Springboard Data Science Career Track Program

Capstone Project 2 - Company Industry Classification Final Report

Irene Yao
January 28, 2019

INTRODUCTION

According to Ernst & Young, US venture capital investment reached US$138b in 2018, spreading across 9,216 deals. With the exponential growth in startup formation and equity investments into those companies, being able to keep track of the entities and follow the money become crucial for both investors and investment service providers. One key step in the decision cycle is to have accurate data on the companies and decide how to analyze the entities from different perspectives including industries, development stages, and geographical locations, etc. In this project, I will focus on the classification of industry segments and how we could catalogue each company into one unique industry code based on the company's business description. The model used for cataloging is built as a supervised multi-classification problem with company description as the feature and industry segment as the label. The automatic classification method will provide information curators and users with a fast and scalable approach to data collection and potentially higher accuracy in cataloging by eliminating the human errors.

DESCRIPTION OF DATASET

The data for this analysis was downloaded from The Open Data 500 by The GovLab. It contains about 500 company profiles compiled by The GovLab through outreach campaigns, advice from experts and professional organizations, and research efforts. Information includes company name, url, founding year, headquarter, business description, and company category, etc. In this project, we will only use the "description" and "company_category" (industry segment) columns. The text description will be transformed into vectors and used as our features. The company_category is our label column.

DATA WRANGLING

The "company_category" column identifies which industry the company belongs to. There are 18 categories in total. A few companies are lacking the company category and those records are removed. Two of the categories contain only 1 company each, and those 2 records are also deleted. Meanwhile, I cleaned up some misspelled industry codes and added a numeric column to encode the industry categories as integers.

	description	company_category	category_id
0	3 Round Stones produces a platform for publish...	Data/Technology	0
1	The company mission is to provide finance to s...	Finance & Investment	1
2	At 5PSolutions, we wish to make all basic info...	Data/Technology	0
3	Abt Associates is a mission-driven, internatio...	Research & Consulting	2
4	Accela powers thousands of services and millio...	Governance	3

After the data wrangling, we obtained 524 records with their business description, company_category and corresponding category_id. There are in total 16 categories (industry segments) left for analysis.

TEXT PRE-PROCESSING

The business description is composed of free texts, which contains html tags, url, accented characters, and special characters, etc. It also contains numbers to describe address, company size, or user base, which are not directly relevant to our analysis. The following steps are thus taken to normalize the texts.

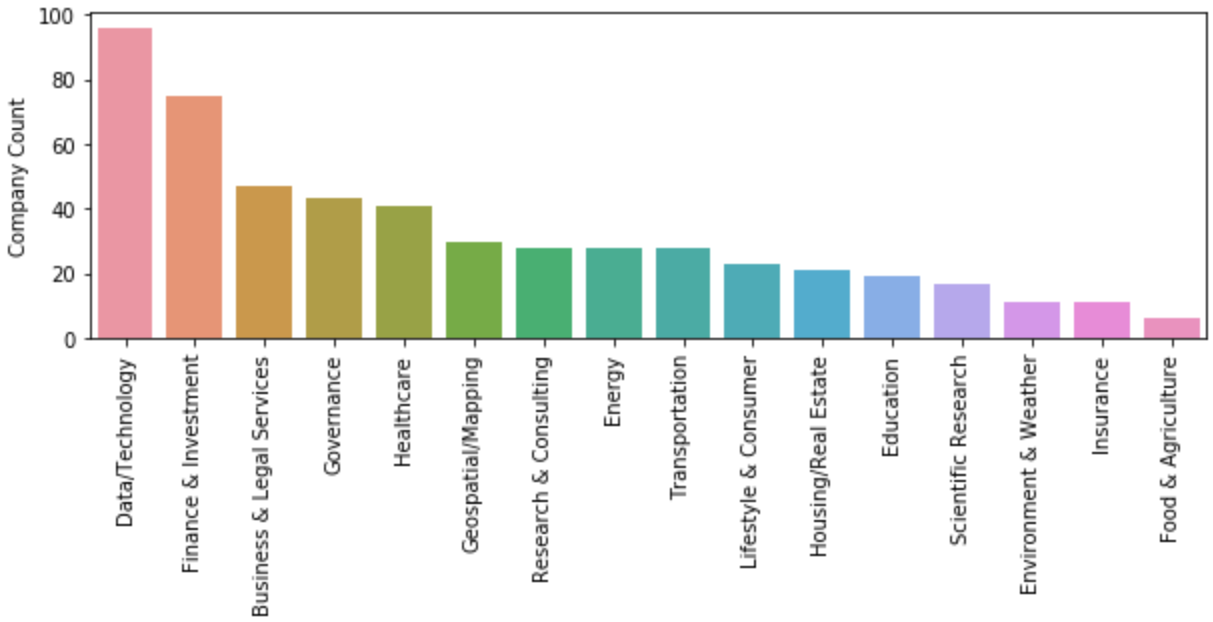
- Remove html tags, urls, numbers, accented characters, special characters, extra spaces, and stopwords
- Expand contractions to change words such as we'll into we will, there're into there are
- Lemmatize text

Normalized description after text-preprocessing

	description	company_category	category_id	normalized_description
0	3 Round Stones produces a platform for publish...	Data/Technology	0	round stone produce platform publish datum web...
1	The company mission is to provide finance to s...	Finance & Investment	1	company mission provide finance small business...
2	At 5PSolutions, we wish to make all basic info...	Data/Technology	0	psolution wish make basic information differen...
3	Abt Associates is a mission-driven, internatio...	Research & Consulting	2	abt associate mission drive international comp...
4	Accela powers thousands of services and millio...	Governance	3	accela power thousand service million transact...

