to test and see if the company descriptions had specific business focuses in them allowing to have measurable differences in the descriptions. Trying to find dictionaries that would allow us to find something interpretable was difficult until we realized how powerful it would be if we could come up with our own that were in essence bias free. We chose business books with specific focuses and turned the important book specific words inside them into the dictionary.

Turning the business textbooks into dictionaries was very useful for our test. Another for loop was used to go to each website and pull specific parts of the books and put them in a vector. To narrow down the words we used from the books we decided that the index, glossary and table of contents would have the most useful words in the book. Again, followed by utilizing regex again and taking the filler words, punctuation and symbols out leaving clean dictionaries to use as data sets. This leads to some strengths and limitations to consider. The limitations are that the data sets could be considered small since only the index, glossary and table of content were used due to memory constraints. As well as the text books were open sourced and not peer reviewed publications from prominent schools. However, the strength of this data set is that it is clean and allows for easy interpretable data, making it perfect for this test.

The below table lists summary statistics for our two types of textual data, both before and after textual analysis processing.

Table 1 - textual data summary statistics

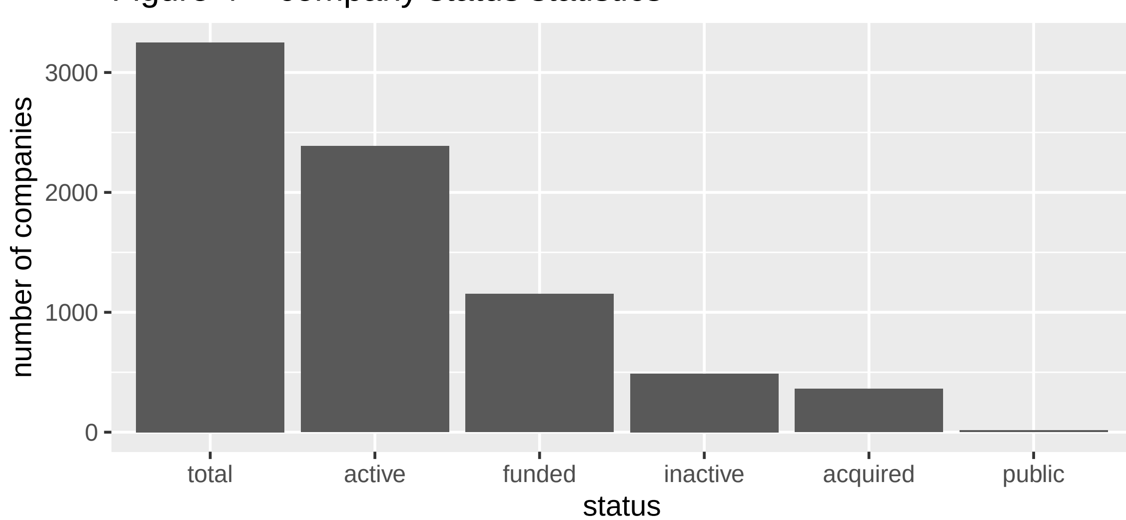
data	document count	words per document						
		minimum	lower quartile	mean	median	upper quartile	maximum	standard deviation
companies, pre-processing	3586	1	28	66.45901	53	93	586	55.06245
companies, post-processing	3251	1	19	38.47155	31	52	287	27.17250
textbooks, pre-processing	5	344	2039	3009.60000	2841	3583	6241	2170.63627
textbooks, post-processing	5	181	1067	1676.80000	1918	2065	3153	1117.73396

Crunchbase startup funding data The final data set is the funding data from Crunchbase. Professor Katie Moon kindly provided us with this data, which is from a platform that aggregates information about private and public companies. This funding data tells us the amounts, rounds and seasons that the companies were funded through Y Combinator. We used this because it is a great indicator to determine the varying level of success. The ultimate success being that the company went public but it also shows when they only went through a few seasons. The below table and figure break out the number of companies per each funding and exit status category. This is a good example of the so-called "Power Law" of venture capital - the vast majority of returns come from a very small subset of companies. In this case, the very small number of acquired (362) and public (15) companies versus the total number of companies that went through the incubator program (3251).

Table 2 - funding and exit data for Y Combinator companies

status	count
total	3251
active	2386
funded	1153
inactive	488
acquired	362
public	15

Figure 4 − company status statistics



Methodology

Once we obtained the required data, our methodology followed three main steps: (1) processing the data into useable form, (2) calculating cosine and Jaccard similarity scores between each company description and each textbook, and (3) use ordered and multivariate regressions to analyse the relationship between the calculated similarity scores and company outcomes. These steps are discussed in detail in the following sections.

Textual data Our two separate sources of textual data required different means to extract and then process into usable formats. We extracted the company descriptions using the rvest package, looping through each company's individual website and scraping the required information. For the five textbooks, we manually extracted just the table of contents, index, or glossary (as available, it varied per textbook) and then imported them into R using the pdftools package.

