MACHINE LEARNING IN FACILITIES & ASSET MANAGEMENT

ZIGENG FANG
University College London

MICHAEL PITT
University College London

SEAN HANNA
University College London

ABSTRACT

In this article, we explore a machine learning approach that helps the Facility Management (FM) manager and FM data analyst to do the FM data clustering and classification automatically. Through experiment the popular machine learning algorithm, we examine how the current available Machine Learning (ML) & Natural Language Processing (NLP) technic can help solve the interoperability issue faced in the field of FM. The finding of this research indicates that: 1. The deep learning network is able to classify the building assets according to their group elements (level three of NRM's Elemental Standard of Cost Analysis) with a high accuracy rate (by more than ninety percent accuracy);2. The Convolutional Neural Network (CNN) Classifier can achieve better accuracy performance than the junior building data analyst; 3. The Unsupervised Skip-gram Gradient Descent Model can cluster the words in the document into different groups; 4. The Unsupervised Skip-gram Gradient Descent Model can reveal the hidden relationship inside the FM data. For future researches and projects, this research enlightens the future direction of applying ML/NLP techniques in the field of FM.

Keywords: Facility management, Machine Learning, Natural Language Processing, Text Classification, Interoperability, RICS NRM: New Rules of Measurement.

INTRODUCTION

During the past 15 years, interoperability is gradually becoming an unavoidable topic in the facility management industry. Concerns over integration and interoperability have been widely discussed in surveys conducted in many countries (El-Saboni, Aouad and Sabouni, 2009; Hartmann, Gao and Fischer, 2008; Rezgui and Zarli, 2006; Anon, 2017). The cost of lacking interoperability can be huge. It was identified by a study published by the US National Institute of Standards and Technology (NIST) (Gallaher, O'Connor and Dettbarn, 2004), an annual loss of $15.8 billion was occurred in the US capital facilities industry in the year 2002, resulted from the lack of interoperability among computer-aided design, engineering, and software systems.

The problem of lacking interoperability occurs not only among different design and software systems but also in strategic asset management processes. With the development of the PPP (Public Private Partnerships) and PFI (Private Finance Initiative) projects in the UK, some facility management industry pioneers have started to concentrate on managing the different building assets strategically as a portfolio base. So that similar building assets can be managed on a wider strategy approach (SAMP) (IAM, 2016). However, a proactive strategic maintenance plan cannot be accomplished without knowing the up-to-date asset condition information. Therefore, methods (e.g. Asset defect survey) are used to assess the current condition of building assets. Projects under portfolio based strategic asset management, in many cases, were surveyed by different in-house or out-source surveyors and includes a number of projects, which creates various interoperability issues. Data analysts are usually required to recode and categorise different assets from different projects according to the standardised asset coding format (e.g. BCIS Code (RICS, 2012b)). The current problem is recoding processes are often error-prone, as they are costly, mechanically, repetitively, and manually conducted categorisation activities. To obtain a better accuracy and liberate data analyst from doing the highly repetitive classification task over and over again. We need a better solution to lower the cost of asset standardisation and solving interoperability problem between different projects. The data interoperability problem steps in not only during the strategic asset management process, when assets from different projects need to be reclassified for the strategic asset management, but also during the existence of different

exchange scheme flavours of different stakeholders. For instance, two different vendors might interpret the same standard in two different ways during the encoding of the same piece of information (Shen et al., 2010). However, given currently available tools, most of these tasks were conducted in a manually time-consuming and expensive manner. Thus, the construction industry needs a more automatic way of clustering building asset into different groups and validating the model more efficiently. So how we can customise the data structure and meet the needs of different stakeholders (e.g. building owner/manager) in a timely manner? It is suggested that the Machine Learning based Text Classification technics is the potential solution.

It was stated by Bird that the classification is choosing the correct class label given certain inputs (Bird, Klein and Loper, 2009). Many of the news that post in website contain their own tags like: “sport”, “technology”, and “economy”, and many of these tagging processes are now done through classification models, in which a lot of manpower and time are saved. In the AECFM industry, building assets and operation tasks are rich in both quantity and type. How to manage this huge amount of information becomes a problem that every AECFM professional needs to be faced with. Also, as previously mentioned, the building information has been passed around different parties, and different buddies need the data to be arranged in their own preferences. Classification model enables us to take a structured approach towards both the arranging and naming of asset information. Now, with the help of ML/NLP technique, the process of retrieving and reusing information can be done even faster and more automatically. So, FM personals can reduce their time expenditure over dull data categorising, asset model quality ensuring, and interoperability issue fixing, instead, spend more time over strategic analysis and management of the building lifecycle model. In addition, due to the fact that the description used to classify the asset is majorly text-based, a feature extract method (e.g. word2vec) is needed to transfer text data into the numerical format so that it can be later inputted into the machine learning classification algorithm. Therefore, the first aim of this research is to test whether the word2vec technics can help provide the word vectors that show the semantic relationship between different words. Then, the second aim of this research is to test whether the CNN Classifier can perform better than the manual classification (at 65.91% for the junior data analyst), given the embedded word vectors space provided by word2vec technics. This study will contribute to the understanding of whether the FM asset defect survey data is applicable for machine learning applications and how FM interoperability problem and cost of strategic asset management can be alleviated through machine learning applications. Since there is currently very limited academic research covering machine learning applications in the FM area. This study can enlighten more researcher in exploring the possibility of applying machine learning technologies in the FM context.

