Occupancy Sensing in Buildings through Social Media from Semantic Analysis

Xing Lu
Student Member ASHRAE

Fan Feng
Student Member ASHRAE

Zheng O'Neill, PhD, PE
Member ASHRAE

ABSTRACT

Building occupancy behavior is becoming increasingly important and playing a significant role in the modeling of building energy performance, intelligent operations of building systems, and design of the future building system. Image-based, threshold and mechanical, motion sensing, and radio-based sensing are commonly used for occupancy sensing in buildings. However, those sensing technologies suffer from high cost, inevitable sensor errors, and scalability issues. Thus, they are not widely implemented in buildings. With the vast development of information technologies in the era of the internet-of-things (IoT), occupant sensing and data acquisition are not limited to traditional approaches. Social media could provide new near-real-time data sources that might contain occupancy information in space and in time. Utilizing public APIs provided by social media services, it is possible to attain the geographic information through either geo-tagged posts from Twitter or Facebook check-in messages. These datasets explicitly indicate the occupant presence and could be used to estimate the occupancy. However, it is well known that most social media users probably are not willing to disclose their location information. To increase the volume of the social media datasets, this paper provides a methodology to detect the implicitly geo-tagged posts that hold valuable occupancy information through a text classification and semantic analysis. The implicitly geo-tagged posts refer to the posts that could be inferred for the human occupancy information, but the user does not add the location to the posts. In this framework, first, the data acquisition tool is utilized to obtain the Tweets containing the textual location information such as a museum, a restaurant, etc. Then, training and evaluation datasets are pre-selected respectively with labels regarding whether the posts contain relevant occupancy information. The Word2Vec, a word embedding algorithm, is used to convert the text to the vectors as the input to the classification model, alongside with some Tweet-content-based features. In addition, several popular classification machine learning techniques are adopted to train the model. The preliminary results show that the proposed methodology could detect the implicitly geo-tagged posts from social media with relatively high classification performance. We also presented a preliminary result of a typical occupancy schedule to be used by the building energy models based on social media data. Results from this preliminary study will lay the foundation of the occupancy sensing through the social media data mining, which aims to provide another data source for occupancy sensing in buildings.

INTRODUCTION

Occupancy behavior in buildings is becoming an important topic of research as building systems become more sophisticated, and people spend more time in buildings (Hong et al. 2016). The occupancy behavior largely impacts modeling of building energy performance, operations of intelligent building systems, and design of the future building system, making occupancy behavior one of the leading influencers of energy consumption in buildings (Yan et al. 2015). In the ASHRAE HVAC Application Handbook (Owen 2015), the occupancy schedule in commercial buildings is in the form of a static schedule for the HVAC design and sizing. However, it is well known that the static occupancy schedule may cause the discrepancy between the design value and actual operational value, in terms of peak load, peak

Xing Lu is a PhD student in the Department of Mechanical Engineering, the University of Alabama, Tuscaloosa, AL. Fan Feng is a PhD student in the Department of Mechanical Engineering, University of Alabama, Tuscaloosa, AL. Zheng O'Neill is an Associate Professor in the Department of Mechanical Engineering, The University of Alabama, Tuscaloosa, AL.

Published in ASHRAE Transactions, Volume 126, Part 1

load occurrence moment, and total load. To accurately sense the occupancy, various types of occupancy sensor (such as image-based, threshold and mechanical, motion sensing, and radio-based sensing) are commonly used in buildings (Dong et al. 2019). However, those sensing technologies suffer from high cost, inevitable sensor errors, and scalability issues. Thus, they are not widely implemented in buildings. With the vast development of information technologies in the era of the internet-of-things (IoT), occupant sensing and data acquisition are not limited to traditional approaches.

The prevalence of social media platforms generates a myriad of publically available social media data, which could potentially provide new near-real-time data sources that might contain occupancy information in space and in time. Utilizing public application program interfaces (APIs) (TwitterDeveloper 2019) provided by the social media services, it is possible to attain the geographic information through either geo-tagged posts from Twitter or Facebook check-in messages, which is depicted in Figure 1(a) and (b). These datasets explicitly indicate the occupant presence and could be used to estimate the occupancy. However, it is well known that most social media users probably are not willing to disclose their location information. Although the datasets from the explicitly geo-tagged posts could be insufficient to represent the occupancy information, the implicitly geo-tagged posts could be a workaround as another social media data source for occupancy sensing. These geo-tagged posts are those that could be inferred for the human occupancy, but the user does not add his/her location to the posts. Figure 1(c) and (d) show two examples of the implicitly geo-tagged posts. We could infer from the Tweet textual semantics that the user is currently in the building, that is, the Art Institute of Chicago. However, there are some cases that the users mentioned the detailed location in the post, but they are apparently not present at a certain location. For examples in the following posts: 'I've always wanted to go to the Art Institute of Chicago. # bucketlist'; 'Hotels near the Art Institute of Chicago https://www.govisitchicago.com/top-hotels-near-art-institute-chicago/.'

