Natural Language Processing for Company Business Descriptions: Classification: TF-IDF with Traditional Machine Learning Methods

Wei Ding, Wusheng Liu, Kwok Ping Ng
Technical University Munich
wei.ding@tum.de, kwokping.ng@tum.de, wusheng.liu@tum.de

1 Content Data Cleaning

Before applying any machine learning method to generate vectors for each website and doing the classification, we cleaned the webpage content as our first step. Figure 1 shows this text cleaning pipeline. Since all the content is crawled from the Internet, it contains HTML tags. So first of all we removed such tags using the package *Beautiful Soup*[1]. After that, we removed punctuation and numbers from the web content with *Regular expression package*[2], which may not be helpful for the category classification and we only kept alphabets. We also converted all the letters to lowercase to get more accurate occurrence rate of each word. Furthermore, we removed English stopwords such as ‘a’, ‘is’ etc using the *NLTK*[3] package. Then we checked the number of characters of each companies’ content and removed those with fewer than 200 characters. Because content with such few characters contains very less information or states that the domain of that websie is for sale.

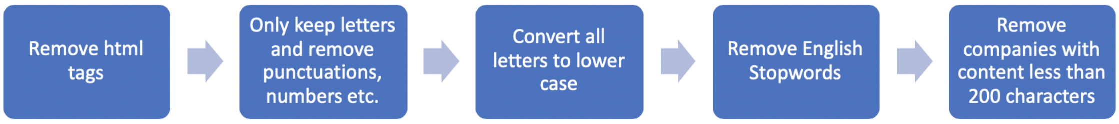

Remove html tags
Only keep letters and remove punctuations, numbers etc.
Convert all letters to lower case
Remove English Stopwords
Remove companies with content less than 200 characters

Figure 1: Pipeline of data cleaning for the website content

2 Website Categories

2.1 Data Regrouping

Each website belongs to industry groups and segments according to *VentureSource*[4] of Dowjones in the provided data file. The total number of segments are 28. The websites were essentially regrouped to 7 groups/categories according to VentureSource again. The reason is simplicity of data prediction. The labels are

- BUSINESS FINANCIAL SERVICES
- CONSUMER GOODS GROUP
- CONSUMER SERVICES GROUP
- ENERGY UTILITIES GROUP
- HEALTHCARE GROUP

[1] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[2] https://docs.python.org/3/library/re.html
[3] https://www.nltk.org/
[4] http://privatemarkets.dowjones.com/Deals/Help/VSPremium/Glossary_IndustryGroups.html

- INDUSTRIAL GOODS MATERIALS GROUP
- INFORMATION TECHNOLOGY GROUP

. Once we find out the best prediction model, we may use the 28 segments for the website prediction.

2.2 Imbalanced Data

Table 1 shows number of companies in each big category. It is very obvious that the number of companies in each category is very imbalanced. The largest category INFORMATION TECHNOLOGY GROUP has 8 times more companies than the smallest category ENERGY & UTILITIES GROUP.

Category	Number of Companies
BUSINESS & FINANCIAL SERVICES	1579
CONSUMER GOODS GROUP	416
CONSUMER SERVICES GROUP	1140
ENERGY & UTILITIES GROUP	219
HEALTHCARE GROUP	1041
INDUSTRIAL GOODS & MATERIALS GROUP	477
INFORMATION TECHNOLOGY GROUP	1885

Table 1: Number of companies in each category

In order to test the influence of imbalanced classes, we tried three different ways to gain our training and testing data used for further process:

- Group 1: directly using imbalanced seven categories
- Group 2: using the smallest category companies number (219) and selecting same number of companies with their web content from other category randomly
- Group 3: only choosing four categories which have more than 1000 companies and selecting randomly 1041 companies in each category.

We then carried out experiments and doing the classification on the data selected by these three ways.