RESEARCH OBJECTIVES

All the above queries and obstacles presented to give the guidance for setting objectives of this research.

Objectives of this paper:

1. To test whether the word2vec model can help reveal the hidden relationship between the keywords inside the FM data and allow the FM text description to be further used by neural based classifier model (Word2vec is defined as the model that provide the vector representation of words, in which can carry the semantic information of the words. The more detailed information of Word2vec and the vector representation of words will be introduced in the Literature Review and Methodology)
2. To test whether the neural based classifier can achieve better accuracy rate against the manual classification accuracy rate (at 65.91%) and find out the best parameters for training convolutional neural network

The Literature Review & Methodology Section overviews the literature on three theories from the domain of machine learning: a) Text Classification b) Word Embedding and c); High-dimensional Data Visualisation, as well as, the structure and guidance of building asset data coding used in the first experiment. Two models used in experiments are then outlined in the Research Design Section. The Background of Collected Data Section offers the research datasets’ background and the Results & Discussion Section presents the discussion and analysis of the research. Finally, the Conclusion Section offers conclusions and suggestions for future research.

LITERATURE REVIEW & METHODOLOGY

TEXT CLASSIFICATION

Text classification plays a crucial role in many applications across different disciplines (e.g. document retrieval, web search, and spam filtering). Machine learning algorithm such as logistic regression or Support-vector-machine (SVM) is the heart of these applications. When inputting the text into the algorithm, they normally need to be first converted into a fixed-length vector to properly function with algorithms. One of the most common fixed-length vector representation is the bag-of-words method that raised by Harris in the year 1954 (Harris, 1954). This method is famous for its simplicity, efficiency

and good accuracy in many of datasets. But in this research, the "word2vec" method is used instead of the traditional bag-of-words method.

Text classification (TC) as previously mentioned has been utilized in many different areas. In web mining area, Phan, Nguyen, & Horiguchi (2008) have used TC to classify the domain disambiguation for 12,340 web search results. Yang & Pedersen, (1997); Toman et al.,(2006); Chen, Huang, Tian, & Qu, (2009); and Uysal & Gunal, (2014) have applied them for classifying News items with sizes varied from (764-19,997). While for the medical area, Moschitti & Basili (2004) have applied TC to classify the disease for 28,145 medical abstracts. For the academic area, Vo & Ock, (2015) use the TC in an 8,100 sized scientific documents' classification according to their titles. Although the size of these training datasets varies from hundreds to tens of thousands, their number of categories are relatively limited (normally less than ten). Only a very few of them have the categories number above fifty, but none of them larger than a hundred (Kobayashi, V.B. et al., 2018).

Among these different applications, the dataset used in Dave's (2003) product review classification research shares the highest similarity against the dataset used in this research. To start with, the length of the product review is similar to the asset description as they carry similar functions: to describe the property of an item. Although The training size of Dave's dataset is over 30,000, which is larger than the dataset used in this research (over 10,000), they both fall into the acceptance level which is larger than the rest of previously mentioned studies. The most significant difference between these two datasets comes to the number of categories. While Dave's dataset has only 7 different categories, this experiment has more than 30 different categories. This is one of the major challenges faced in this classification task. In terms of the experiment result, Dave's team achieved 76 percent in their most-confident tercile, which can be used with the manual classification result as the benchmark for this research. Due to the fact that there is no previous academic research applies text classification in the FM context, this research is trying to test whether the text classification can also be applied in FM.

WORD EMBEDDING

Word Embedding is defined as a natural language processing technique that aiming at mapping semantic meaning into multi-dimensional geometric space. The embedding process connects the numeric vector to the word captured in the dictionary, so that, the Euclidean Distance (or Cosine Distance) between two vectors can seize part of semantic relationship embedded in the corpus for selected words.

One of the most popular benchmarks for indicating the quality of the embedding is solving word analogies. Finding the linear relations between word pairs (such as *king:man :: woman:queen*) is one of the famous examples given by word analogies. Mikolov's study in the year 2013 showed the proportional analogies (a is to b as c is to d) can be resolved by vector calculation as (c - a + b = d) (e.g. king - man + woman = queen) (Mikolov, Yih and Zweig, 2013). Subsequent research has also been carried out to evaluate the performance of word embeddings (Mikolov, Yih and Zweig, 2013). "A 'good' word embedding should encode linguistic relations in a way that they are identifiable through linear vector offset", stated by Aleksandr (Aleksandr, Anna and Satoshi, 2016).

For FM data, "window" and "pump" are assumed to carry significantly different semantical meanings. Therefore, it is fair to assume that their representing vectors will be very far apart in the embedding space. While for "door" and "ironmongery", they are more likely to fall into the same FM asset category, thus, their vectors' embedding space positions should be close to each other.