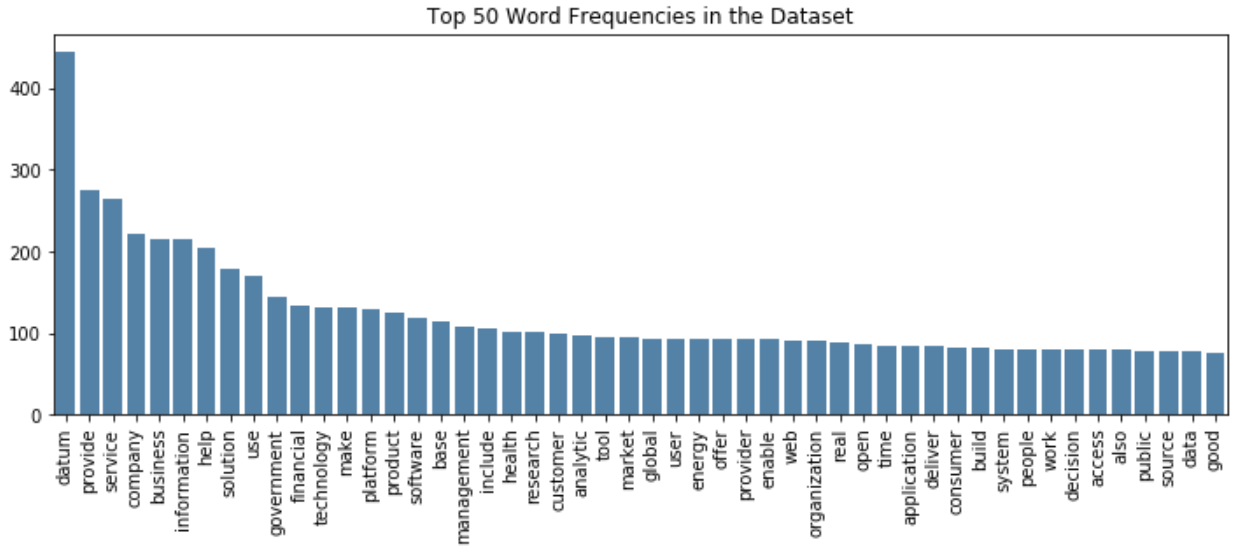
EXPLORATORY DATA ANALYSIS (EDA) & STATISTICAL INFERENCE

Before diving into building machine learning models, we will look at the number of companies in each category class.



We notice that the number of companies per category is imbalanced, with most of the companies in the Data/Technology and Finance & Investment group. Since conventional machine learning algorithms tend to bias towards the major class, we will keep the imbalanced distribution in mind when training the data.

Next, I will show a word count to quickly get an idea of the most common words that appear in the description.

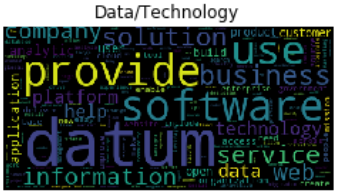


The above chart shows the top 50 most frequent words in the company summary. Among them, datum, provide, service, company, and business are the top 5 words used to describe the

companies. In order to see if any of the words are particularly used for any industries, I will generate word clouds for all 16 industries.

Finance & Investment

Data/Technology

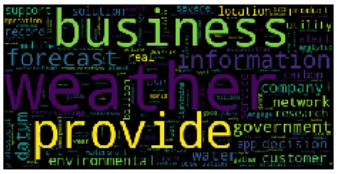


Governance

Research & Consulting

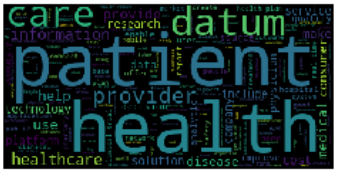
Business & Legal Services

Environment & Weather



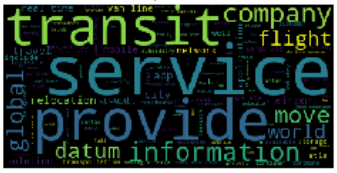
Lifestyle & Consumer

Healthcare

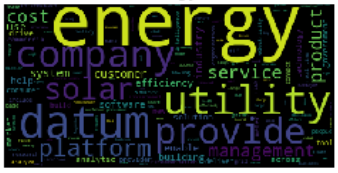


Insurance

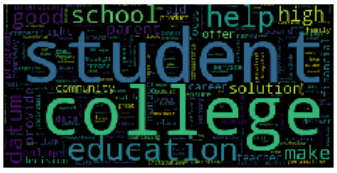
Transportation



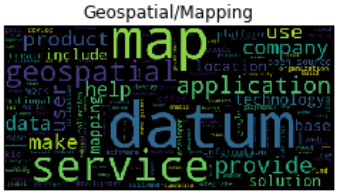
Energy



Education



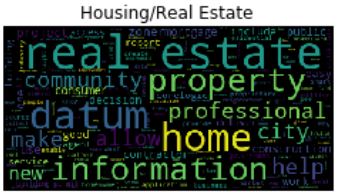
Geospatial/Mapping



Scientific Research

Food & Agriculture

Housing/Real Estate



We found that some of the words such as provide, service, company, and business are quite common in the description of all industries, yet don't reveal much information about the type of service the company provides. These may be the words we need to eliminate from the analysis.