3 Generated Vectors

3.1 TF-IDF

TF-IDF stands for term frequency-inverse document frequency. The TF is number of a word appears in a document over the total number of words in the document.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}} at\ t^{th}\ word\ and\ d^{th}\ document.$$

The IDF is the logarithm of the total number of documents over the number of documents which the word appears.

$$idf_t = log(\frac{D}{D_t}).$$

Then TF-IDF means these 2 terms are multiplied together.

$$score_t = tf_{t,d} * idf_t.$$

Therefore, TF-IDF is a matrix of words vs documents. Its entries represents the score of the word appearing in the document. This higher score of a word means the word higher appearing frequency in the document(TF) but little appearing in other documents (IDF). Then this word has higher possibility to be recognized as an identifier in current document.

We use the TF-IDF as word vectors for classification and prediction by different prediction models.

3.2 Words With High Scores

Through computing TF-IDF scores for all words from the websites, we got a vector for each company. Length of the vector is number of all the words. The value is TF-IDF score of each word in that website content. By averaging TF-IDF scores with regard to the categories, below are some interesting keywords we found among the 50 keywords with highest scores for each category.

Labels	**Keywords**
BUSINESS & FINANCIAL SERVICES	marketing,need,right,account,based
CONSUMER GOODS GROUP	order,platform,product,available,provide
CONSUMER SERVICES GROUP	best,like,experience,new,personal
ENERGY & UTILITIES GROUP	work,shall,make,solutions,energy
HEALTHCARE GROUP	health,help,provide,online,personal
INDUSTRIAL GOODS & MATERIALS GROUP	work,care,technology,products,business
INFORMATION TECHNOLOGY GROUP	systems,development,mobile,software,security

Table 2: Keywords in each category

We selected 5 symbolic ones for each category among the 50 high value words. For example, "system", "software" are the keywords of INFORMATION TECHNOLOGY GROUP and "health" indicates the HEALTHCARE GROUP. However some high scores words are obtained in every category like "information". So only the keywords could not identify the category. The machine learning classifier would also consider the distribution of values in the vectors to identify belonging each category.

4 Experiments and Results

We performed K-Means,K Nearest Neighbour(K-NN) and Support Vector Machine(SVM) on group 1-3 data. Using accuracy and weighted F1 score to compare them. Following are the results:

	K-Means		KNN		SVM	
	Weighted F1	Accuracy	Weighted F1	Accuracy	Weighted F1	Accuracy
Group 1	0.28	0.32	0.37	0.36	**0.62**	**0.62**
Group 2	0.23	0.28	**0.65**	**0.64**	0.55	0.56
Group 3	**0.33**	**0.36**	0.57	0.57	**0.62**	**0.62**

Table 3: Weighted F1 score and accuracy of different prediction models

SVM performed best on group 1 and 3 data because their data volume were more sufficient. So the sample size influenced the accuracy of the prediction when using SVM. K-NN has the both highest accuracy and F1 score in group 2 data because each category has the same sample size. Due to same reason, each category of group 1 data is imbalance therefore it gave the lowest accuracy and F1 Score. Among 3 models, K-Means achieved the worst performance. It required to calculate the distance of the features/ word vectors to classify the categories.It can be seen that features are overlapping in Figure 3. Then the distance in-between cannot be utilized. Therefore K-Means failed the prediction.

4.1 Unsupervised Learning

4.1.1 K-Means

K-Means is one of the most commonly used unsupervised learning algorithms. It works by partition n items into K clusters so that all the items are categorized to the cluster with the nearest mean. K-Means minimizes the within-cluster variances. For this project, we used the MiNiBatchKmeans from sklearn library with default settings.

From Table 3 we can see the performance of K-means of Group 1 and Group 3 are quite similar.