a) Geo-tagged posts b) Check-in posts c) Implicit geotagged posts example 1 d) Implicit geotagged posts example 2

Figure 1 Explicit geotagged posts: (a) Geotagged posts (b) Check-in posts;

Implicit geotagged posts (c) Example 1 (d) Example 2.

Text classification and semantic analysis could be utilized to help us identify the right implicitly geo-tagged posts, which contain the occupancy information. Text classification problems have been widely used and addressed in many real applications such as information retrieval, sentiment analysis, recommender systems, etc. (Kowsari et al. 2019). In this case study, to increase the volume of the social media datasets in the building occupancy applications, we present a methodology to detect the implicitly geo-tagged posts from the social media that hold valuable occupancy information to sense the occupancy in buildings at the building level. We also presented a preliminary result of a typical occupancy schedule to be used by the building energy models based on social media data. The paper is organized as follows. Section 2 shows the methodology of the study and Section 3 presents a case study following the method. Section 4 summarizes the key takeaways.

METHODOLOGY

The proposed methodology to sense occupancy through social media has four key elements: data collection and pre-processing, feature generation, classifier formulation, and result evaluation, as illustrated in Figure 2. We will briefly describe each element in this section. Readers may refer to references for more details regarding the principle and implementation of the text classification (Aggarwal and Zhai 2012; Kowsari et al. 2019; Shivam 2018).

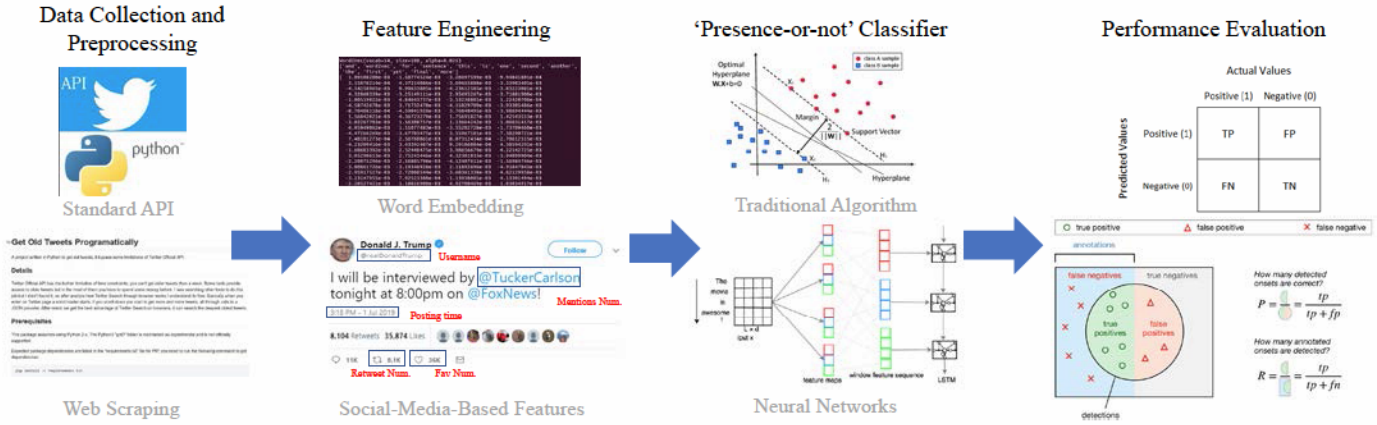

Figure 2 Schematics of the proposed methodology.