To get a better idea of how the words are distributed, let's explore the words in the corpus. I will use the TfidfVectorizer to extract the features. Since some of the words such as the company names are very unique to their own descriptions, I will add the parameter min_df=2 to filter for words that appear in at least 2 documents. I'm also interested in both the most frequent unigram and bigram words.

After vectorizing the words, we see that each of the 524 descriptions is represented by 3624 features, showing the tf-idf score for different unigrams and bigrams. We can then use sklearn.feature_selection.chi2 to find the terms that are the most correlated with each of the industry segments.

Most Correlated Unigrams and Bigrams for each industry

Business & Legal Services

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	legal	23.169131	individual business	9.046163
1	lawyer	12.846582	customer need	4.508171
2	patent	6.770733	background check	4.442675

Data/Technology

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	datum	5.181395	business intelligence	3.637256
1	demographic	4.376913	complex datum	3.262470
2	software	3.898663	enterprise datum	3.205333

Education

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	student	70.875412	student college	12.910004

1	college	58.340190	help parent	12.012365
2	aid	26.508216	school teacher	11.038416

Energy

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	energy	100.064870	energy datum	13.45330
1	solar	35.900085	energy management	13.22840
2	utility	19.881284	energy efficiency	12.73332

Environment & Weather

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	weather	61.796405	solution business	13.416328
1	forecast	18.590747	app deliver	10.378810
2	carbon	18.334691	environmental datum	8.251736

Finance & Investment

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	financial	15.885951	institutional client	5.241050
1	investment	15.266900	individual investor	4.464934
2	investor	13.557337	credit rating	4.316367

Food & Agriculture

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	food	74.594469	find local	11.749247
1	farmer	54.048532	available market	10.048238
2	farming	36.232088	health technology	9.834192

Geospatial/Mapping

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	map	27.497008	map datum	12.421722
1	geospatial	18.389584	datum collection	10.503421
2	navigation	14.181180	location base	6.459677

Governance

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	citizen	14.391516	local government	9.677492
1	legislative	13.112537	civic technology	4.936080
2	government	11.836985	code america	4.665531

Healthcare

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	patient	47.189338	health plan	12.348129
1	health	32.869214	clinical trial	10.433997
2	care	29.668346	care provider	9.159891

Housing/Real Estate

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	estate	58.731613	real estate	60.793055
1	real	17.540759	estate agent	15.532455
2	permit	16.622051	estate professional	12.267281

Insurance

	Unigram	Unigram Chi2	Bigram	Bigram Chi2
0	insurance	111.967348	insurance financial	27.926749

1	casualty	22.463505	auto insurance	23.635972
2	watercraft	17.426867	property casualty	22.463505

Lifestyle & Consumer

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	weight	17.431761	weight loss	10.247805
1	eat	14.866746	consumer report	9.891763
2	yahoo	9.532937	use mobile	7.813908

Research & Consulting

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	consulting	18.853410	management consulting	14.863290
1	measurement	12.282955	consulting firm	13.851545
2	firm	8.904541	global management	12.374758

Scientific Research

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	chemical	33.077659	drug discovery	15.980704
1	cancer	20.902822	genomic datum	15.639425
2	scientific	17.293334	life science	12.200810

Transportation

	Unigram	**Unigram Chi2**	**Bigram**	**Bigram Chi2**
0	relocation	31.149619	van line	22.358012
1	move	24.893401	relocation company	15.656647
2	storage	23.541693	storage service	15.356687

The Chi2 score tests for independence between the features and the outcome to the predicted, in this case the independence between the unigrams/bigrams and the industry segments. Large values of Chi2 indicate that the observed frequencies of the words of a particular industry segment are far apart from the expected frequencies. Thus, the unigrams/bigrams with higher Chi2 scores are more relevant to the corresponding industry.

By skimming through the terms, we found that most of the words in each category seem to be representative of the industry.