In a well-structured embedding space, the semantic relationship between the two words should be precisely captured. For instance, the "path" (displacement) vector that comes from "failure" and "repair" should capture the semantic relationship between two concepts. In this case, the relationship between "failure" and "repair" is "severity × costs". In concept, after construct embedding space the following vectorial identity should be able to be revealed: failure + (severity × costs) = repair approximately. As a result, the questions like "Does the building asset need repair?" can be answered by applying the relationship vector.

T-SNE FOR WORD VECTOR VISUALISATION

"T-SNE" is a technique that variated from Stochastic Neighbour Embedding the published by Hinton and Roweis in the year 2002. It's proved to be able to provide better visualization by reducing the tendency to crowd points together in the centre of the map (Maaten and Hinton, 2008). It is suggested that t-SNE is better than existing techniques at creating a single map that reveals structure at many different scales (Maaten and Hinton, 2008). This is especially important for the high-dimensional data that spreads on several related, but different, low-dimensional manifolds. Therefore, it is more capable of capturing the local structure of the high-dimensional data. In the meantime, it can also reveal global structure like the presence of clusters at several scales. This technique is used for providing visualisation of the word vector space.

NEW RULES OF MEASUREMENT (NRM)

The New Rules of Measurement (NRM) is considered as a comprehensive guide to good cost management of construction projects and maintenance works (RICS, 2013). It was published by Royal Institute of Chartered Surveyors (RICS) Quantity Surveying and Construction Professional Group. Besides providing a consistent approach for estimating, cost planning, procurement and whole-life costing with a standard set of measurement rules, it also enhanced the understanding of measurement rules for the construction project stakeholders (Lee et al., 2011; RICS, 2012). The adopting of NRM facilitates consistency and benchmarking and helps avoid disputes potentially.

According to RICS's Elemental Standard Form of Cost Analysis (SFCA) (2012b), building element is defined as '…a major physical part of a building that fulfils a specific function or functions irrespective of its design, specification or construction'. A clearly defined and standardized element will help building professionals to store and exchange cost information correctly.

There are four levels of analysis categories in BCIS's SFCA. The first level is called "Total building", which consists of two sub-set: Building and External works. The second level of analysis is named "Group elements (Concise)" which includes the following elements: "Substructure; Superstructure; Internal finishes; Fittings and furnishings; Services; Prefabricated Buildings and Building Units; Work to Existing Building; External Works; and Facilitating Works (SFCA)" (2012b). While the level of details used in this research is level 3 - "Detailed elements (Detailed)". For this project the following elements are included: "Builder's Work in Connection with Services; Ceiling Finishes; Communication, security and control systems; Disposal Installations; Electrical Installations; External Drainage; External; Fixtures; External Services; External Walls; Fencing, Railings and Walls; Fire and Lightning Protection; Fittings, Furnishings and Equipment; Floor Finishes; Frame; Fuel Installations and Systems; Heat Source; Internal Doors; Internal Walls and Partitions; Lift and Conveyor Installations; Roads, Paths, Paving's and Surfacing's; Roof; Sanitary Installations; Services Equipment; Soft Landscaping, Planting and Irrigation Systems; Space heating and air conditioning; Specialist Installations; Stairs and Ramps; Substructure; Ventilation Systems; Wall Finishes; Water Installations; Windows and External Doors". Lastly, the most detailed level of analysis in BCIS's SFCA is called: "Sub-elements (Amplified)" which is not covered in this research.

RESEARCH DESIGN

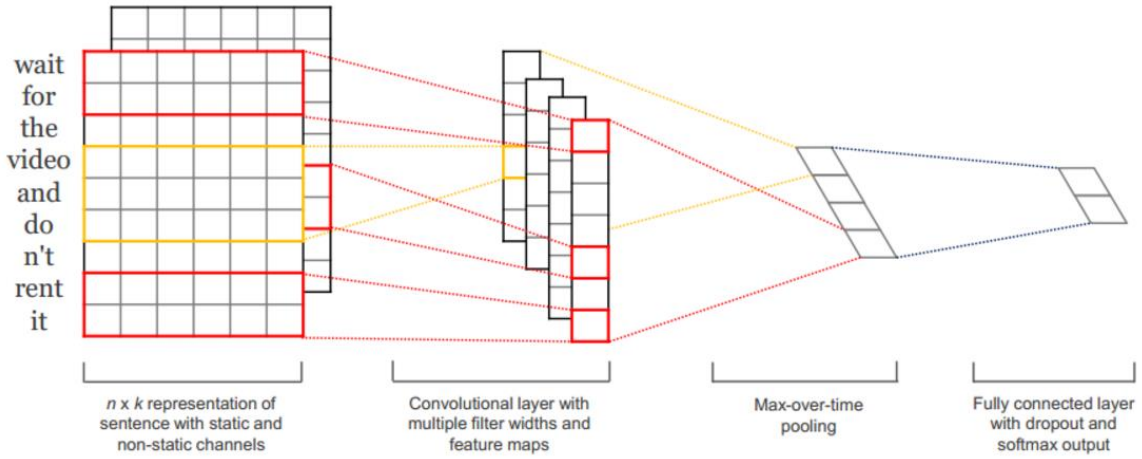
In order to achieve the two above objectives mentioned in the Research Objectives Section, two different but correlated models were designed to conduct two tasks: 1. The first task is to train the multidimensional embedding space, by referring to the structure of the Word2vec Embedding Network, with the Home Depot Project Dataset. In addition, this model will also be used to produce the 2-dimensional visualization of word embedding and prepare the default words embedding for the second task; 2. The second task is to classify the input asset into different categories by feeding the CNN text classification model with tagged input data with the pre-trained multidimensional embedding space (matrix).