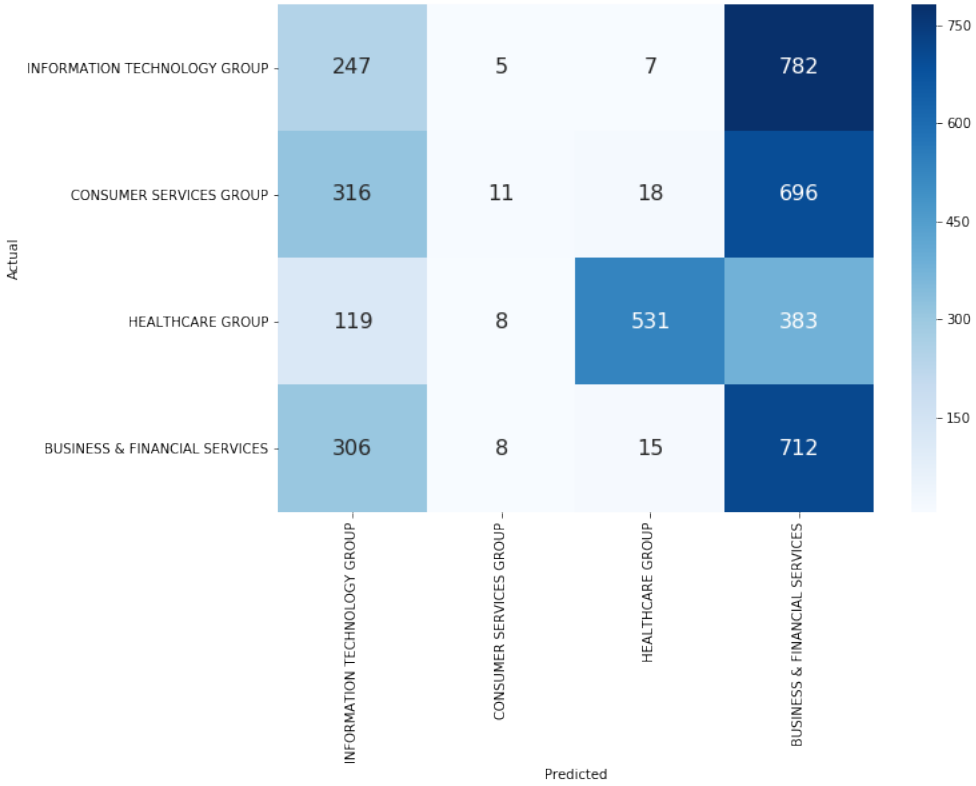


Figure 2: The confusion matrix of K-means result from Group 3

Group 3 outperformed Group 1 slightly. Although the result is not good, it still exceeds the result if we classify the website randomly, which is 0.14 for Group 1 and 0.25 for Group 3. Figure 2 shows the confusion matrix of the K-means result from Group 3. From the matrix we may notice that K-means tends to classify more companies into BUSINESS & FINANCIAL SERVICES and less data into CONSUMER SERVICES GROUP, while classifying for HEALTHCARE GROUP is relative reasonable.

In order to analyse why the results are not ideal. We also plotted the data into two dimensional space using PCA and T-SNE. Both methods are dimensionality reduction methods. PCA is better at capturing global structure of the data while T-SNE explains better the relations between neighbor. We showed scatterplots with 300 randomly sampled website of their TF-IDF score in Figure 3. The colors in the above figure represent the predicted categories by K-means while the colors in the below figure show the actual category with the content distribution. We can see a lot of overlapping data points in the below figure. This is mainly the reason for the unsatisfied performance of K-means. Since K-means only calculated the mean distance, so it is hard to tell nearby data points apart.