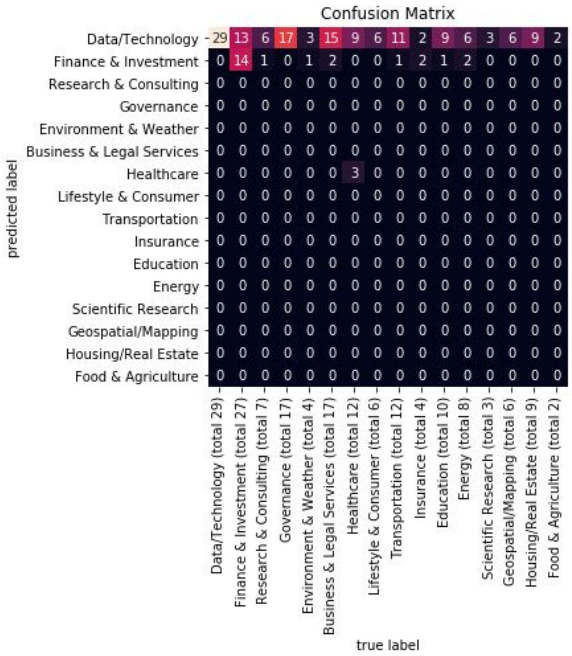
MACHINE LEARNING

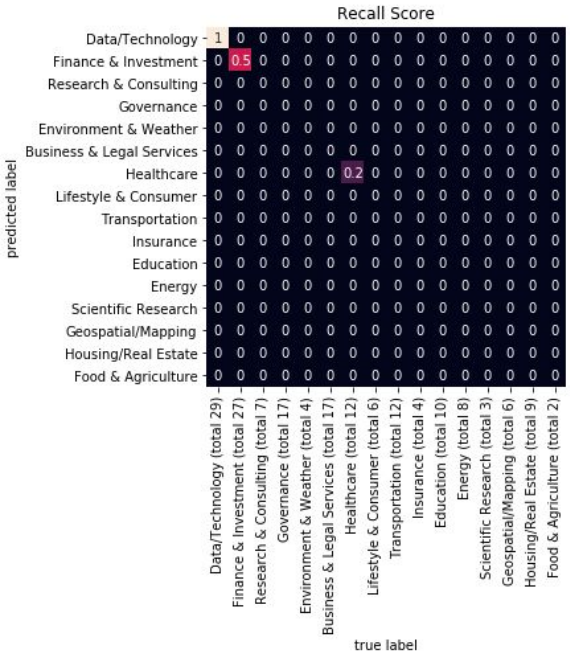
Baseline Model

Now we have all the features (the word vectors) and the labels (company_category), it's time to build our baseline model for the documents. Let's first try Naive Bayes. Since MultinomialNB is usually considered the most suitable algorithm for word counts, we will start with it.

```
## train the classifier
clf = MultinomialNB().fit(X_train_tfidf, y_train)
```

Result:
Accuracy on training set: 0.4472934472934473
Accuracy on test set: 0.2658959537572254





As shown above, the algorithm didn't perform well on either the training data or the test data. From the heatmap on the left side, we can see that regardless of the features, most of the companies are predicted as Data/Technology. We remembered from earlier that this is an imbalanced data and the result could bias towards the major class. To solve this, I will apply over-sampling technique to balance the data.

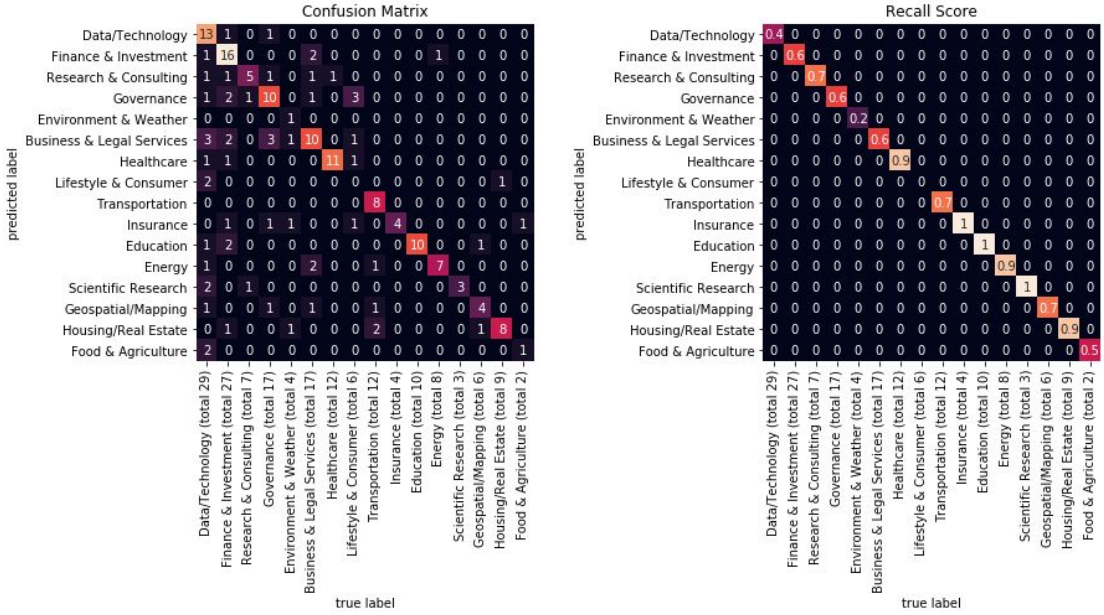
Class size after balancing the dataset

```
class_check_woSMOTE:  Counter({0: 67, 1: 48, 5: 30, 6: 29, 3: 26, 13: 24, 2: 21, 11: 20, 7: 17, 8: 16, 12: 14, 14: 12, 10:
9, 4: 7, 9: 7, 15: 4})
class_check_words:  Counter({0: 67, 1: 67, 2: 67, 3: 67, 4: 67, 5: 67, 6: 67, 7: 67, 8: 67, 9: 67, 10: 67, 11: 67, 12: 67,
13: 67, 14: 67, 15: 67})
```

Now we can see that each class has 67 samples in it. The dataset is balanced. We will fit the MultinomialNB again on the balanced data.

```
clf_balanced = MultinomialNB().fit(X_sm, y_sm)
```

Result:
Accuracy on training set: 0.9715099715099715
Accuracy on test set: 0.6416184971098265