The initial textual data, once imported into R, was still not in a format useful for analysis. We followed several steps to convert the data into a usable format:

(1) We removed all numbers, whitespace, punctuation, and any blank or "NA" rows to reduce the data to just words.

(2) Next, we "lemmatized" each word using the textstem::lemmatize_words function. This function uses a dictionary based on the Mechura 2016 English lemmatization list. This step reduced the words to common base forms, reducing the complexity of the data and making it easier to analyze.

(3) Finally, we calculated term frequencies for each document, and then used this information to build a document-term matrix of all the company descriptions, textbook extracts, and terms. The matrix contains a set of vectors for each textbook extract and company description, with each value corresponding to a specific term and the frequency it is found in each document.

Funding and exit data We based the current status of each company based on information scraped from the Y Combinator website - each company is described as either "Inactive" (gone out of business), "Active", "Acquired", or "Public". We were fortunate in this project in that the Y Combinator and Crunchbase websites use a similar naming convention for companies, making joining the textual data with the funding data an easy step. We filtered the Crunchbase funding data down to the company name and the number of funding rounds received and joined this data to the list of companies using the dplyr::left_join function.

The Crunchbase data only included companies that have received startup funding, enabling us to define a second logical variable for whether or not a company had received funding.

Similarity scores The final processing step required was to calculate the similarity between each company description and the textbook extracts, i.e. how similar a company's initial descriptive language is to a business topic. We did this using two measures, the cosine and Jaccard similarity scores between each company description and each textbook.

We calculated the cosine similarity score for each company description using the lsa::cosine function. This function takes two arguments, the respective vectors for each company description and textbook, and returns a cosine similarity score. In mathematical terms, this score is the dot product of the two vectors divided by the product of their lengths.

The Jaccard similarity score is similar to the cosine similarity score, but only considers unique words per document instead of frequency. To calculate this score, we first converted each vector into a logical vector, based on whether or not the term was present in the respective document. The Jaccard similarity score is then calculated as the intersection of the two vectors divided by the union of the two vectors.

The mean cosine and Jaccard similarity scores for each company, grouped by company status, are listed in the below table.

Table 3 - similarity scores by company status

		Mean cosine similarity scores				
status	count	entrepreneurship	finance	leadership	marketing	strategy
Acquired	362	0.0738666	0.0255525	0.0403263	0.0272614	0.0498762
Active	2386	0.0784647	0.0298825	0.0418272	0.0335066	0.0526428
Inactive	488	0.0651645	0.0222868	0.0358728	0.0272202	0.0461308
Public	15	0.1074728	0.0321027	0.0614199	0.0352015	0.0725202

		Mean Jaccard similarity scores				
status	count	entrepreneurship	finance	leadership	marketing	strategy
Acquired	362	0.0004728	0.0002122	0.0002551	0.0000659	0.0002961
Active	2386	0.0004464	0.0002053	0.0002590	0.0000649	0.0002824
Inactive	488	0.0003831	0.0001778	0.0002160	0.0000570	0.0002429
Public	15	0.0006702	0.0002840	0.0004078	0.0001092	0.0004442

Data integration and regression analysis Combining these sources of data gave us a table with X independent variables (the cosine and Jaccard similarity scores for each company - how closely a company's description matched a business topic) and two dependent variables. The first dependent variable was logical: whether or not a company received funding. The second dependent variable was ordinal, based on a company's exit status: "Inactive", "Active", "Acquired", or "Public", in that order. We ran two regressions using this data to determine if there was a relationship between the textual data and (1) whether a company received funding and (2) its viability as a company as measured by exit status.

Results

It is difficult to draw robust conclusions from our two regressions. The results of each regression are listed in the below tables, with p-value .05 statistically significant independent variables in bold.

For the first regression, similarity scores versus a company's exit status, the estimates of the coefficient for each variable in the regression formula varied widely in both sign and value. Only one variable, the marketing cosine similarity score, was statistically significant at a p-value of .05. We could interpret this as with an one unit increase in the marketing cosine similarity score, the log odds of a company progressing through each status increases by 0.30. However, due to the wide variety in the rest of the results it is impossible to draw a solid conclusion from this.

Table 4 - exit status regression table

term	estimate	log odds	std.error	t statistic	p value	coef.type
ent_c	1.4308724	4.182347e+00	1.3944003	1.026156e+00	0.3048180	coefficient
fin_c	-0.5290431	5.891685e-01	1.4805517	-3.573283e-01	0.7208460	coefficient
ldr_c	**3.1763422**	**2.395896e+01**	**1.5903768**	**1.997226e+00**	**0.0458006**	**coefficient**
mkt_c	-1.2159323	2.964335e-01	0.8584770	-1.416383e+00	0.1566634	coefficient
str_c	-3.2406770	3.913740e-02	1.9529217	-1.659399e+00	0.0970354	coefficient
ent_j	1267.7589882	Inf	0.0037398	3.389887e+05	0.0000000	coefficient
fin_j	419.0829791	1.012571e+182	0.0030495	1.374271e+05	0.0000000	coefficient
ldr_j	-2442.6144174	0.000000e+00	0.0041386	-5.902061e+05	0.0000000	coefficient
mkt_j	745.5850996	Inf	0.0013264	5.621172e+05	0.0000000	coefficient
str_j	1078.0305844	Inf	0.0043240	2.493123e+05	0.0000000	coefficient
Inactive|Active	-1.3724071	2.534960e-01	0.0779986	-1.759528e+01	0.0000000	scale
Active|Acquired	2.4571785	1.167183e+01	0.0868109	2.830496e+01	0.0000000	scale
Acquired|Public	5.8110501	3.339696e+02	0.2673527	2.173552e+01	0.0000000	scale