WORD2VEC EMBEDDING NETWORK MODEL

The aim of the Word2vec Embedding Model is to put all of words into the same space. So that the semantic similar words can be clustered together. How to set up this embedding is the question needs to be answered by this model. In the model of word2vec, the sentence will be loaded into the embedding space randomly. After that, the nearest words will be predicted through the logistic regression algorithm. These words will, then, be compared with the real adjacent words of this input word, and the result will be used to correct the input word's position in the vector space. Thus, a high-dimensional word embedding space can be generated. Moreover, to visualise this high-dimensional embedding space, t-SNE is used in this case to reduce the embedding space's dimension and display words vectors in 2-D format.

CNN TEXT CLASSIFICATION MODEL

Figure.1: Model architecture with two channels for an example sentence (Kim, 2014)



The architecture of the CNN Text Classification Model is built based on Kim's (2014) paper. However, there are several differences between Kim's model and the model used in this research. The first major difference is the convolutional layer in this research model doesn't have multiple feature maps. Therefore, there will be only one feature map (and one filter width) in this model for each training. The second major difference is, due to the length of the documents used in this research, the dropout technique hasn't been applied in this research. The third major difference is that the static channel is not been used in this channel. The reason for doing this is to fine-tune the value of word vectors (this allows the value of word vector to be affected by the backpropagation). In the input layer, the matrix will be used to store word vectors that corresponding to the "word" or "short phrase" in the sentence. If the sentence has n number of words, given the dimension of vector equals to k, then the size of this matrix will be $n \times k$. While for the convolutional layer, the size of the convolutional window ($h \times k$) is determined by two parameters, in which, h represents the number of the words and k, again, is the dimension of the word vector. In the pooling layer, either the average value or the maximum value of the feature map will be extracted and output in the form of a one-dimensional vector. Finally, in the fully connected (SoftMax) layer, the SoftMax function will be used to do the task of classification based on the previously calculated result. For the final classification, the output will compare the probability of each categories to determine the predicted class for input text.

BACKGROUND OF THE COLLECTED DATA

THE BACKGROUND OF RESEARCH DATA

The research dataset for classification is obtained through the asset defect surveying of a large hospital project. The asset description obtains through asset surveying will be used as the only supporting information that enables both manual and Machine-learning classification processes. The manual classification task is conducted firstly by two junior data analysts (one with one year; the other one with three-year facility management data analyst experience) from the strategic asset management team, then, the result of the classification is further corrected by the senior operation manager, in which the corrected version was regarded as the "correct dataset" for training.

The number of data entry for this case study is 11,156, with 32 different categories. The dataset used in this case study has 11,157 number of data entries with 32 categories in total. If compared with the above-mentioned examples in the Literature Review and Methodology section, this dataset is not the largest dataset, however, its relative moderate size and categories number still provide a strong support for its feasibility.

In this dataset, eighty percent of the randomly chosen project dataset will be used as the training dataset for training, while, the rest twenty percent of the data will be used as the test dataset to evaluate the performance of the CNN classification model. The average accuracy (at 65.91%) achieved by data analysts is used as the accuracy benchmark for this task. Although this accuracy might be lower than the asset surveying professional with more experience. However, as the most of similar task is conducted by data analysts or administration officer, comparing the machine learning classification with the data analysts' classification is still meaningful.

To build the embedding (word vector) space, this research uses the Home Depot Product Search Relevance dataset provided in Kaggle as the training dataset for word2vec embedding model. Home Depot Inc. (or Home Depot) is the largest American home improvement retailer, where its product portfolio highly overlapped with many of the building asset items. The intention for using this dataset is trying to utilize the wording relationship within the test string of Home Dept dataset to provide default word vectors for the embedding space that used later for neural based text classification (Home Depot, 2016).

RESULTS & DISCUSSION

To better evaluate how this research answers two objectives raised in the introduction section, the structure of this section will be divided into two parts: 1. Results of Word2vec Embedding Network and 2. Results of CNN Text Classifier.

RESULTS OF WORD2VEC EMBEDDING NETWORK

In this experiment, the Home Depot Project dataset is used to generate the embedding space for later classification. The objective of the embedding process is to help set default positions of words from the project dataset, in the vector space. So that semantic similar words are gathered together. This process will enable many further applications (e.g. finding synonyms, clustering words from the same category, or obtaining derivatives through vector addition and subtraction).