Data Collection and Preprocessing

In this section, we will present the method of collecting and cleaning social media datasets. One way of collecting the data is through the social media official API. The U.S. social media giants Twitter, Facebook, Reddit all have their proprietary APIs. However, this approach has some limitations for free and standard users. Take the Twitter Standard Search API as an example; the free standard tier allows the return of at maximum 100 relevant Tweets in the seven days. The data fidelity is incomplete compared to the paid categories. The paid access could allow the developer access to the full-fidelity data from as early as 2006, along with direct account management support, and dedicated technical support to help with an integration strategy. Another way of collecting data is through web-scraping. As aforementioned, official APIs have the limitation of time constraints; therefore we cannot get tweets older than a week. However, the web-scraping tools such as GetOldTweets (Jefferson 2017) could provide us history posts. The basic underlying principle is summarized as follows. When we enter on the Twitter page, a scroll loader starts. If we scroll down, we begin to get more and more tweets. The GetOldTweets tool exactly mimics this process. In this way, we could get the best advantage of Twitter Search on browsers and deeply search the oldest tweets.

All data needs to be cleaned before the feature extraction and feeding to the classifier, which can help to reduce the noise in the text data. Most text data from social media contain many unnecessary words such as stop words, misspelling, slang, etc. Many text-processing techniques are suggested such as tokenization, stop word elimination, case lowering, slang and abbreviation paraphrase, spelling correction, stemming, lemmatization, etc.

Feature Engineering

In this section, we will discuss the selection of feature extraction techniques. In this step, the raw text data will be transformed into the feature vectors, and different categories of feature vectors will be combined to help improve the accuracy of the classifier. The main features could be categorized into (1) Weighted words, (2) Word embeddings, and (3) Social media-based features. For the weighted words feature extraction, the Bag-of-Words (BoW) model (Wallach

2006) and Term Frequency-Inverse Document Frequency (TF-IDF) are two basic approaches. They are easy to compute. However, they do not capture the position and the meaning in the text. A word embedding is a form of representing words and documents using a dense vector representation. The location of a word within the vector space is learned from the text and is based on the words that surround the word when it is used. Word embedding models could capture the semantics of the word and each word will be mapped to an N dimension vector of real numbers. Word2Vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014), and FastText (Bojanowski et al. 2017) are the three most common pre-trained models to keep the syntactic and semantic information of each sentence. Apart from the pre-trained word embedding, we could also learn the word embedding as a part of fitting a machine learning model.

Social media-based features are statistical features based on the characteristics of the social media posts such as the presence of URLs, the presence of hashtags, hashtag count, favorite count, repost count, etc. Different combinations of these features will be fed into a classifier.

Classification Models

In this section, we will outline the classification techniques for the text classification and semantic analysis. The Naïve Bayes, logistic regression, and support vector machines (SVMs) are traditional but still commonly used for classification. Bagging models (such as a random forest model) and boosting models (such as XgBoost Model) are the ensemble models based on the tree-based models. Different neural network architectures are well used in the text classification applications, such as shallow neural networks and deep neural networks (e.g., Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Hierachical Attention Networks (HANs)). Selection of the classification models and tuning the model parameters are both critical to get the best fit model.

Evaluation

The evaluation metrics of the text classifiers measure the performance of making the right classification decision from different methods. Generally, four metrics are widely used: accuracy, precision, recall, F1-score based on the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), as illustrated in Eq. (1)-(4). The significance of these four elements may vary based on the classification application. It is noted that compared to the accuracy, the last three metrics are more meaningful in terms of the effectiveness of the text classifiers because the accuracy is insensitive to variations in the number of correct decisions due to the large value of the numerator (TP+TN) (Kowsari et al. 2019). In this case study, the negatives (people-not-present labels) account for a relatively large fraction of datasets, which may cause issues on the performance of the classifier. To decrease the impact of the large denominator (TN), we could eliminate some negative-labeled data to make a large proportion of the positive-labeled data in the whole data corpus.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

CASE STUDY: THE ART INSTITUE OF CHICAGO

To verify the feasibility of the proposed methodology, we use a museum building, the Art Institute of Chicago as a case study building. The Art Institute of Chicago, founded in 1879 and located in Chicago's Grant Park, is one of the oldest and largest art museums in the United States. It opens daily from 10:30 a.m. to 5:00 p.m. except on Thursdays until 8:00 p.m. The reason why we select a museum building in this preliminary study because people have a higher probability of making posts about their visits through social media and thus, we could obtain more datasets. First, we will discuss the details regarding the data collection and preprocessing. Then the selection for different features and classifiers are detailed, followed by the subsection of results and discussions.

