Statistics Singapore Newsletter

ISSN 2591-7382

ISSUE 2, 2022

Using Big Data to Profile Singapore's Internet Economy

by Tow Joon Han and Roger Erh
Business Statistics Division
Singapore Department of Statistics

Introduction

The internet permeates many aspects of Singapore society and economy, from the way people interact to how companies and businesses operate. Given the prevalence of internet usage for business activities, there is growing demand for a better understanding of the internet economy.

The Singapore Department of Statistics (DOS) embarked on a project to text mine web-based data to study how enterprises in Singapore make use of their corporate websites. National Statistical Offices such as Statistics Netherlands[1] have also worked on similar projects to use Big Data from the web to measure the internet economy.

In this project, enterprises are broadly classified into three categories according to the presence of a corporate website and the corresponding usage (Table 1). Enterprises under categories B and C make up the internet economy.

Data Collection

To classify the enterprises under the respective categories, the Uniform Resource Locator (URL) or the website address of each enterprise (if the enterprise has a website) will first have to be obtained. The URLs are gathered from various sources, such as surveys, administrative data, online directory and the Singapore Network Information Centre (SGNIC) which is the domain registry of website addresses ending with ".sg".

Table 1
CATEGORISATION OF ENTERPRISES ACCORDING TO USAGE OF THEIR CORPORATE WEBSITES

Internet Category	Definition
A	Enterprises without websites[2]
B	Enterprises with websites but do not generate revenue directly from their websites Example Car dealership website which does not allow for car purchases online
C	Enterprises with websites and generate revenue directly from their websites Example Hotel or airline website with booking/ purchasing features

Classification Method

A supervised machine learning classifier is then used to classify the enterprises' URLs into the respective categories based on the features extracted. As the training and testing process requires a set of labelled records, a careful matching of the enterprises' URLs to their respective categories is prepared to fit and tune the algorithm. A total of 2,100 URLs are prepared with an 80-20 split into the training and test datasets respectively. Of the different classifiers explored, a Random Forest Classifier is chosen as it performed the best in terms of test set accuracy (Table 2).

1 *Measuring the internet economy with big data.* Netherlands: CBS; 2020.

2 Enterprises which solely rely on social media and third-party online platforms were outside the scope of DOS's project. As such, these enterprises would be classified under Category A.

Table 2
PERFORMANCE OF CLASSIFIERS EXPLORED

Classifier	Test Set Accuracy
Random Forest	79%
Gradient Boosting Machine	77%
Voting Classifier	77%
Logistic Regression	72%
Neural Network	71%
AdaBoost	70%
Support Vector Machine	68%
Naïve Bayes	57%

In addition, the Random Forest Classifier offers ease of interpretation through its readily visible feature importance. Feature importance is a simple metric that indicates the relative contribution of each feature to the classifier's predictions.

For instance, for the selected classifier, the word 'Shop' has a feature importance score of 0.044 which is more than seven times the average feature importance score of 0.006. This means that the word 'Shop' is highly relevant in the classification as compared to an averagely important feature.

This allows a summary insight into the classifier's predictions. The feature importance is calculated using a machine learning package in Python (Scikit-Learn). Feature words with notable feature importance are highlighted in Table 3.

Table 3
FEATURE IMPORTANCE OF SELECTED WORDS

Feature Word	Feature Importance
Shop	0.044
Cart	0.041
Price	0.027
Facebook	0.021

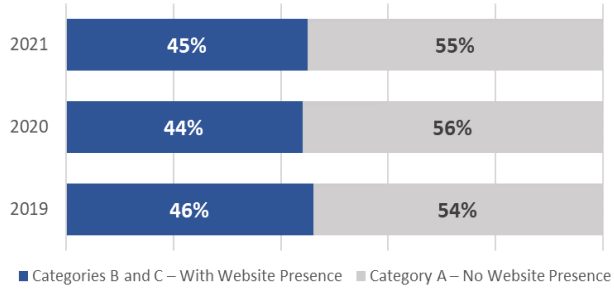
After deploying the classifier on the enterprises' URLs to obtain the predicted internet category for each enterprise, the information was merged with survey and administrative data on firm characteristics (e.g. firm activity, age, size, etc.) for further analysis.