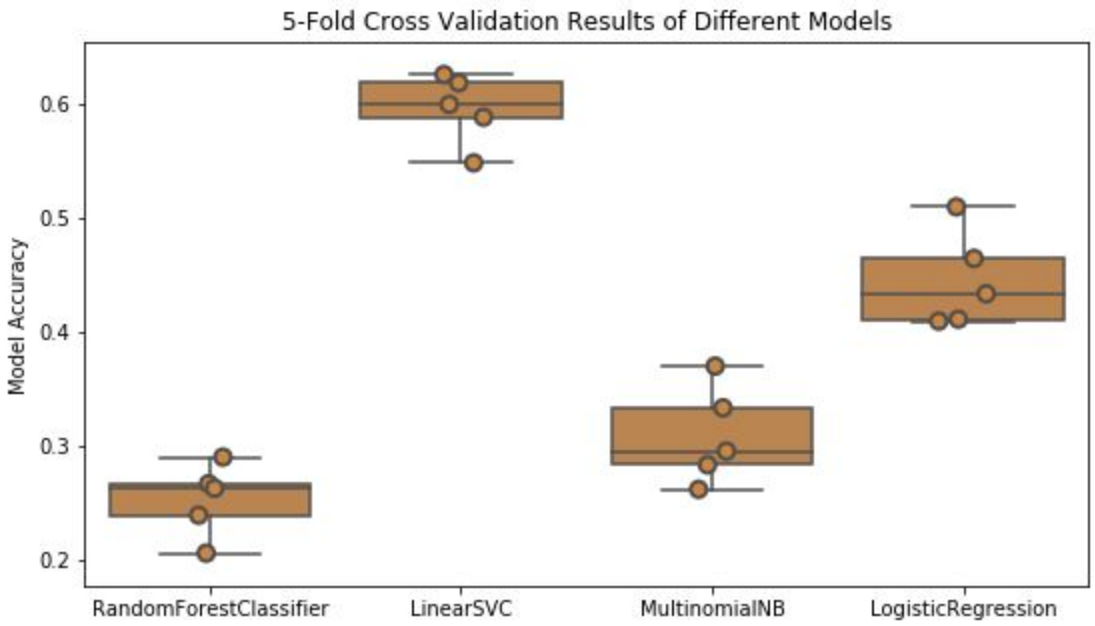


With balanced classes, the accuracy has increased from 0.266 to 0.642. More values start to emerge on the diagonal of the heatmap, which indicates correct prediction. Let's test out the model on some samples.

Sample Description	Actual Class	Predicted Class
aquicore energy intelligence solution combine software next generation meter technology cost effectively monitor analyze real time energy datum across commercial real estate property aquicor platform centraliz critical energy datum provide intuitive graphical interface streamline management team collaboration hardware allow cost effective rapid deployment across hundred property matter month regardless building age size bms system aquicor platform enable organization make timely decision improve staff productivity reduce energy waste increase tenant satisfaction meet strategic energy goal	Energy : category_id 11	11
blackrock offer mutual fund closed end fund manage account alternative investment individual institution financial professional mission create good financial future client	Finance & Investment : category_id 1	1
high medical bill lead personal bankruptcy usa primary lack price quality transparency medical industry two provider maryland charge significantly different price procedure quality care patient typically know cost procedure receive billcomparedcare found goal provide patient ability compare cost care among provider within area user may upload input medical bill information anonymously without phi personally identifiable information comparedcare medical bill repository patient actively search cost medical procedure within area comparison well read physician review submit patient one stop shop medical pricing patient	Healthcare : category_id 6	6

Model Selection

There are many other models that could be used for text classification. The following evaluation is conducted to compare several different model performance on this dataset, including Random Forest, Linear SVC, Multinomial Naive Bayes (the baseline model), and Logistic Regression.



Test accuracy of different models

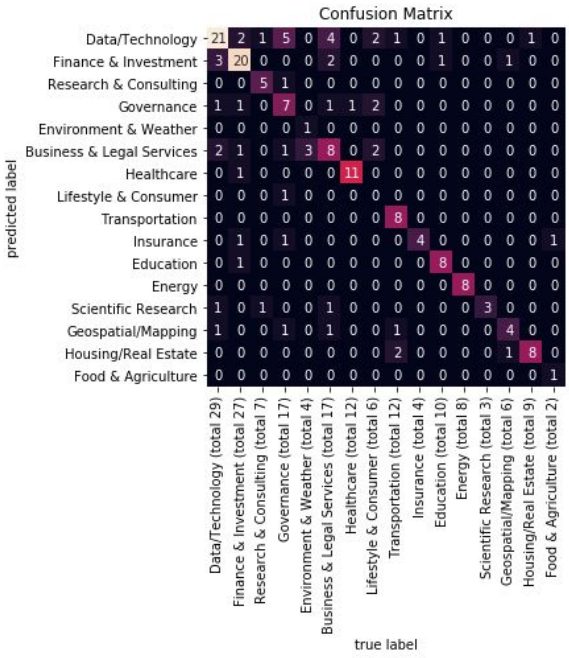
```
model_name
LinearSVC                 0.596554
LogisticRegression        0.445803
MultinomialNB             0.308688
RandomForestClassifier    0.252768
Name: accuracy, dtype: float64
```

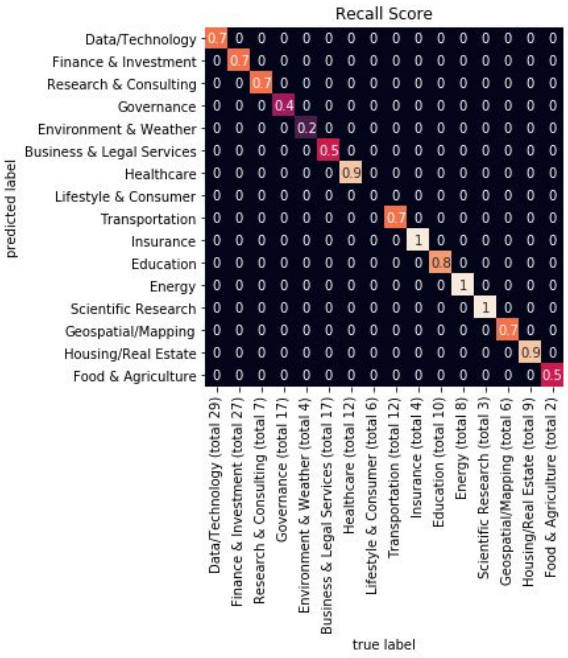
The above chart and table suggest that among the four models tested, LinearSVC performs better than the other three classifiers with a median accuracy around 59%. We will then apply LinearSVC on our data to evaluate the results.