Basic procedures of word embedding generating and visualizing processes are summarised as follows:

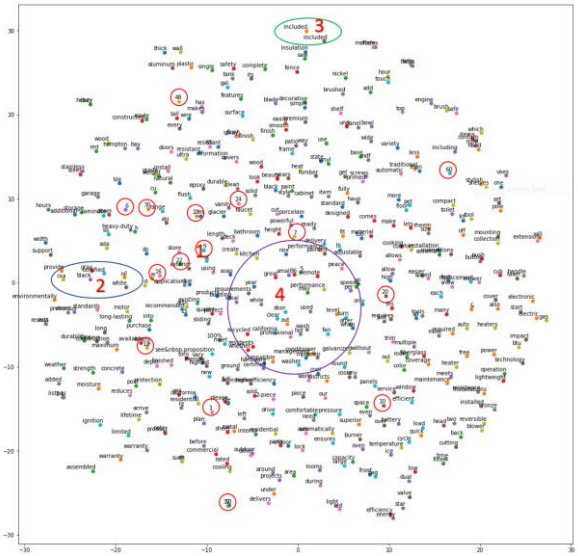
- The first step is to pre-process and load the words from the Home Depot project dataset into the model.
- The second step is to take out the words from the sentences (e.g. window) and put them into the embedding vector space. Initially, word vectors' positions in the space are all set randomly.
- While, the next step is to use methods like logistic regression to predict and further correct word vectors' positions in the vector space by referring to other word vectors in the space. After repeating this process many times, the embedding space can then be obtained.
- The last step is to reduce the dimension of the vector space to 2 through the t-SNE (t-distributed Stochastic Neighbour Embedding). So that, word vectors can be visualized in 2-D format.

RESULTS OF 2-D EMBEDDING BY USING T-SNE PACKAGE

In author's hypothesis, the whole "word cloud" can be separated into different "sub-clouds" that keep a certain distance from each other. In each "sub-cloud", words share similar pattern with each other. Some of these patterns can be recognised by professionals that have experience in AEC/FM industry. In this case, the pattern of numbers, units, colours, some adjectives that have the similarity (e.g. adjectives that describe the quality of the asset), and typos are supposed be found in the clustered pattern. In the following paragraphs, two typical clustering examples will be used to demonstrate different experiment outcomes.

A POORLY CLUSTERED EXAMPLE

Figure.2: The 2-D word embedding clouds diagram generated through t-SNE and Matplot package



Unlike Figure.3, Figure.2 is an example of poorly clustered 2-D word embedding clouds diagram. The most obvious evidence is that the allocation of "numbers" and "units" are not concentrated in a specific area, instead, they were scattered everywhere. Moreover, it is also not possible to find the similar sized group like the second group showed in Figure.3. Instead, small patterns that can be found in Figure.2 are smaller in size (compared with group 2 & 3 in Figure.3). It was noticed that the proper tuning of the embedding network and t-SNE package's parameters are important for determining the quality of clustering and visualising the FM-related knowledge.

A WELL CLUSTERED EXAMPLE

Figure.3: The 2-D word embedding clouds diagram generated through t-SNE and Matplot package

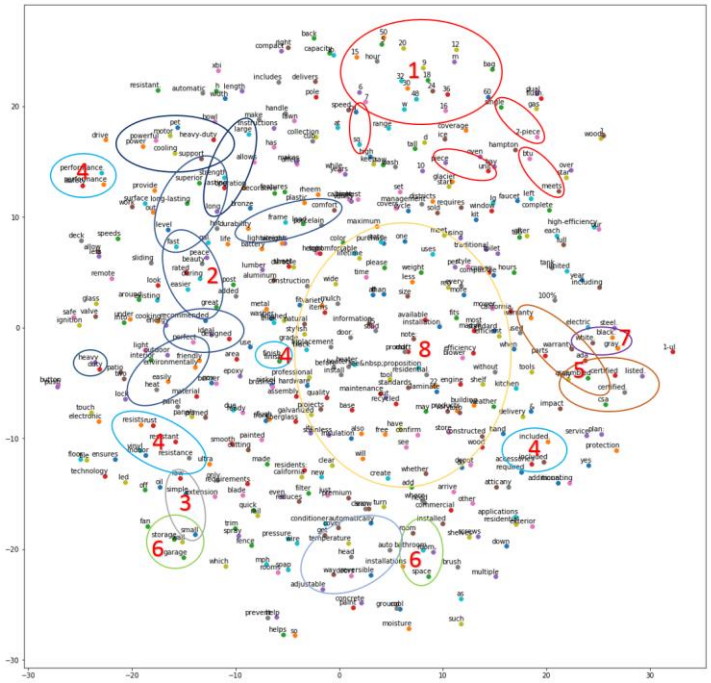


Figure.3 is one of well-clustered examples obtained from the grid analysis of embedding model. All plotted words were selected based on its frequency of appearance. Five hundred of most frequent words were selected to generate the visualising diagram, with the removal of fifty meaningless stopping words (e.g. "the", "a" …). In general, data points have not been separated into different "sub-clouds" that keep a certain distance from each other, as expected by the assumption, which potentially indicates that words in the Home Depot Dataset cannot be easily linearly clustered. The fact that these visualising words were selected based on word frequency is another potential contributor to this overall clustering shape. Apart from the above reasons, the Home Depot dataset, unlike many datasets used in other t-SNE examples, do not have pre-classified labels, therefore it is not possible to visually cluster them into different groups directly through their labels.

