high sunlight. After the sunlight calculation is complete, designers or engineers decide whether the sunlight level for each region is good enough or not. If yes, designers start specifying plant species on each greenery region. Only plant species that are suitable for the sunlight condition of a greenery region are listed for designers or engineers to select. After that, designers or engineers send a request for estimating CO_2 capture quantity and checking if current design meets the requirement of regulation.

Suggested estimation method

To consider plant species, this study uses the IPCC estimation method. The IPCC method includes more data of plant species for reducing estimation uncertainty. For countries that have collected domestic plantation data, this method can further reduce the uncertainty of estimation results, and for those do not, IPCC provides default data. On a building site, the carbon stock change in soil can be ignored because most of the dead wood and litter are removed, and usually, there is no wood harvesting on a building site. Therefore, we can use the following equation to replace Eq. (1):

$$\Delta C = \Delta C_{AB} + \Delta C_{BB} \tag{6}$$

This equation calculates the annual carbon stock change in biomass on building sites. The other equations remain the same as Equation (4) and (5).

Sunlight analysis

For each greenery area, cumulative sunlight through a year is calculated and categorized based on plant data used. For example, Taiwan government provides an illustrated guide of Taiwan's native plants for green building design (ABRI, 2010). This guide contains suitable insolation range for each plant species and divides it into three levels, low, medium, and high (Table 2). The cumulative sunlight amount (kWh/ m^2) of a horizontal greenery area on the same building site without any shadow through a calendar year is set as 100% amount of sunlight. Several commercial software is available for solar sunlight analysis, such as Insight (Autodesk, 2018) and AECOsim (Bentley, 2018). Many studies (Salimzadeh et al., 2018) have described how to use a BIM model to conduct sunlight analysis.

Table 2. Suggested sunlight level

Sunlight level	Amount of sunlight (%)
High	> 70
Medium	40-70
Low	< 40

Plant database

For each country or region, an organization is needed to provide service of integrating CO_2 capture-related data along with photos and properties of plants. Figure 3 shows an example of the relationships of entity sets stored in plant database. For Taiwan, the Department of Forestry at National Taiwan University have studied domestic plant data, including CF, I_v, D, R, and BEF_I, for IPCC method (Department of Forestry and Resource Conservation, National Taiwan University, 2014). These data cover common plant species and need further study for more

species. The guide of Taiwan's native plants for green building design provides suitable insolation range, climate zone, and other properties of common native plant species. The default CO_2 capture data for plant types, which is provided by IPCC or Taiwan government, are stored in an individual database table. When there is no CO_2 capture data for a plant species, the default data for its type are used as an alternative.

