Natural Language Processing for Company Business Descriptions: TF-IDF Analysis and Updated Methods

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

1 Category Top Score Keywords

In this week, we made a deeper search into the TF-IDF scores of each category. We first selected N words with the highest TF-IDF keywords in each category. Some words, like "information", "service", "data", appeared in multiple categories. We considered these common high score keywords shared by multiple categories also as stopwords and removed them from the original web contents. Then we selected 8,000 words with highest term frequency as a basis of our vectors to each website and recalculated TF-IDF scores for each website. For the value of N, we chose 30, 50,100 and 200.

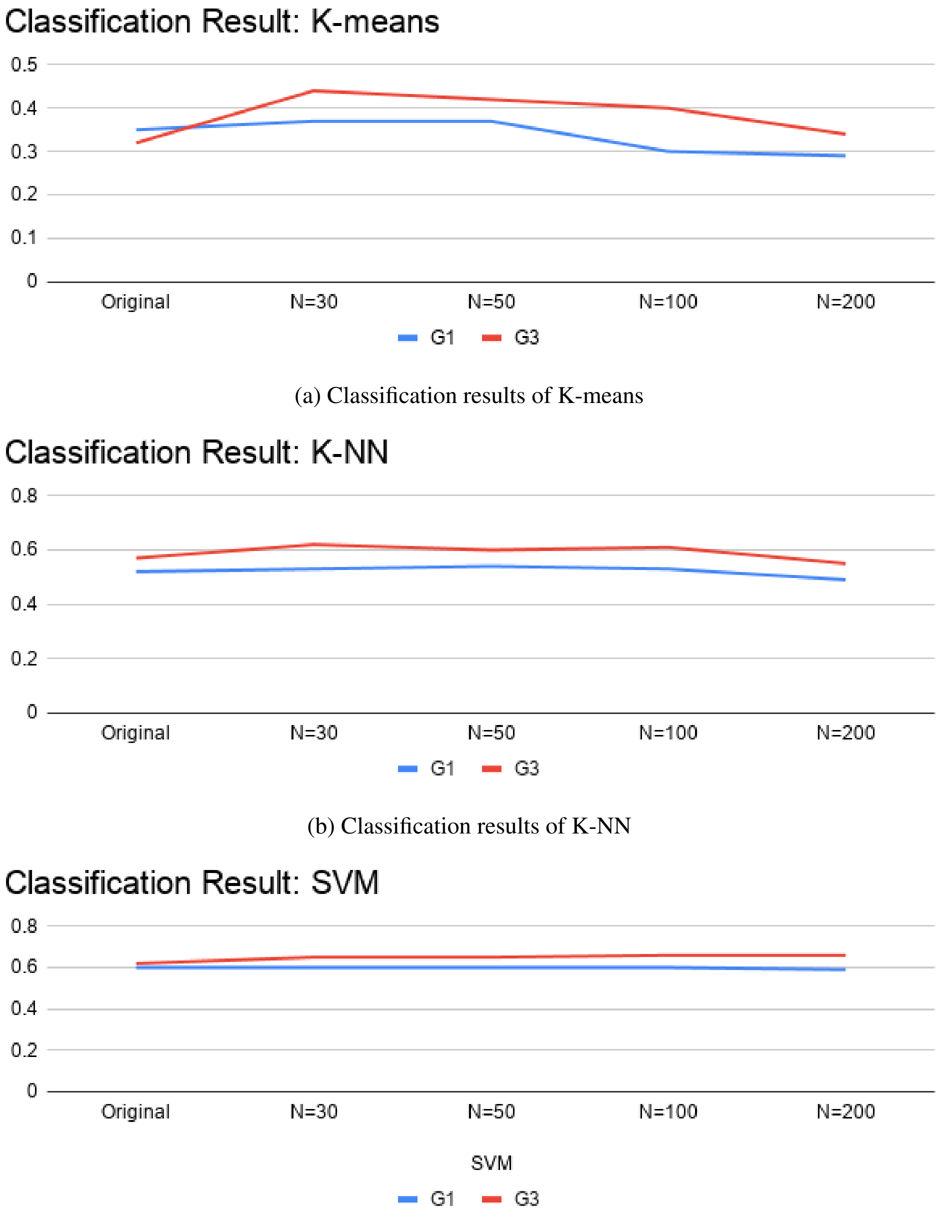
Table 1 shows top 10 words with the highest average TF-IDF scores in each categories after removing overlapping words appeared in top 50 keywords in each category. It means that these words contain most information about each category. From this table we can see quite clearly that these keywords are relative straightforward for indicating each categories.