However, there are still many patterns that can be recognised through the embedding model. In Figure.3, the word cloud is visually clustered into 8 different feature groups. The first group is marked with red ovals. This group includes mostly numbers (e.g. 18, 32, 30…) and units (cu., sq., w…). They were well concentrated in the dense cloud's upper right corner. This group is a good indicator of the quality of clustering as numbers and units have larger quantity and easier to identify. The second group comes to words that represent the "good" or "great" features of the building asset. In this case, "good" and "great" are defined as a generalized term. For instance, some words, like "powerful" and "heavy-duty", represent the good capability of the FM equipment. While words like "beauty", "easier", "great", "ideal", "friendly", "comfortable", and "environmental" are also typical words that people generally use when commenting something positively. In Figure.3, those words can be easily identified and sit on the left-hand side of the diagram. The third group includes words that stand for "small" or other negative impressions. The number of words for this group is low, which indicates that these "negative" words were infrequently used in the Home Depot Project dataset. This might be a potential indication of the risk of abusing the praised word in product descriptions. Group four is consisted of sub-groups of words that share the similar spelling or typo. The identification of this group indicates that the embedding word vector can help find and cluster the potential misspelled or the highly replaceable word in the dataset. This finding proofs the potential of embedding network in helping data analyst conduct dataset quality assurance checking. The group five and six are also good examples of showing the semantic connection between closely located words. For instance, in many of FM contracts, the "certified" asset is always followed by clauses of "warranty". While in group five, it also indicates the potential opportunity for data analysts to learn certain "FM knowledge" from the good clustering example. In practice, it is very hard for junior data analysts, who do not

have the FM related experience, to link the "csa" with "certification". However, in group five's example, given the help of the diagram, it is easier for people to link "csa" and "certified" together as they are close to each other. Furthermore, group seven demonstrates the unsupervised neural network's ability to cluster the same type of adjective of building asset into one group ("colour": "black", "white", or "grey"). In the end, the group eight is used to demonstrate that although the "cluster" could find the certain 2-D pattern given embedding space, there are still many "not easily classified" words that are left along without proper grouping. But overall, the result indicates that there is a lot of useful knowledge can be learned from the well-clustered pattern, and these patterns will help FM personals to better understand the character of its dataset. The result of point clouds diagram also indicates that the vectorised representation of words can carry the semantic relationship of Home Depot dataset. Therefore, these vectors can help setting-up the default word embedding value used for CNN classification network.

RESULTS OF CNN TEXT CLASSIFIER

In this experiment, the text classification problem was solved by using pre-trained word embedding and a convolutional neural network.

Basic procedures of text classification processes are summarised as follows:

- The first step is to convert all text samples in the HCP dataset into sequences of word indices. The "word index" here refers to an integer ID that represents the sequence of a word in dictionary built based on HCP dataset. In this step, parameters like the number of words considered in the dataset and the maximum length of words in a word document are set.
- The second step of this programme is to prepare an "embedding matrix" that stores embedding word vectors for words in the "word index" that prepared in the word2vec embedding model (e.g. vector i corresponding to the word of index i in the "word index").
- The next step is to set-up an Embedding layer and load the embedding matrix into this layer. It is noticed that the weights of the embedding layer are set to be frozen, therefore, the weights of the embedding layer will not be updated throughout the training.
- Finally, a convolutional neural network is built with 32 output categories.

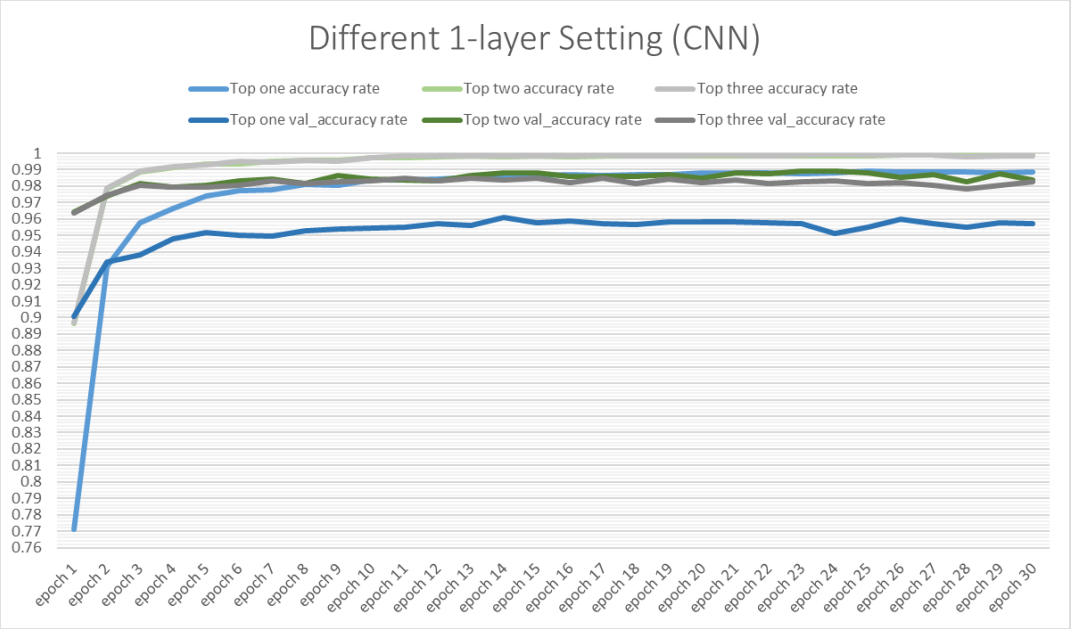
RESULT OF THE BEST TUNED PARAMETER SETTING FOR THE CNN TEXT CLASSIFICATION

In this experiment, many different settings were tested. Here is the summary of the different testing results:

- For the testing of different activation function settings, performances of tanh, softplus, sigmoid, relu, elu, and linear activation function were compared. linear activation function seems to perform better in both training and validation datasets.
- For the testing of different batch sizes for convolutional layers, performances of different batch sizes were tested (from batch size 1 to 64). The batch size 1 turns out to achieve the best performance.
- For the testing of different filter sizes for convolutional layers, performances of different filter sizes were compared (from filter size 4 to 64). The filter size 64 leads to the highest accuracy learning rate.
- For the testing of different kernel sizes for convolutional layers, performances of various kernel sizes were compared (from kernel size 2 to 7). In this case, kernel size 5 turns out to be the most suitable kernel size for CNN layer.
- For the testing of different layer settings, performances of several swallow layer CNN structures were tested. The results show the performance of different layer settings in this case is not important, as they both end-up having the very close accuracy rate. However, the single CNN layer can provide relatively faster learning speed.
- For the testing of different Pooling layer settings, performances of average-pooling and max-pooling layer were compared. As a result, the max-pooling method outperforms the average-pooling method in bother accuracy growth rate and steadiness.
- For the testing of different types of optimizers, performances of RMSprop, SGD, Adagrad, Adadelta, Adam, Adamax, and Nadam optimizer were tested. As expected, Nadam optimizer outperforms the rest of the optimizers by having the fastest accuracy growth rate, followed closely by RMSprop optimizer and Adam optimizer. However, RMSprop optimizer has a more stable learning curve.

As a result, we finalized our optimized setting as follows: a 1-layer convolutional layer structure with the following parameters: 1-batch size; 64-filter size; 5-kernel size; linear activation function; maxpooling; and RMSprop optimizer.

Figure.4: The accuracy learning curve of the 1-layer CNN model based on training data with different accuracy measure methods



For evaluating the performance of our CNN model, we use three different kinds of accuracy rating, as shown in Figure.4. The "top one accuracy rate" is the accuracy rate of the classifier to correctly predict the category according to the input asset description, by only taking the highest predicted category label. Similarly, the "top two and three accuracy rates" indicate accuracy rates, when the corrected category label exists within two or three of the highest probability classes. For instance, for the top three accuracy rate, if the corrected predicted label is ranked the second or third in their probability ranking, we still treat as it predicts the correct class. The reason for using these different accuracy ratings is because, in practice, even if the predicted label cannot correctly provide the asset label, the model can still help data analyst to conduct the quality assurance process by providing the other potential options. For the top one accuracy rating measurement, the validation accuracy of the CNN classification model ends up at 95.74%, which is much higher than the manual labelling accuracy rate at 65.91% that achieved by data analyst given the BCIS coding manual guideline. The whole classification process takes about 13 minutes to finished which is neglectable compared with the duration of manual classification.

CONCLUSION

Overall, results from this research indicate that the deep learning network can classify building assets according to BCIS NRM level-3 group with a considerable high accuracy rate (with top one accuracy rate at 95.79%, top two accuracy rate at 98.03%, and top three accuracy rate at 98.21%) over the test dataset. The outcome of "word2vec" embedding experiment also shows that the Unsupervised Skip-gram Gradient Descent Model can cluster the words in the document into different groups and the embedding "word2vec" can be utilised in the CNN Classifier. The result confirms that the CNN Classifier can achieve better accuracy performance than the manual classification done by junior building data analysts at 65.91% with a shorter classification time at only 13 mins rather than the hours of manual classification process previously. Furthermore, the classification accuracy of this experiment is also higher than the Dave's experiment result (at 76 %). Therefore, it can be concluded that the ML based auto-classification process can potentially provide AEC/FM with a timely solution in tackling with data interoperability and standardization problem when doing the strategic asset management. By improving the accuracy of classification and providing the asset information with more "cheap" labels, facility data analyst and managers can easily converter the data and information from asset surveyors and consultants into their preferred cost analysis format. So that the price for utilizing data from different projects is reduced, which leads to a better interoperability for the facility management's data environment.

Although both two objectives of this experiment have been achieved. There still some limitations in this experiment. Firstly, since this research is conducted as a pilot study to test the applicability of the text classification over asset defect survey dataset, the text classification method and the classification dataset used is relatively limited. Therefore, in future studies, more ML based methods can be tested to compare the performance of different ML methods and more project datasets should be tested to further validate the reliability of the ML based text classifier over building asset portfolio. Besides text-based asset description, there are also other different types of asset information that can be further utilised (e.g. photo data, numerical data). Can theses data also be further utilised to speed up other processes or provide more intelligent solutions in the other AECFM fields?

REFERENCES

Aleksandr, D., Anna, G. and Satoshi, M. (2016). Word Embeddings, Analogies, and Machine Learning: Beyond King - Man +Woman = Queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp.3519–3530.