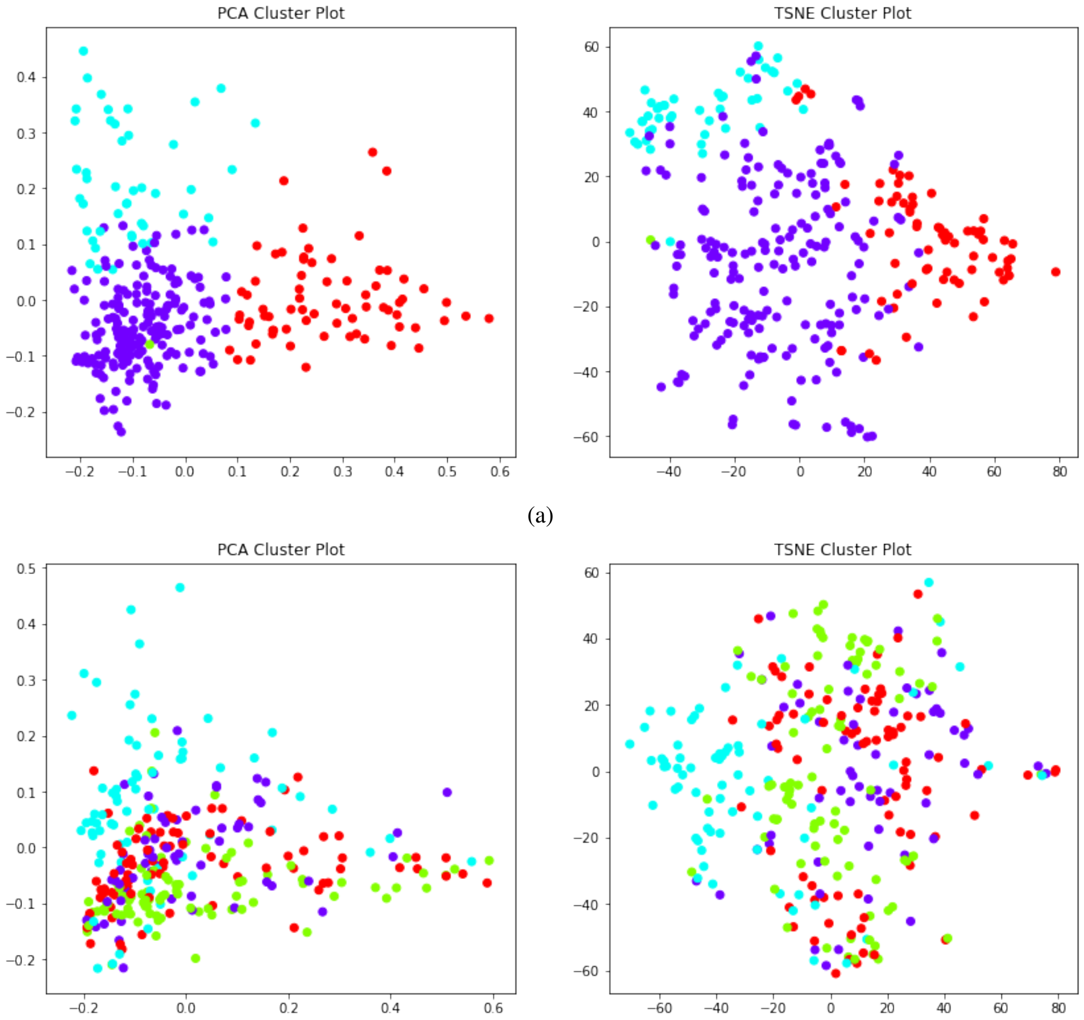
4.2 Supervised Learning

For supervised learning, two methods are currently being tried out, which are K Nearest Neighbours and Support Vector Machine. The whole labeled data are separated with a fraction of 80% for training and 20% for testing.

4.2.1 K Nearest Neighbours(KNN)

K Nearest Neighbours algorithm works by finding the K nearest neighbours in training data and use the labels of the nearest neighbours to predict the labels for the test cases. There are a few options to choose for the distance calculation. For a start, we will use the default distance function ‘Minkowski’ of the sklearn library. Further tunning and grid search shall be performed as next step to improve the performance.

Among all the three groups, Group 2 has the best test performance with an accuracy of 64% and F1



(a)

(b)