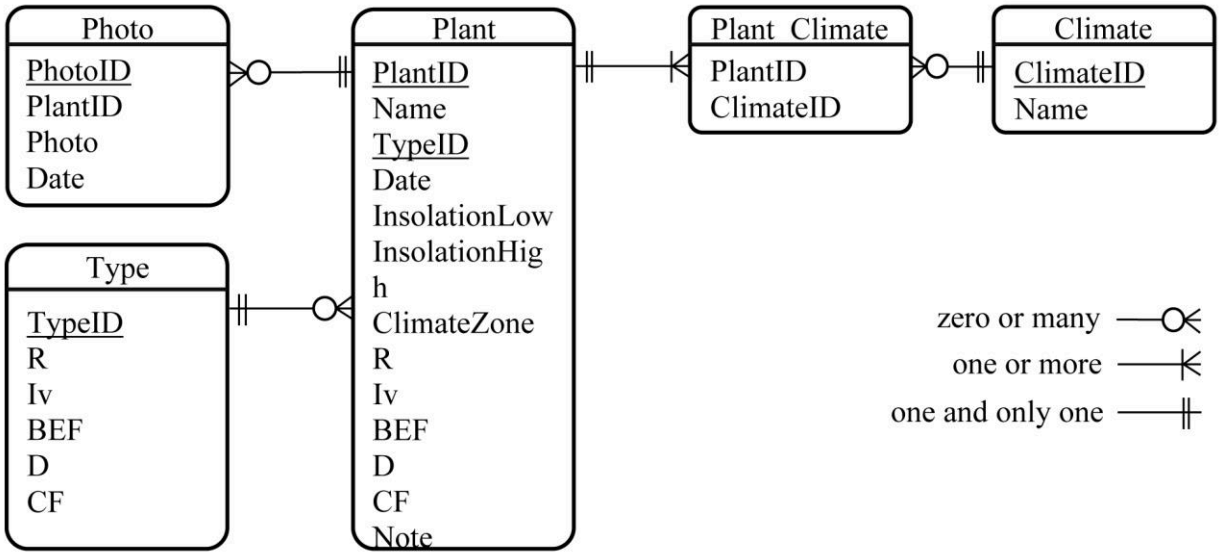

Figure 3. Entity-Relation diagram of plant database

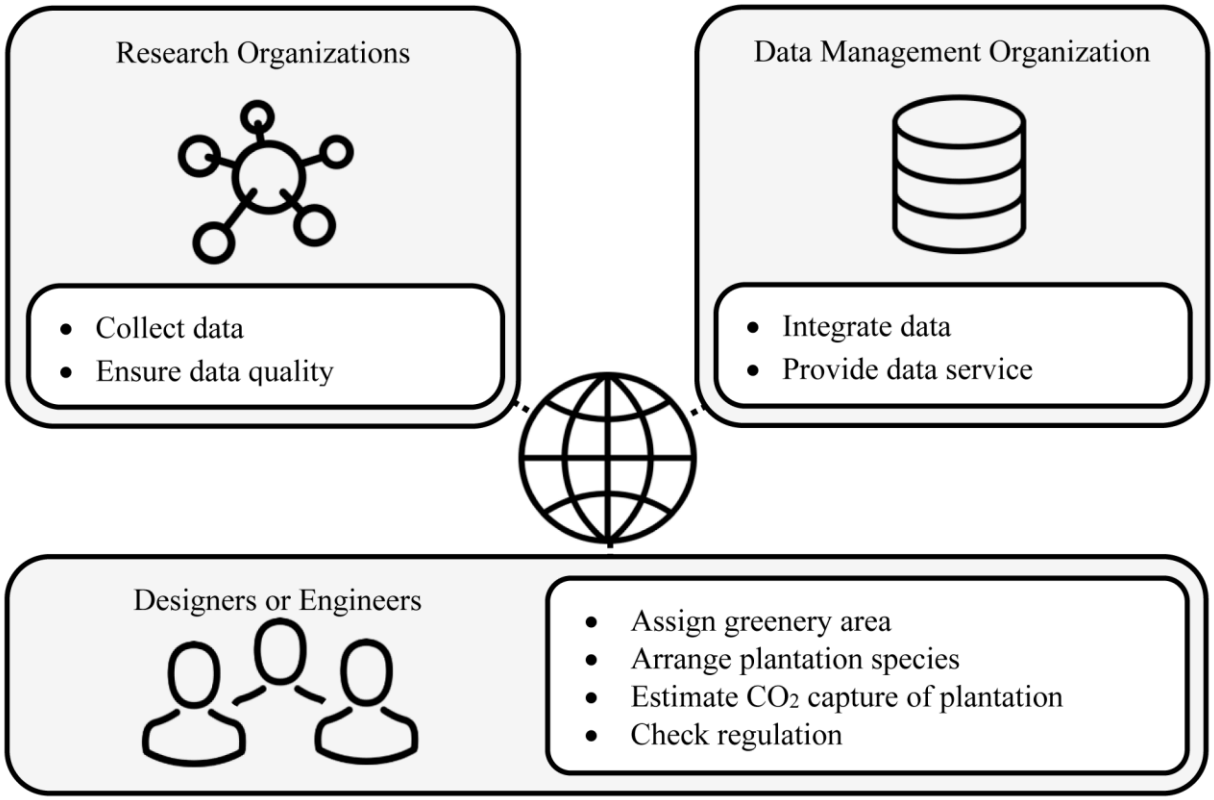

Figure 4. Roles and activities in BIM-enabled design method

Implementation of the proposed method

Figure 4. shows the roles and activities of the proposed design method. One or more research organizations are needed to be responsible for collecting CO_2 capture data of common plant species or types and ensure the quality of the data. To guarantee the consistency of plant data, a centralized data service is suggested so that designers and engineers within a country or region can always retrieve the most updated plant data for estimating CO_2 capture for designed green plantation.