Data Collection and Preprocessing

We collected history posts in April and May 2019 using the Twitter official API (TwitterDeveloper 2019) approach and the counterpart using the GetOldTweets approach (Jefferson 2017). We searched the relevant Tweets using the keywords 'art,' 'institute,' 'Chicago.' We compared the data from the two data sources and found that the data are smaller in the second approach. However, we found that the data from GetOldTweets neglected the retweet posts and the posts that are existing in history. The others are the same for these two methods because the retweet posts and the existing posts in the history will normally not indicate the presence of the people, it would be good to use the GetOldTweets approach since it is free of charge and have a similar data volume.

On top of that, we collected all the available ~30,000 history posts from December 2016 to June 2019 using the GetOldTweets approach, and manually labeled the latest 3,000 history posts which indicated whether the user was present or not. It is found that the positive (people-presence) data only occupies ~15% of the labeled data. To balance the proportion of the true positives and true negatives, we are using all the true positives. The total number of the training and validation datasets is 1,000. For the data preprocessing, we lower the case of the posts, conduct the tokenization, and then remove the stop words.

Feature Selection and Classifier Formulation

For the feature selection, we generated the word embeddings using pre-trained Word2Vec, where each word is presented by a high dimension vector. The dimension of the vector space is 300. For each Tweet, the aggregated vector is weighted by the value of the TF-IDF.

Table 1. Summary of the selected features

Categories	Selected Features	Type	Dimension Num.
Word embedding	300-dimension word embeddings weighted by the TF-IDF	Float	300
Tweet-content	Whether the posted time is within opening hours	Binary	1
	Whether the username of users contain the keywords like 'art'	Binary	1
	Whether the domain name in the Url contains check-in apps	Binary	1
	Hashtags in the Tweet Count	Integer	1
	Mentions in the Tweet Count	Integer	1
	Favourite/like Count	Integer	1
	Retweet Count	Integer	1
	Posted Hour	Integer	1
	Posted Day of the Week	Integer	1

In addition, we also considered social media content-based features. The posted time is a critical feature because the valid 'presence' posts must be made within the range of opening time. Many Tweets are synchronized from other applications such as Facebook, Swarmapp, Artic, Foursquare, etc. The domain name with check-in app 'Swarmapp' could have more probability for the people presence rather than art institutes application such as 'artic.' Therefore,

Published in ASHRAE Transactions, Volume 126, Part 1

whether the domain name is a check-in application name could be an important feature. The counts of favorites, retweets, hashtags, and mentions could also be essential features for identifying the features. When visiting and making a post in a museum, people may mention some official accounts and persons of significance to share the joy and findings. In addition to the aforementioned features, the username of the users could also be a critical feature. For example, some users are official accounts and they would not normally make a check-in post. Therefore, we check if the usernames have strings such as 'art,' 'archeo,' 'Chicago,' 'museum,' etc. Finally, combining the word embeddings and the other selected Tweet-content-based features, we select 309-dimension vectors for each data point. Table 1 shows a summary of the selected features.

For the classifier formulation, we tested the performance of different categories of classifiers. We selected the SVM (a traditional classifier), the random forest (an ensemble classifier), and the shallow neural network (that contains three types of layers). The training/testing data ratio is 8:2.

Result Analysis and Preliminary Occupancy Schedules

The performance metrics of different classifiers are listed in Table 2 in terms of accuracy, precision, recall, and F1-score. It can be seen that the accuracy of the different classifiers is in a similar range, with the random forest and neural network slightly being higher. This is as expected because we have a large number of true negatives when calculating the accuracy using Eq. (1).

Table 2. Summary of the performance metrics of different classifiers

Performance Metric	SVM	Random Forest	Shallow Neural Network
Accuracy	0.8485	0.9091	0.9091
Precision	0.6000	0.8333	0.7500
Recall	0.8571	0.7143	0.8571
F1-score	0.7059	0.7692	0.8000

As mentioned in the last section, the precision and recall are more meaningful in the evaluation of the effectiveness of the text classifiers. Although the Random Forest has a relatively high score of precision, it has a lower score of the recall score. This means the classification algorithm could not recognize the 'presence' of the user and label it as the 'not present.' Since we need to know the number of the valid presence of the people in buildings, it is desirable to see a higher recall score. In terms of the F1-score, the neural network performs the best with a score of 0.8. F1-score is an overall metric combining the precision and the recall. It can be seen that the neural network performs slightly better than the other two classifiers.