LinearSVC Model

```
## fit the model
model = LinearSVC().fit(X_sm, y_sm)
```

Result:
Accuracy on training set: 1.0
Accuracy on test set: 0.6763005780346821





Classification Report

	precision	recall	f1-score	support
Data/Technology	0.58	0.72	0.65	29
Finance & Investment	0.74	0.74	0.74	27
Research & Consulting	1.00	0.71	0.83	7
Governance	0.69	0.53	0.60	17
Environment & Weather	1.00	0.25	0.40	4
Business & Legal Services	0.47	0.47	0.47	17
Healthcare	0.92	0.92	0.92	12
Lifestyle & Consumer	1.00	0.17	0.29	6
Transportation	1.00	0.75	0.86	12
Insurance	0.57	1.00	0.73	4
Education	0.90	0.90	0.90	10
Energy	1.00	1.00	1.00	8
Scientific Research	0.43	1.00	0.60	3
Geospatial/Mapping	0.50	0.67	0.57	6
Housing/Real Estate	0.73	0.89	0.80	9
Food & Agriculture	1.00	0.50	0.67	2
micro avg	0.71	0.71	0.71	173
macro avg	0.78	0.70	0.69	173
weighted avg	0.74	0.71	0.70	173

The accuracy score has increased to 0.676 by using LinearSVC. The model have improved the recall score for both Data/Technology and Finance & Investment and kept the score for other classes almost intact.

So far, we have built the corpus based on the training data after the train test split. What if we build the vocabulary before the split? Since we only have 524 samples, it might make sense to use all the documents to build the corpus. If in the future we are able to obtain more profiles, we could then retrain the corpus to include more words.

LinearSVC with Corpus Built on Entire Document

Corpus Size of Training Set: 2505
Corpus Size of Entire Document: 3624

```
## fit the model
SVC = LinearSVC().fit(X_sm, y_sm)
```

Result:
Accuracy on training set: 1.0
Accuracy on test set: 0.7052023121387283

By including all the documents in the corpus, we increased the corpus size from 2505 features to 3624 features. The result was a 3% improvement in the test accuracy, which reached 0.705.

Feature Tuning

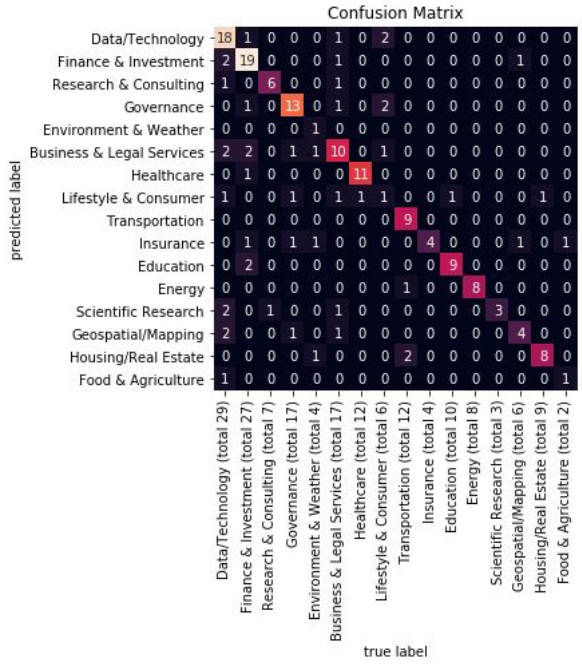
There are several parameters I would like to look into, k_neighbors in SMOTE, the penalty parameter C in LinearSVC, and the parameters of TfidfVectorizer. After testing for various k_neighbors, the value k_neighbors equals 1 generated the best result. Thus here, I will use k_neighbors equals 1 to randomly duplicate some of the records in the minority class. As mentioned before, we will use min_df=2 in TfidfVectorizer to exclude some of the words that are unique to a particular business description. We also remember from earlier that some words such as provide, service, company, and business are seen in many of the descriptions and don't really correlate to any of the industries. Thus, I will also add max_df=0.4 to exclude words that appear in more than 40% of the documents. Then I tested C for several values in the range of 0.04 to 1 to choose the optimal parameter.