Anon, (2017). *Construction 2020: A Vision for Australia's Property and Construction Industry*. [online] Available at: http://www.construction-innovation.info/images/pdfs/Construction_2020.pdf [Accessed 24 Aug. 2017].

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Chen, J. et al., 2009. Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3 PART 1), pp.5432–5435. Available at: http://www.nlp.org.cn. [Accessed September 12, 2018].

Dave, K., Lawrence, S. & Pennock, D.M., 2003. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*, Available at: http://www.cs.cornell.edu/people/pabo/movie- [Accessed November 15, 2018].

El-Saboni, M., Aouad, G. and Sabouni, A. (2009). Electronic communication systems effects on the success of construction projects in United Arab Emirates. *Advanced Engineering Informatics*, 23(1), pp.130-138.

Gallaher, M., O'Connor, A. and Dettbarn, J. (2004). *Cost analysis of inadequate interoperability in the US capital facilities industry*. NIST Report No. GCR 04-867. US Department of Commerce Technology Administration, National Institute of Standards and Technology.

Harris, Z. (1954). Distributional structure.

Hartmann, T., Gao, J. and Fischer, M. (2008). Areas of Application for 3D and 4D Models on Construction Projects. *Journal of Construction Engineering and Management*, 134(10), pp.776-785.

Hinton, G. and Roweis, S. (2002). Stochastic Neighbor Embedding. *Advances in Neural Information Processing Systems*, (15), pp.833–840.

Home Depot. (2016)., Home Depot Product Search Relevance | Kaggle. Available at: https://www.kaggle.com/c/home-depot-product-search-relevance/data [Accessed November 16, 2018].

IAM, 2016. Capital Investment, Operations and Maintenance Decision Making.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp.1746-1751.

Kobayashi, V.B. et al., 2018. Text Classification for Organizational Researchers: A Tutorial. *Organizational Research Methods*, 21(3), pp.766–799. Available at: http://journals.sagepub.com/doi/pdf/10.1177/1094428117719322 [Accessed August 24, 2018].

Kursat Uysal, A. & Gunal, S., 2014. The impact of preprocessing on text classification. Available at: http://dx.doi.org/10.1016/j.ipm.2013.08.006 [Accessed September 12, 2018].

Lee, G., 2011. What Information Can or Cannot Be Exchanged? *Journal of Computing in Civil Engineering*, 25(1), pp.1–9. Available at: https://ascelibrary-org.libproxy.ucl.ac.uk/doi/pdf/10.1061/%28ASCE%29CP.1943-5487.0000062 [Accessed August 16, 2018].

Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, pp.2579-2605.

Mikolov, T., Yih, W. and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*.

Moschitti A, Basili R.. 2004 Complex linguistic features for text classification: A comprehensive study[C]//European Conference on Information Retrieval. Springer, Berlin, Heidelberg, pp.181-196.

Phan, X.-H., Nguyen, L.-M. & Horiguchi, S., 2008. *Learning to classify short and sparse text & web with hidden topics from large-scale data collections*, Available at: http://portal.acm.org/citation.cfm?doid=1367497.1367510 [Accessed September 12, 2018].

Rezgui, Y. and Zarli, A. (2006). Paving the Way to the Vision of Digital Construction: A Strategic Roadmap. *Journal of Construction Engineering and Management*, 132(7), pp.767-776.

Royal Institution of Chartered Surveyors, 2012a. NRM 1: Order of cost estimating and cost planning for capital building works. In *NRM 1: Order of cost estimating and cost planning for capital building works*. p. 8. Available at: www.ricsbooks.com [Accessed August 16, 2018].

Royal Institution of Chartered Surveyors, 2012b. *Elemental Standard Form of Cost Analysis. Priciples, Instructions, Elements and Definitions*, Available at: https://www.rics.org/Global/BCIS_Elemental_Standard_Form_of_Cost_Analysis_4th__NRM__Edition_2012.pdf [Accessed June 14, 2018].

Royal Institution of Chartered Surveyors, 2013. NRM 2: Detailed measurement for building works. In *NRM 1: Detailed measurement for building works*. p. 2. Available at: www.ricsbooks.com [Accessed August 16, 2018].

Shen, W., Hao, Q., Mak, H., Neelamkavil, J., Xie, H., Dickinson, J., Thomas, R., Pardasani, A. and Xue, H. (2010). Systems integration and collaboration in architecture, engineering, construction, and facilities management: A review. *Advanced Engineering Informatics*, 24(2), pp.196–207.

Toman, M., Tesar, R. & Jezek, K., 2006. Influence of word normalization on text classification. *Proceedings of InSciT*, (January 2006), pp.354–358. Available at: http://www.kiv.zcu.cz/research/groups/text/publications/inscit20060710.pdf.

Vo, D.T. & Ock, C.Y., 2015. Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications*, 42(3), pp.1684–1698. Available at: http://dx.doi.org/10.1016/j.eswa.2014.09.031 [Accessed September 12, 2018].

Waterhouse, R. & Philp, D., 2016. National BIM Report. *National BIM Library*, pp.1–28.

Email contact: zi.fang.15@ucl.ac.uk