In Table 3, the detailed classification results of the testing sets are presented using the shallow neural network. The labels '1' and '0' represent the status of the people-present (positive) and people-not-present (negative). Majorities of the labels belong to be '0' (i.e., people not present). For the labels '1', the results show that the method could basically distinguish them from most of the '0' labels (not-present labels). Indexes 59 and 168 were mislabeled, but their prediction scores are above 0.1. In addition, Index 96 was mislabeled to be '1' although they should be '0'. The Tweets that are easy to be semantically differentiated such as Indexes 198 and 30 have a high prediction score.

Table 3. Demonstration of the classification results using the shallow neural network

Index	Posted Time	Tweets	Label	Prediction	Score
136	4/1/2019 3:20	Art Institute of Chicago will be hosting Gregg...	0	0	0.00257
139	3/19/2019 4:03	The Art Institute of Chicago is hosting Every...	0	0	0.02752
198	2/8/2019 11:27	I'm at The Art Institute of Chicago - @ artins...	1	1	0.60308
59	3/12/2019 15:38	Art Institute was amazing! # rembrandt # beaut...	1	0	0.16830
96	3/30/2019 16:43	Cut Piece, de Djanira. Performance, Art Instit...	0	1	0.63309
23	3/27/2019 10:48	Hopper @The Art Institute of Chicago https://w...	1	1	0.50097

30	3/28/2019 13:38	I'm at The Art Institute of Chicago - @ artins...	1	1	0.73036
54	3/20/2019 5:10	Can't wait to see this babe in May. # wcw # tr...	0	0	0.00174
39	4/7/2019 16:01	Got to spend an afternoon this weekend with Va...	1	1	0.70794
66	4/4/2019 21:00	Criticized for Failing to Consult Indigenous G...	0	0	0.02370
67	3/28/2019 8:21	Thanks so much for this Art Institute of Chica...	0	0	0.05156
88	3/25/2019 14:43	Wall-Floor Positions, de Gustave Klimt. Video,...	0	0	0.04427
63	4/12/2019 8:10	School of the Art Institute of Chicago has nam...	0	0	0.27143
168	3/14/2019 15:15	A. Lincoln # artinstituteofchicago # chicago #...	1	0	0.48824
86	3/24/2019 9:43	Autorretrato aos 13, de Giotto. Desenho, Art I...	0	0	0.04016
184	3/23/2019 1:00	The Art of Reading at the Art Institute of Chi...	0	0	0.01482
55	4/4/2019 10:16	Art Institute of Chicago delayed exhibition of...	0	0	0.00174
25	3/27/2019 11:33	Art museum I'm ready to come home tbh. Work to...	1	1	0.65585
72	3/13/2019 0:13	I Like America and America Likes Me, de Alexan...	0	0	0.00103
158	4/9/2019 21:12	Art Institute of Chicago where Swami Ji delive...	0	0	0.01881
60	4/3/2019 14:40	In a move museum leadership is calling unprec...	0	0	0.07430
110	4/9/2019 20:13	I'm too sad to tell you, de Joseph Beuys. Vide...	0	0	0.03527
199	3/18/2019 11:22	@ JohnMu Just about every result page for the ...	0	0	0.03667

While the results from this preliminary study show the feasibility of the proposed methodology to sense the occupancy in buildings at the building level, a lot of future work could be done to improve the result of the overall framework. The future and ongoing work are listed as follows. First, there is a need to have more labeled datasets, especially positive data. The currently labeled datasets are likely insufficient to contain comprehensive circumstances. The performance would be enhanced with more labeled data because the performance of the more complex neural network will manifest itself with larger datasets. Second, there is a need to explore more delicate text cleaning, which could help to eliminate more noise for the word embedding and classifier. Third, there is a need to investigate better word embedding algorithms. We are using the pre-trained Word2Vec from Google, and we encountered the circumstance that the words are not in the corpus. Recent contextualized word representations could be a workaround to enhance the word embedding. Fourth, there is a need to examine more powerful classification models and the hyper-parameter tuning in the classifier models. The deep neural network models such as CNN and RNN are believed to have a better performance in coping with the natural language processing. Also, several network parameters should be fine-tuned to get the best fit model by a hyper-parameter tuning. Finally, there is a need to study the uncertainty from labeling the data. We label the data based on the factors such as human recognition of the textual content, the posted time vs. the opening time, the posted pictures along with the Tweets. However, there are cases that it is hard to identify whether people are present or not. As another example, the posts may occur within the normal hours of operation but not occur at the time of occupancy, which results in inaccurate timestamp data. Furthermore, the ratio of users to non-user of social media should be observed due to the age bracket of the social media users. Such uncertainty will all lead to inaccurate extracted occupancy schedules.