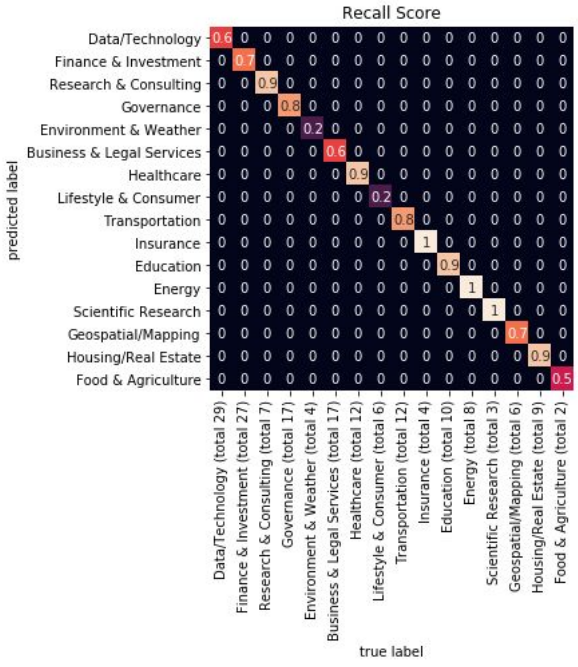
```
Accuracy on training set (c=0.04):  0.9878731343283582
Accuracy on test set: c=0.04) 0.6878612716763006
Accuracy on training set (c=0.05):  0.9888059701492538
Accuracy on test set: c=0.05) 0.7167630057803468
Accuracy on training set (c=0.06):  0.9906716417910447
Accuracy on test set: c=0.06) 0.7167630057803468
Accuracy on training set (c=0.07):  0.9925373134328358
Accuracy on test set: c=0.07) 0.7225433526011561
Accuracy on training set (c=0.08):  0.9925373134328358
Accuracy on test set: c=0.08) 0.7167630057803468
Accuracy on training set (c=0.1):  0.9934701492537313
Accuracy on test set: c=0.1) 0.7109826589595376
Accuracy on training set (c=1):  1.0
Accuracy on test set: c=1) 0.7109826589595376
```

Among all the values tested above, the best C is 0.07 and we will use the best C to build our final classifier.

```
## fit the model with best parameter
best_model = LinearSVC(C=0.07)
best_model.fit(X_sm, y_sm)
```

Result:
Accuracy on training set: 0.9772079772079773
Accuracy on test set: 0.7225433526011561





Classification Report

	precision	recall	f1-score	support
Data/Technology	0.82	0.62	0.71	29
Finance & Investment	0.83	0.70	0.76	27
Research & Consulting	0.75	0.86	0.80	7
Governance	0.76	0.76	0.76	17
Environment & Weather	1.00	0.25	0.40	4
Business & Legal Services	0.59	0.59	0.59	17
Healthcare	0.92	0.92	0.92	12
Lifestyle & Consumer	0.14	0.17	0.15	6
Transportation	1.00	0.75	0.86	12
Insurance	0.44	1.00	0.62	4
Education	0.82	0.90	0.86	10
Energy	0.89	1.00	0.94	8
Scientific Research	0.43	1.00	0.60	3
Geospatial/Mapping	0.50	0.67	0.57	6
Housing/Real Estate	0.73	0.89	0.80	9
Food & Agriculture	0.50	0.50	0.50	2
micro avg	0.72	0.72	0.72	173
macro avg	0.69	0.72	0.68	173
weighted avg	0.76	0.72	0.72	173

So far, we have reached a test accuracy of 0.723, with 125 out of 173 profiles predicted correctly. However, for those profiles that are misclassified, it would be interesting to see what those are caused by.

'Lifestyle & Consumer' predicted as 'Data/Technology' : 2 examples.

	company_category	**normalized_description**
181	Lifestyle & Consumer	fuzion app mission end wage gap woman minority within year challenge wage gap aequita mobile cloud solution enable user take control career make informed decision
99	Lifestyle & Consumer	cloudspyre llc boutique custom software development company specialize develop mobile cloud web base software solution build ios android blackberry app help people learn world around cloudspyre also partner company programming side creative project cloudspyre expertise develop website large variety application platform favorite build ruby rail website custom backend website cloudspyre app build ruby rail

'Data/Technology' predicted as 'Finance & Investment' : 2 examples.

	company_category	normalized_description
102	Data/Technology	collective ip found singular mission accelerate commercialization global rd marketplace uniquely surface idea technology inventor quickly catalyze connection buyer seller asset revolutionary way accomplish goal build comprehensive accurate organization technology emerge university company inventor
40	Data/Technology	barchartcom full service provider future equity foreign exchange market datum barchart provide wide range market datum product solution established organization industry demand accuracy innovation barchart goal form partnership deliver comprehensive solution success client financial agricultural energy medium industry

'Lifestyle & Consumer' predicted as 'Governance' : 2 examples.

	company_category	normalized_description
515	Lifestyle & Consumer	workhand free community worker trade show great connect job coworker find tool equipment display license meeting place american worker build maintain fix haul hero spend day get hand dirty make country run workhand show capable building fix maintain haul
510	Lifestyle & Consumer	wemakeitsafer mission dramatically reduce number product relate injury illness death occur year worldwide save life help build strong socially responsible financially stable company

'Data/Technology' predicted as 'Business & Legal Services' : 2 examples.

	company_category	normalized_description
108	Data/Technology	computer package inc provide intellectual property management system patent annuity payment service exclusive patent audit service improve ability manage acquisition andor divestiture efficiently
111	Data/Technology	connotate transform web datum content high value information asset feed content product grow market business intelligence enable mass datum aggregation migration integrationconnotate patent intelligent agent technology empower business user programmer quickly create data set new application content product connotate customer experience productivity gain reduced cost mitigated risk informed decision making strategic competitive advantage

'Finance & Investment' predicted as 'Business & Legal Services' : 2 examples.