Key Findings

45 Per Cent of Enterprises in Singapore Had a Corporate Website in 2021

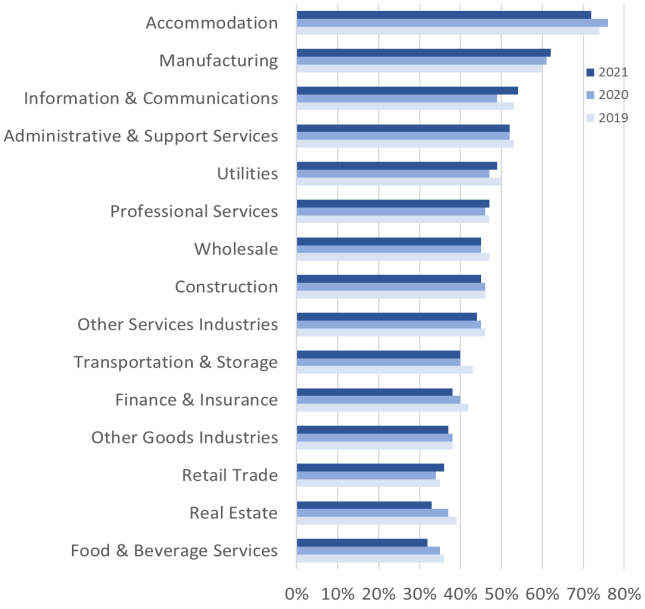
About 45 per cent of enterprises in Singapore had a corporate website in 2021, with the proportion remaining relatively stable over the past three years (Chart 1).

Chart 1
SHARE OF ENTERPRISES
WITH CORPORATE WEBSITE, 2019-2021



Among the industries, Accommodation, Manufacturing and Information & Communications had a higher share of enterprises with corporate websites (Chart 2). On the other hand, Retail Trade and Food & Beverage Services had a relatively lower share of enterprises with corporate websites.

Chart 2
SHARE OF ENTERPRISES
WITH CORPORATE WEBSITE, 2019-2021

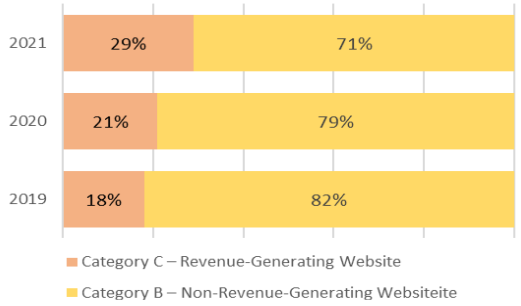


Enterprises in these industries may depend more on third-party food online marketplaces or food delivery platforms to engage their consumers online, instead of having their own websites to do so.

Increasing Share of Websites With Revenue-Generating Features over the Years

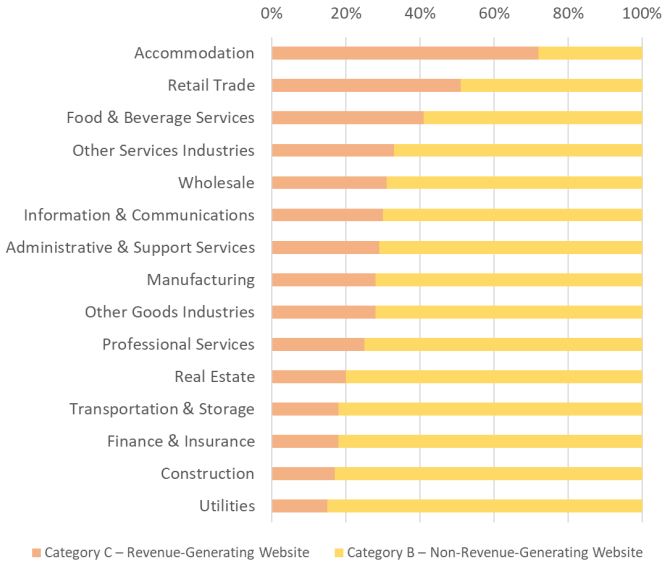
Among enterprises with corporate websites, 29 per cent of them had websites with revenue-generating features in 2021. This represented a 8-percentage point increase from 2020 (Chart 3).