CONCLUSION

In this paper, a BIM-enabled design method for green plantation on building sites is proposed for providing a more realistic estimation on CO_2 capture of plantation. This method takes advantage of sunlight simulation in 3D digital space so that designers and engineers can select suitable plant species for greenery region with different sunlight condition on building sites. The estimation method provided by IPCC is adopted because its formulations take into account more local plant data. A centralized data service providing designers and engineers consistent and up-to-date plant parameters of common plant species is recommended. This method not only enables designers and engineers to consider sunlight condition for choosing suitable plant species but also reduces the uncertainty of estimation results than current green building certification systems.

REFERENCES

ABRI (Architecture and Building Research Institute). (2017). *Green Building Evaluation Manual.* ABRI Press. New Taipei, Taiwan.

ABRI. (2010). *Illustrated Guide to Taiwan Native Plants for Application in Green Building Design*. ABRI Press. New Taipei, Taiwan (in Chinese).

Autodesk. (2018). Better Building Performance. Retrieved from Autodesk Inc., website: https://insight360.autodesk.com/oneenergy, accessed on December 1, 2018.

Bentley. (2018). AECOsim Energy Simulator. Retrieved from Bentley Systems Inc., website: https://www.bentley.com/en/products/product-line/building-design-software/aecosim-energy-simulator, accessed on December 1, 2018.

Department of Forestry and Resource Conservation, National Taiwan University. (2014). *Establishment of Forest Greenhouse Gas Inventory System and Trial Calculation in Accordance with MRV Principle* (translated). Department of Forestry and Resource Conservation, National Taiwan University. Taipei, Taiwan (in Chinese).

IPCC. (2006). *2006 IPCC Guidelines for National Greenhouse Gas Inventories: Volumes 1 and 4*. IGES. Hayama, Japan.

OMG. (2011). *Business Process Model and Notation (BPMN).* Object Management Group.

Sacks, R., C. Eastman, G. Lee, and P. Teicholz. (2018). *BIM handbook: a guide to building information modeling for owners, designers, contractors, and facility managers, Third edition*. John Wiley & Sons, Inc. Hoboken, New Jersey.

Salimzadeha, N., F. Vahdatikhakib, and A. Hammadc. (2018). BIM-based Surface-specific Solar Simulation of Buildings. Proceeding of 35th International Symposium on Automation and Robotics in Construction, Berlin, Germany, July 20-25, 2018.

Identifying Damage-Related Social Media Data during Hurricane Matthew: A Machine Learning Approach

Faxi Yuan[1] and Rui Liu, Ph.D.[2]

[1]Ph.D. Student, M. E. Rinker, Sr. School of Construction Management, Univ. of Florida, Gainesville, Florida. E-mail: faxi.yuan@ufl.edu
[2]Assistant Professor, M. E. Rinker, Sr. School of Construction Management, Univ. of Florida, Gainesville, Florida. E-mail: liurui@ufl.edu

ABSTRACT

Previous research used keywords like Hurricane Matthew/Sandy to filter the disaster- and damage-related social media data. However, various Twitter data containing these keywords were not describing the disaster events or their impacts. Meanwhile, machine learning demonstrates its potential for classifying social media data. Nevertheless, very limited existing research employs this approach for identifying damage-related social media data. This paper introduces the machine learning approach for identifying the damage-related social media data. Naïve Bayes, support vector machine (SVM), and decision tree are employed for training the classifier. The 10-folder cross-validation method is utilized for evaluating the performance of these three classifier models. Naïve Bayes model demonstrates the most reliable results. This paper provides a new solution for filtering the damage-related social media data during natural disasters. The manually annotated Twitter data can be used for classifying social media data in future disaster events.