	company_category	normalized_description
506	Finance & Investment	webfiling aim reinvent business reporting company provide cloud base collaboration solution eliminate version control issue important challenging report include external filing wdesk use role base access single document datamodel technology improve speed review approve sophisticated linking validation feature reinforce accuracy compliance throughout reporting process wdesk secure enable professional take control business reporting datum
43	Finance & Investment	mandasoftcom provide berkery nof information service provide merger acquisition datum via hosted searchable database base berkery noye information industry weekly report resource merger acquisition datum many top executive company shape future information although anyone interested would find mandasoftcom useful service gear towards c level executive want quickly find review information industry transaction company eye toward find match product service

'Finance & Investment' predicted as 'Education' : 2 examples.

	company_category	normalized_description
103	Finance & Investment	ecmc initiative college abacus free service allow user compare project financial aid package across school identify school within budget use net price calculator npcs build us college mandate high education opportunity act college abacus build system generate estimate post secondary institution united state
315	Finance & Investment	nerdwallet nerd create great tool crunch number give result unfiltered unbiased matter money question banking insurance health care investment education housing travel shopping offer data drive tool impartial information help make solid decision moneywe clear tool user friendly unbiased use number base analytic approach give objective results personalize customize result base financial preferences complete site list product service make money include everything findsome way help guard money choose credit card help find one save give good rewards find deal help shop smart bargain coupon invest money help avoid rip figure money stay safe grow stay get healthy help find affordable healthcare good insurance good hospitals pay college help find scholarship calculate loan compare college include law school mba program

'Data/Technology' predicted as 'Scientific Research' : 2 examples.

	company_category	normalized_description
120	Data/Technology	crowdanalytix demand crowd source service provide datum science expertise analytic manager team enterprise professional service firm crowdanalytix operate crowdsourc platform large grow community independent datum scientist solve customer problem use datum science contest publicly access datum crowdanalytix solution manager responsible manage partner project community completioncrowdanalytix headquarter silicon valley receive investment lead venture capital firmsfor information please visit wwwcrowdanalytixcompartner
340	Data/Technology	mission orlin research inc produce software product multiply power efficiency empirical research social science goal expand access complex datum source advance method scientific investigation use analyze

'Data/Technology' predicted as 'Geospatial/Mapping' : 2 examples.

	company_category	normalized_description
427	Data/Technology	social explorer easy use demographic website web base application create fast intuitive visually appeal map report social explorer anyone internet connection access work census dataour mission build simple easy use fun demographic website world
163	Data/Technology	factual datum company help developer publisher advertiser build relevant personalize mobile experience use context location

'Transportation' predicted as 'Housing/Real Estate' : 2 examples.

	company_category	normalized_description
482	Transportation	uber aim evolve way world move seamlessly connect rider driver app make city accessible open possibility rider business driver operate city today uber rapidly expand global presence continue bring people city closer
441	Transportation	spothero provide parking location chicago area expand city around us spothero mission bring exceptional service incredible deal city across united state one time add city around country

In the tables presented above, companies misclassified as Business & Legal Services contain words such as intellectual property and patent, companies categorized as Education do provide services to colleges and education institutes. Most of the misclassifications make sense and are difficult for the algorithm to place them in the correct class.

CONCLUSIONS & FUTURE WORK

To conclude, we have constructed a relatively accurate classifier to determine the industry segment of a company based on its business description. To build the final classifier, we gradually improved the features by balancing the data and tuned the hyperparameters of the text vectorizer. We also experimented different models and different hyperparameters of the models to evaluate the results. The best-performing model was achieved by using LinearSVC as our classifier with the corpus built on the entire document before the train test split. Under the best model, we were able to achieve 72.25% accuracy on our test set with majority of the recall score above 0.7. The categories Environment & Weather and Lifestyle & Consumer both have a recall of 0.2. These are the segments with very small sample size and we would hope to improve the prediction on those two segments with more data. To find all the models we worked on and the corresponding test results, refer to the link here.

For the analysis in this report, we were only able to obtain around 500 records with majority of the samples belong to two classes, Data/Technology and Finance & Investment. An important information we have gleaned from the project is the advantage to compute the vectorization using the whole corpus while doing a train and test split. Because of that, at this point, we believe that the best approach is to continue collecting the data and retrain the model using the new vocabulary obtained until the model reaches an accuracy that cannot be improved any further. Additionally, steps such as expanding the feature set and using ensemble methods to combine different classifiers have the potential to further improve the results.

Given more time, we would try other applications to the text analysis such as Latent Dirichlet Allocation (LDA) to infer concepts from the company's description in a non-supervised way. We can use these concepts to extract a list of keywords that are specific to each topic. Those keywords could be tagged to the company profiles for a broader range of research analysis.

RECOMMENDATIONS TO THE CLIENT

Currently, data collection on company profiles and investments involves a large amount of manual work and labor time. The automatic cataloging described in this report could drastically improve the efficiency of information tagging and data input. The machine learning approach could be utilized by data providers, investment firms, and service providers such as law firms and accounting firms. For clients who are going to apply the algorithm, it is recommended to train the model with large enough data to increase the accuracy of the prediction.

At the initial phase of using this model, it is recommended to implement a transition period where we would have human to check the results and retrain the model accordingly. As seen in the previous analysis, we gained a 3% increase in accuracy when we included roughly 1120 more words in the corpus. It is highly possible that the model would perform even better if we retrain it with more data. It will eventually reach a point when adding more company profiles

won't generate much more words to add to the dictionary or adding more words don't produce better results. Once the model reaches its stable and better status, we could then gradually phase out the human work and realize true automation.