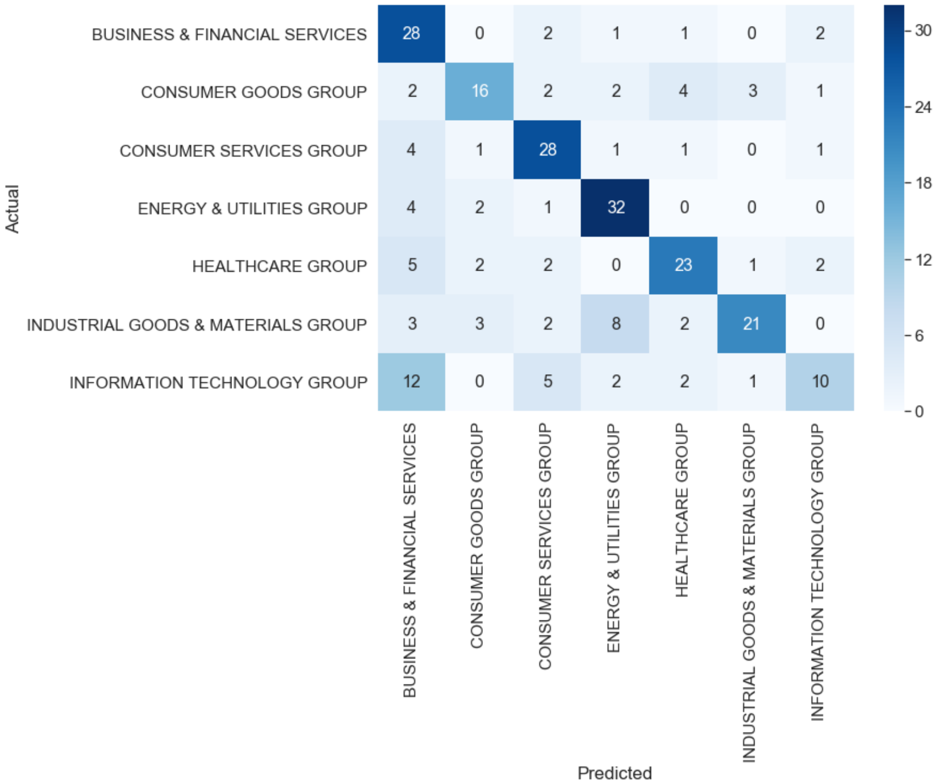
Figure 3: Visualization of TF-IDF vectors into two dimensions using PCA and T-SNE (a) Different colors represent the predicted labels from K-means (b) Different colors show the actual labels

score of 65%. The confusion metrics for Group 2 is shown in Figure 4 above. It can be seen from the heatmap that the highest numbers for each categories are mostly on the diagonal entries, which suggested that they are being correctly predicted. The only case where the model had a less than 50% accuracy is on the INFORMATION TECHNOLOGY GROUP.

4.2.2 Support Vector Machine(SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm capable of performing classification, regression and even outliers detection. In 2-dimensional space, the linear support vector machine works by drawing a line to separate two classes and make sure that it is as far away from the closest samples as possible. In higher dimensions, SVM works by drawing hyperplanes to separate data so that the hyperplane has the largest distance to the nearest training-data point of any class. For simplicity, we will use the SVM with a linear kernel as a start. Different kernels will be tried out at the next step to include non-linearity and fine tune the model.

Performing SVM on group 1 imbalance data has the best performance among all the prediction models because it has the highest accuracy and F1 score(both are around 62%). The diagonal of the confusion matrix (Fig 5) is showing number of corrected predicted categories. Most of them are the largest entry among the row or column except for CONSUMER GOODS GROUP due to lack of data