INTRODUCTION

The real-time characteristics of social media data promote its popular applications for situation awareness such as damage assessment during the disasters (Yuan and Liu 2018a, 2018b; Kryvasheyeu et al. 2016). Specifically, the previous research commonly employs keywords such as disaster type and disaster event name to filter the disaster- and damage-related social media data (Ovadia 2009; Shelton et al. 2014; Zou et al. 2018). However, a large number of social media data containing keywords such as Hurricane Matthew is found to have no relationship with the disaster events and their impacts. An example tweet is "Snap crackle pop #hurricanematthew #aftermath #jax #fl @ Southeast Jacksonville https://www.instagram.com/p/BLTX7mqDTIn/". This tweet does not describe the disaster event or its impacts. Considering this kind of tweets into the disaster-related tweets can reduce the reliability of using the ratio between the number of disaster-related tweets and the number of general tweets to reflect the disaster impacts (e.g., Zou et al. 2018; Guan and Chen 2014). The advanced development of machine learning methods provides another solution to identify the disaster- and damage-related social media data.

Machine learning has been widely used for text classification in ACM (The Association for Computing Machinery) and IEEE (The Institute of Electrical and Electronics Engineers) fields (e.g., Kaur and Kumar 2015; Gonçalves et al, 2013; Forman 2003; Pang et al. 2002). Though machine learning demonstrates its potential in classifying social media data (Gu et al. 2016; Gal-Tzur et al. 2014), there is very limited existing research employing machine learning approach for identifying the damage-related social media data during natural disasters. To fill in this gap, this paper proposes to use a machine learning approach to identify the damage-related social

media data.

The damage-related social media data are the messages describing casualties or damage caused by a disaster (Imran et al. 2013). The 'damage' in this research can be damages on infrastructures (e.g., trees, roads, electricity, gas and water) and humans (physical injuries and psychological impacts). For example, "Downed trees and limbs at the house but the Schnauzers are safe. That's what really matters. https://www.instagram.com/p/BLR4fRIBK1e/" is a damage-related tweet.

This research establishes the process map for using the machine learning approach to classify the social media data into two classes, damage-related and not-related. This paper starts with manual annotation of the randomly selected Twitter data posted across Florida. Using the manually annotated Twitter data (with labels 'damage-related' or 'not-related'), we extract the TF-IDF (Term Frequency-Inverse Document Frequency) features of the two classes at 2- to 3-gram level. Thereafter, three classifier models including Naïve Bayes, Support Vector Machine (SVM) and Decision Tree are employed for training the classifier. The 10-folder cross-validation method is used for evaluating the performances of these three models. As a result, the most reliable classifier model is selected, and its performance is evaluated simultaneously.

RELATED RESEARCH

Social media analysis for damage assessment

The advanced development of crowdsourcing platforms promotes its popular applications in the disaster research fields (e.g., Yuan and Liu 2018c; Spence et al. 2016). Social media, as a crowdsourcing platform, is especially widely used by the scholars due to its well representativeness of US population (Wang and Taylor 2014) and its real-time characteristic (Sriram et al. 2010).

To be specific, recent studies employing social media data for situation awareness mainly fall in damage assessment (e.g., Guan and Chen 2014). The existing research mainly investigates the relationship between Twitter indexes and the disaster damages. The Twitter indexes include disaster-related ratio (e.g., Zou et al. 2018) and damage-related ratio (Yuan and Liu 2018a). The disaster- or damage-related ratios is the ratio between the number of disaster- or damage-related tweets and the number of general tweets. These studies employ the pre-defined keywords to filter the disaster- or damage-related social media data (Kryvasheyeu et al. 2016). A large number of Twitter data containing the pre-defined keywords such as Hurricane Matthew does not describe the disaster events or their impacts. To resolve this limitation, this paper introduces the machine learning approach.