Chart 3
SHARE OF CORPORATE WEBSITES WITH REVENUE-GENERATING FEATURES



Among the industries, enterprises in Accommodation, Retail Trade and Food & Beverage Services were more likely to have features on their corporate websites that generate revenue, which suggested that customers were able to purchase goods and/ or services from their websites (Chart 4).

Chart 4
SHARE OF CORPORATE WEBSITES WITH REVENUE-GENERATING FEATURES BY INDUSTRY, 2021



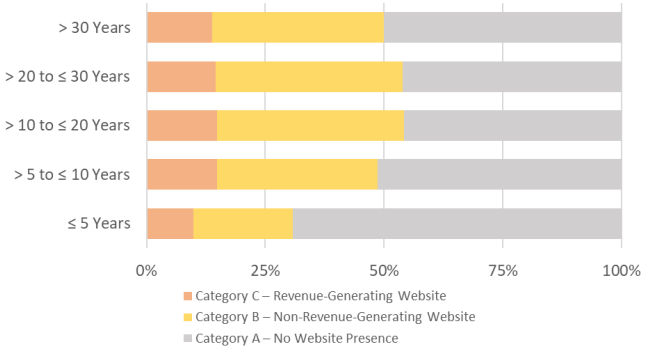
In 2021, enterprises in the Accommodation industry had the highest proportion of revenue-generating websites at 72 per cent, followed by those in Retail Trade (51 per cent) and Food & Beverage Services (41 per cent).

As these industries are mainly consumer facing (i.e. B2C), the enterprises could leverage their corporate websites to facilitate online purchases. On the other hand, industries such as Construction and Utilities[3] had the lowest share of revenue-generating websites, which may be attributed to the nature of their businesses.

Older Enterprises More Likely Than Younger Enterprises to Have Websites

Based on the profile of enterprises by age, majority of firms that were more than 10 years old in 2021 had corporate websites (Chart 5). On the other hand, only 31 per cent of enterprises aged 5 years or less had a corporate website. This implied that younger enterprises may depend on social media or third-party platforms to engage their consumers online, which is not in the current project scope.

Chart 5
SHARE OF ENTERPRISES WITH CORPORATE WEBSITES BY AGE GROUP, 2021



Within the Retail Trade industry, enterprises aged 5 years or less had a higher share of corporate websites with revenue-generating features as compared to enterprises in other age groups (Chart 6).

Chart 6
SHARE OF CORPORATE WEBSITES WITH REVENUE-GENERATING FEATURES BY AGE GROUP AND SELECTED INDUSTRY, 2021

3 The Utilities industry comprises enterprises engaged in electricity, gas, water, sewerage and waste management (includes materials recovery) activities.

This suggested that while younger enterprises in the Retail Trade industry were less likely to have a corporate website, those with one were likely to have websites that are revenue-generating (i.e. allow for online purchases). However, this was not observed for other industries.

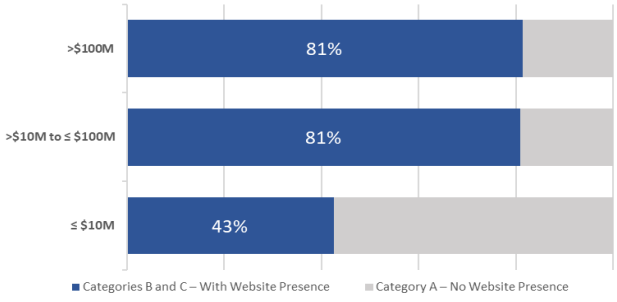
For the Accommodation industry, older enterprises tended to have revenue-generating websites compared to their younger counterparts. As for the Food & Beverage Services industry, the share of revenue-generating websites also increased with age but there were many firms aged 30 years and over which did not use corporate websites to generate revenue.