Labels	Keywords
BUSINESS & FINANCIAL SERVICES	clients,financial,companies,platform,cloud,sales, media,insurance,agreement,learn
CONSUMER GOODS GROUP	food,order,ingredients,foods,quality,water, design,free,available,shall
CONSUMER SERVICES GROUP	app,free,agree,rights,shall,order, available,users,day,agreement
ENERGY & UTILITIES GROUP	energy,solar,gas,oil,power,fuel, water,renewable,wind,drilling
HEALTHCARE GROUP	medical,patients,clinical,care,patient,healthcare, cancer,treatment,dr,drug
INDUSTRIAL GOODS & MATERIALS GROUP	manufacturing,packaging,quality,equipment,design, high, steel,materials,aerospace,production
INFORMATION TECHNOLOGY GROUP	cloud,users,network,solution,platform,applications, app,sales,learn,enterprise

Table 1: Top 10 Keywords in each category after removing the common words shared in first 50 highest keywords in each category

We also used the updated TF-IDF scores to carry out the experiments again. Previous results suggested that G2 is having the worst performance over all the three models. This might be due to the fact that the sample size of G2 is too small. As a result, we decided to continue the experiment with dataset G1(original data with all categories) and G3(only 4 categories with each more than 1000 companies) only moving forward.

Figure 1 shows the comparison of the model accuracy by removing the common words from N top key words for dataset G1 and G3. As you can see, the model accuracy did improve after removing the common words. The accuracy is highest for K-means and KNN when N=30 or 50 while the performance of SVM is relatively stable regardless of the value of N.



(a) Classification results of K-means

(b) Classification results of K-NN

(c) Classification Results of SVM

Figure 1: Comparing the accuracy of different methods after removing the common words from N top keywords for dataset G1 and G3

Meanwhile, we could observe from Table 2 that the size of vectors for each website also influences the classification results. Especially for group 1, the classification accuracy using K-NN has increased more than 15% if we shrink the size of vectors from the size of the website content unique words(150,000) to 8,000 high term frequency words.

	Vector Size	
	150,000	8,000
Accuracy	0.36	0.52

Table 2: Comparison of Group 1 with different vector size using K-NN method

2 Other Updates

2.1 Result After Choosing Tabs

As we analysed the tab names in first week of the development, occurrence of words in tabs are taken account into the prediction during last two weeks, The tabs are selected by filtering the occurrence of words larger than 100 and the length of the word larger than 1. The original of company number were 6698. After filtering, companies left are 6388.

	K-Means		KNN		SVM	
	Weighted F1	Accuracy	Weighted F1	Accuracy	Weighted F1	Accuracy
Group 1 Original	0.28	0.32	**0.51**	**0.51**	**0.62**	**0.62**
Group 1 Choosing Tabs	**0.33**	**0.33**	0.5	0.5	0.6	0.6

Table 3: Weighted F1 score and accuracy of group 1 data whether choosing tabs

By comparing the previous group 1 original data, effect of filtering important tabs are not observable. It is because K-means results of filtering tabs are slightly better than original data but K-NN and SVM are a little bit worse. Then our assumption about important tab containing key information of the company is not true. Other tabs could also contain relevant information which is able to classify companies.

2.2 Result After K-folder Cross Validation

In order to make sure that the result from the models can be generalized to other data, we decided to adopt a technique called K-folder cross validation for the model training and validation. K-folder cross validation is widely used in applied machine learning to estimate how a model would perform on unseen data. Empirical results have suggested that a K value of 5 or 10 would yield a test error rate that suffer neither from excessively high bias nor from very high variance. For our project, we decided to use k value of 5. The result for the SVM model after applying 5-folder cross validation is shown in Table 4 below. You can see that the k-folder cross validation result is close to our original model test result(62%), which suggested that our model can be well generalized on unseen data.

	SVM	
	Weighted F1	Accuracy
Group 1 Cross Validation	0.6	0.6
Group 3 Cross Validation	0.65	0.65

Table 4: Weighted F1 score and accuracy of group 1 and group 3 of SVM Model

3 Next Steps

We will 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.