Machine learning for social media data classification

Machine learning is widely used for text classification and text mining by the scholars from various research fields (e.g., Gu et al. 2016; Kaur and Kumar 2015). Machine learning is classified into supervised, unsupervised and semi-supervised approaches (Kaur and Kumar 2015). The difference between supervised and unsupervised machine learning approach is the need of training data. The former requires training data to train the classifier while the later does not need any training data (Gonçalves et al. 2013). The semi-supervised machine learning approach is the combination of the supervised and unsupervised approaches (Kaur and Kumar 2015).

The commonly used machine learning method for social media data classification relies on

the supervised classification approaches (Pang et al. 2002). One of the advantages of the supervised machine learning approach is its ability to adapt and create trained models for specific purposes and contexts (Gonçalves et al. 2013). To identify the damage-related social media data from the original Twitter data posted during Hurricane Matthew, this paper selects the supervised machine learning approach.

METHODOLOGY

To identify the damage-related social media data generated in Florida during Hurricane Matthew, this paper proposes the machine learning approach to achieve the classification of damage-related and non-related data. Three classifier models including Naïve Bayes, Support Vector Machine (SVM) and Decision Tree are introduced. The evaluations of their performances are conducted based on the precision, recall and F1-scores. The process map is presented in Figure 1.

Twitter data collection

To collect the Twitter data, the authors developed a web crawler based on the Twitter Advanced Search API. Specifically, this paper adopts the same study period as the authors' previous research (Yuan and Liu 2018a), from 4 October 2016 to 8 October 2016. With the web crawler, this paper collects the Twitter data posted by 67 counties across Florida during the study period. The county seats and the largest cities of these three counties, including Jefferson, Lafayette and Liberty cannot be identified in Twitter Advanced Search. This study collects 45,659 tweets for the other 64 counties in Florida.