Larger Enterprises Were More Likely to Have a Corporate Website as Compared to Smaller Enterprises

81 per cent of enterprises with operating revenue exceeding $10 million had a corporate website in 2021, almost doubled the 43 per cent recorded by enterprises with operating revenue of $10 million and below (Chart 7). It is possible that smaller-size enterprises might prefer to tap on social media or third-party online platforms.

Chart 7

SHARE OF ENTERPRISES WITH CORPORATE WEBSITES BY SIZE OF OPERATING REVENUE, 2021



Enterprises With Corporate Websites Had Higher Productivity Compared to Those Without

In 2020, enterprises with corporate websites had an average productivity[4] (as measured by average nominal value-added per worker (VAPW)) of $103,000, while enterprises without recorded an average productivity of $54,000 (Chart 8). Across all industries, enterprises with corporate websites reported a higher average VAPW as compared to enterprises without (Chart 9). Enterprises with corporate websites in the Administrative & Support Services, Accommodation and Real Estate industries reported much higher average VAPW compared to enterprises without. On the other hand, within the Wholesale Trade industry, the average VAPW were similar for both enterprises with and without corporate websites.

Chart 8

AVERAGE PRODUCTIVITY OF ENTERPRISES WITH AND WITHOUT CORPORATE WEBSITES, 2020

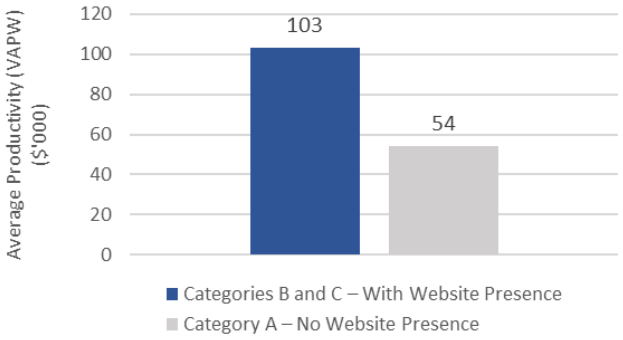
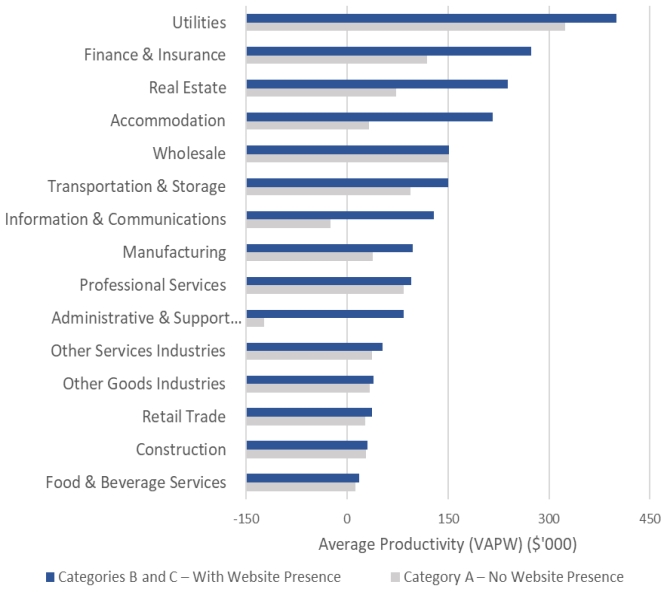


Chart 9

AVERAGE PRODUCTIVITY OF ENTERPRISES WITH AND WITHOUT CORPORATE WEBSITES BY INDUSTRY, 2020



Conclusion

The project demonstrated the feasibility of leveraging Big Data and machine learning in profiling and measuring Singapore's internet economy. The information obtained through text mining online sources can be combined with survey and administrative data to provide further insights on the profile of enterprises in Singapore with and without internet presence.

4 Refers to average productivity per enterprise and is compiled based on employing enterprises.