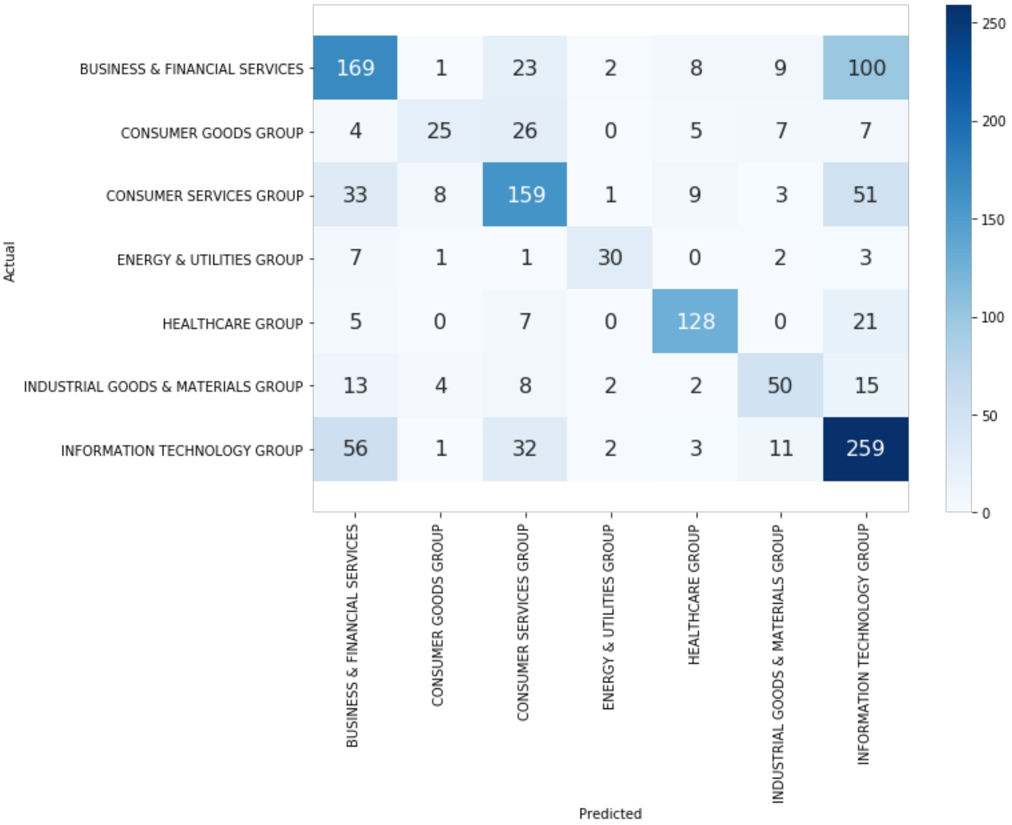
Actual \ Predicted	BUSINESS & FINANCIAL SERVICES	CONSUMER GOODS GROUP	CONSUMER SERVICES GROUP	ENERGY & UTILITIES GROUP	HEALTHCARE GROUP	INDUSTRIAL GOODS & MATERIALS GROUP	INFORMATION TECHNOLOGY GROUP
BUSINESS & FINANCIAL SERVICES	28	0	2	1	1	0	2
CONSUMER GOODS GROUP	2	16	2	2	4	3	1
CONSUMER SERVICES GROUP	4	1	28	1	1	0	1
ENERGY & UTILITIES GROUP	4	2	1	32	0	0	0
HEALTHCARE GROUP	5	2	2	0	23	1	2
INDUSTRIAL GOODS & MATERIALS GROUP	3	3	2	8	2	21	0
INFORMATION TECHNOLOGY GROUP	12	0	5	2	2	1	10

Figure 4: The confusion matrix of K Nearest Neighbours result from Group 2

in this category. The sample size are influencing the accuracy.So INFORMATION TECHNOLOGY GROUP has the most corrected predicted values due to sufficiency of data volume.

5 Next Steps

The main objective of the next step is to fine tune and improve the models listed above. We will use tab to select the website content. We will also be working on other word vectorization techniques like Word2Vec or Doc2Vec and deep learning prediction models to see if a better performance can be achieved.



	BUSINESS & FINANCIAL SERVICES	CONSUMER GOODS GROUP	CONSUMER SERVICES GROUP	ENERGY & UTILITIES GROUP	HEALTHCARE GROUP	INDUSTRIAL GOODS & MATERIALS GROUP	INFORMATION TECHNOLOGY GROUP
BUSINESS & FINANCIAL SERVICES	169	1	23	2	8	9	100
CONSUMER GOODS GROUP	4	25	26	0	5	7	7
CONSUMER SERVICES GROUP	33	8	159	1	9	3	51
ENERGY & UTILITIES GROUP	7	1	1	30	0	2	3
HEALTHCARE GROUP	5	0	7	0	128	0	21
INDUSTRIAL GOODS & MATERIALS GROUP	13	4	8	2	2	50	15
INFORMATION TECHNOLOGY GROUP	56	1	32	2	3	11	259

Figure 5: The confusion matrix of SVM result from Group 1