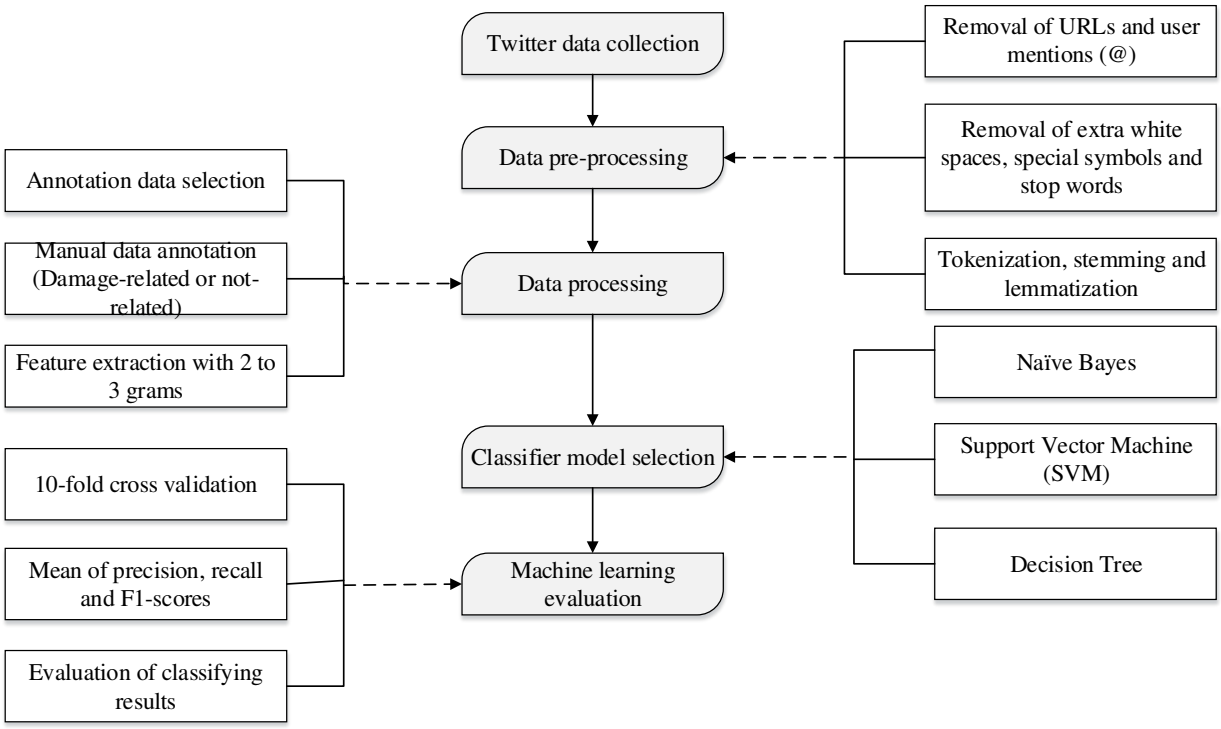

Figure 1. Process map for identifying damage-related social media data

Data pre-processing

The collected Twitter data (i.e., tweet) has the limit of 140 characters and massive amount of noise including slang words, misspelled words, short URLs and many others (Garg and Kumar 2016; Kaur and Kumar 2015). To increase the accuracy of classification process, this section conducts the following actions to pre-process the Twitter data with the algorithms of *sklearn* python library: 1) Remove of URLs and user mentions (i.e., @Twitter user name); 2) Remove extra white spaces, special symbols and stop words (e.g., 'a', 'an' and 'the'); and 3) Perform tokenization, stemming, and lemmatization. The tokenization process splits each tweet into words, lowercases the words and removes the punctuation. The stemming process reduces the words into their root form. The lemmatization process changes the words in third person to first person and changes the verb words in past and future tenses to present.

Data processing

This section starts with the selection of training data from the collected Twitter data. This paper selects 10% from the collected Twitter data for classifier training and test. Hence, 4,600 tweets are randomly selected and distributed to eight annotators for classifying the Twitter data into two classes, damage-related and not-related. The 'damage-related' tweets are labeled with '1' while the 'not-related' parts are labeled with '0'.

The eight annotators' feedback shows that Twitter users use 2 and 3 terms frequently to describe the damage-related contents such as "lose power" or "out of water". Therefore, this paper uses 2- to 3-gram level Term Frequency-Inverse Document Frequency (TF-IDF) vectors as the features. The TF-IDF is calculated as Eq. (1), (2) and (3). The TF-IDF vectors of the training Twitter data are further used as input for the classifier models.

$$\text{TF-IDF}(t) = \text{TF}(t) \times \text{IDF}(t) \quad \text{Eq. (1)}$$

$$\text{TF}(t) = \frac{\text{The number of times term t apperas in a document}}{\text{The total number of terms in the document}} \quad \text{Eq. (2)}$$

$$\text{IDF}(t) = \log \frac{\text{The total number of documents}}{\text{The number of documents with term t}} \quad \text{Eq. (3)}$$

Where, t refers to the term in the documents; a document is a tweet; TF(t) is the normalized term frequency; IDF(t) is the inverse document frequency.

Classifier model selection

Naïve Bayes is commonly used in social media data classification (Imran et al. 2013; Kaur and Kumar 2015). To identify the damage-related social media data, this paper introduces not only Naïve Bayes but also the other two classifier models, SVM and Decision Tree. Their performances are evaluated below.

Machine learning evaluation

The 10-fold cross-validation is a common method used for evaluating the performance of machine learning approaches (Mullen and Collier 2004; Imran et al. 2013). Therefore, this paper employs the 10-fold cross-validation method for evaluating the performances of the classifier models including Naïve Bayes, SVM and Decision Tree. To calculate the values of precision, recall and F1-scores, this research first defines the metrics for describing the compassion results

of predicted results and the manual label results in Table 1. Thereafter, the precision, recall and the F1-socres are calculated from the 10-fold cross-validation process. According to the mean of precision, recall and F1-socres for the three classifier models, this research selects the Naïve Bayes classifier model for our case and evaluate the machine learning performances.