Using the formulated classifier, we labeled the historical ~30,000 Tweets. We added up all the positive labeled data in the same time slot (an hour) on the same day type using these historical Tweets. We assumed that people would stay for three hours before and after the posting time based on the visit duration from Google Map App. In this way, we obtained the weekly typical occupancy schedules to be used as inputs for the building energy models, as shown in Figure 3, which shows two different types of typical occupancy daily schedules. It is noted that this is only a preliminary result which needs further and more comprehensive investigations to alleviate the uncertainties associated with assumptions.

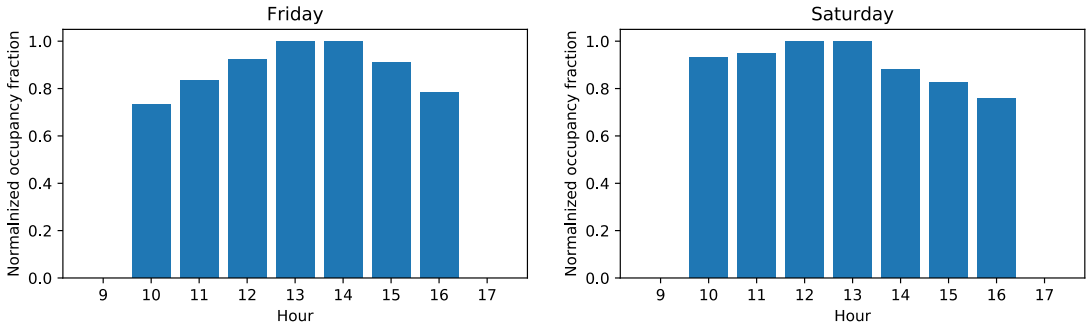

Figure 3 Preliminary typical occupancy schedules for two day types extracted from social media data.

CONCLUSIONS AND FUTURE WORKS

In this paper, we explored a methodology to perform occupancy sensing in buildings at the building level through social media using text classification and semantic analysis. The methodology consists of four elements, including data collection and pre-processing, feature generation, classifier formulation, and result evaluation. As a preliminary study, the Art Institute of Chicago is selected as the case study building. The pre-trained word embedding model Word2Vec is used for the semantic feature extraction along with other social-media-content-based features. Different types of classifiers such as SVM, random forest, and shallow neural network are selected and compared in terms of the classification performance. The results show that the presented methodology could detect the implicitly geo-tagged posts from social media with a relatively high accuracy and classification performance. The neural networks perform slightly better than the other traditional classifiers. We also list some experience on conducting this explorative research as well as some future work. Next steps include improving the proposed methodology, validating, and testing the occupancy schedule estimation based on social media data for the building energy performance modeling.

REFERENCES

Aggarwal, C. C., and C. Zhai. 2012. *Mining text data*: Springer Science & Business Media.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Dong, B., V. Prakash, F. Feng, and Z. O'Neill. 2019. A review of smart building sensing system for better indoor environment control. *Energy and Buildings, 199*, 29-46.

Hong, T., S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati. 2016. Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings, 116*, 694-702.

Jefferson, H. (2017). Get Old Tweets Programatically. from https://github.com/Jefferson-Henrique/GetOldTweets-python

Kowsari, K., K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. 2019. Text classification algorithms: A survey. *Information, 10*(4), 150.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. *Distributed representations of words and phrases and their compositionality*. Paper presented at the Advances in neural information processing systems. 2013

Owen, M. 2015. ASHRAE Handbook: HVAC Applications. *Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers*.

Pennington, J., R. Socher, and C. Manning. *Glove: Global vectors for word representation*. Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014

Shivam, B. (2018). A Comprehensive Guide to Understand and Implement Text Classification in Python. from https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

TwitterDeveloper. (2019). Standard search API. from https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

Wallach, H. M. *Topic modeling: beyond bag-of-words*. Paper presented at the Proceedings of the 23rd international conference on Machine learning. 2006

Yan, D., W. O'Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi. 2015. Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings, 107*, 264-278.