For the second regression, similarity scores versus whether or not a company received startup funding, the results were marginally more clear. Four variables, the entrepreneurship, finance, and marketing cosine similarity scores and the entrepreneurship Jaccard similarity score had statistically significant estimates of their coefficient. This could be interpreted as a one unit increase in these similarity scores corresponding to a respective increase or decrease in the log odds of a company receiving funding. Again however, the wide variance in results makes it difficult to draw solid conclusions. If there were truly a strong relationship, one would expect to see similar results for each cosine or Jaccard similarity score. Instead, they again vary greatly in both sign and value, leading to inconclusive results.

Table 5 - funded regression table

term	estimate	std.error	z-statistic	p.value
(Intercept)	**-0.6021217**	**0.075806**	**-7.9429338**	**0.0000000**
ent_c	**-3.1122222**	**1.462765**	**-2.1276304**	**0.0333677**
fin_c	**-7.7819262**	**1.848884**	**-4.2089853**	**0.0000257**
ldr_c	1.7989371	1.750830	1.0274768	0.3041960
mkt_c	**-2.6269333**	**1.027653**	**-2.5562452**	**0.0105809**
str_c	3.6325517	2.087376	1.7402482	0.0818154
ent_j	**1027.8160713**	**357.883129**	**2.8719322**	**0.0040797**
fin_j	-482.8442896	541.929932	-0.8909718	0.3729443
ldr_j	-963.1932609	500.930394	-1.9228086	0.0545041
mkt_j	-129.7534025	1015.312243	-0.1277966	0.8983100
str_j	590.4033507	517.827425	1.1401547	0.2542219

Conclusion

The results of our analysis are inconclusive - we were unable to find evidence for our hypothesis. The regression results varied widely across all variables and makes it impossible to draw strong inferences about a relationship between initial word choice and subsequent startup success.

To improve on this analysis, future research should seek better sources of initial descriptive language. The company descriptions are likely somewhat skewed or biased in that they are meant for public consumption and advertising. In addition, these descriptions are limited in terms of textual data - the longest company description was 586 words before processing, and 287 after processing. A better representation of initial descriptive language choice would be a dataset consisting of original applications for the Y Combinator incubator program. This would provide a far more accurate version of choice of initial descriptive language than the proxy we used for our analysis.

In addition, a larger dataset could be built by incorporating applications from multiple incubator programs - e.g. TechStars. Due to processing and memory limitations, our dictionaries were limited to just the table of contents, glossary, and indices from business textbooks. More accurate dictionaries could be built using entire textbooks or Wikipedia pages. This would provide a larger dataset from which to build similarity scores.

References

R libraries We used the below libraries in our analysis and to develop this report:

tidyverse / tm / MASS / xml2 / rvest / urbnmapr / here / dplyr / broom / kableExtra tidytext / textstem / sjPlot / pdftools / textclean / widyr / text2vec / lsa

Company descriptions All company descriptions were scraped from their respective pages at the ycombinator.com/companies website.

Textbook extracts We extracted the table of contents, indices, and glossaries (availability varied by textbook) from the below open source textbooks to build our dictionaries:

–Burnett, J. (n.d.). Introducing marketing. Open Textbook Library. Retrieved April 21, 2022, from https://open.umn.edu/opentextbooks/textbooks/introducing-marketing

–Coleman, W., & Halbardier, A. (n.d.). Principles of Management. openstax.org. Retrieved April 21, 2022, from https://openstax.org/details/books/principles-management

–Laverty, M., & Littel, C. (n.d.). Entrepeneurship. OpenStax. Retrieved April 21, 2022, from https://openstax.org/details/books/entrepreneurship

–Reed, K. B. (n.d.). Strategic management. Open Textbook Library. Retrieved April 21, 2022, from https://open.umn.edu/opentextbooks/textbooks/mastering-strategic-management

–Taylor, J., Robison, L., Hanson, S., & Black, J. R. (2021, January 28). Financial Management for small businesses, 2nd Oer edition. Financial Management for Small Businesses 2nd OER Edition. Retrieved April 21, 2022, from https://openbooks.lib.msu.edu/financialmanagement/

Funding data

Crunchbase funding data provided by Katie Moon, Professor of Finance at University of Colorado Boulder.