Table 1. The metrics describing the compassion results of predicted results and the manual label

Predicted label	Manual label	
	Damage-related (1)	Not-related (0)
Damage-related (1)	True Positive	False Positive
Not-related (0)	False Negative	True Negative

RESULTS

Twitter data

This paper collects 45,659 general tweets posted during Hurricane Matthew across Florida. The geographic distribution of the tweets number is presented in Figure 2. The polygon colors, ranging from light green to red, represent that the number of general tweets changes from small to large values. Figure 2 shows large counties such as Miami-Dade (county seat: Miami), Hillsborough (county seat: Tampa), and Orange (county seat: Orlando) have large number of tweets. Considering cities acting as human activity centers (Chapin 1975), Twitter activities gathering in large counties in this case is reasonable (Yuan and Liu 2018a).

Data pre-processing results

This paper pre-process the Twitter data following the steps mentioned in 'Data pre-processing' section. An example is presented in Table 2.

Table 2. Data pre-processing example

Original tweet	Preparing extra #water... #hurricainematthew prep @ Bartram Park https://www.instagram.com/p/BLPDbhojh4X/
After pre-processing	'prepare' 'extra' 'water' 'hurricanematthew' 'prep'

Manual annotation of Twitter data

4,600 tweets are randomly selected for the training and test. They are manually labeled as '1' for the damage-related class and '0' for the not-related class. As a result, 505 tweets are annotated as '1' while the other 4,095 tweets are labeled with '0'. The direct use of the selected 4,600 tweets with labels can cause the Unbalanced Dataset issues due to the little size of class '1' data. The over-sampling method is introduced to resolve this issue (Chawla et al., 2002). By duplicating tweets of the 505 tweets, this paper adds 3,590 tweets into the data sets for training and test. Thereafter, we randomly rank the 8,190 tweets including 4,095 damage-related and 4,095 not-related, for the 10-fold cross validation.

Performance evaluation

The 8,190 tweets are divided into 10 equal size data sets after the random rank. Each data set

has 819 tweets. Then, this research selects nine data sets for training the classifier and uses the remaining one data set for the classifier test. This process continues for 10 cycles and each cycle uses different data sets for the classifier training and test. The means of precision recall and F1-socres for the classifier models including Naïve Bayes, SVM and Decision Tree in the 10 cycles are calculated and presented in Table 3.

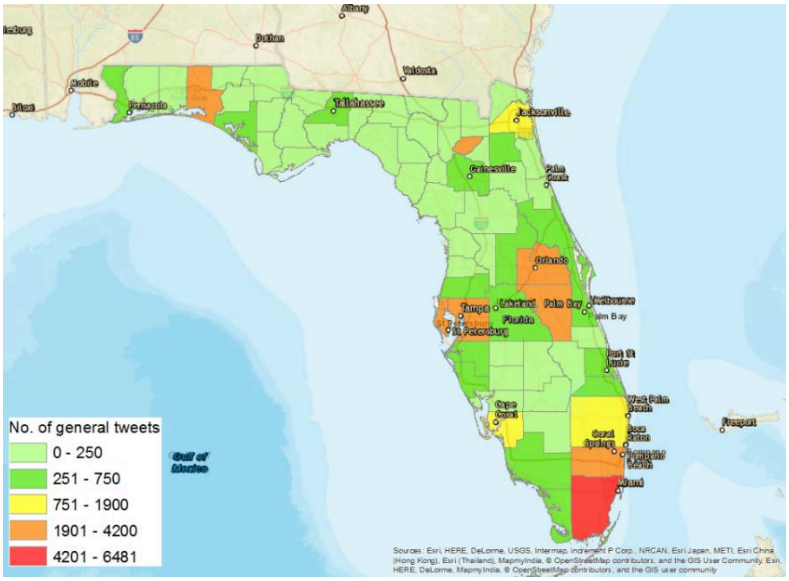

Figure 2. Geographic distribution of the number of general tweets across Florida (adapted from Yuan and Liu (2018a))

According to Table 3, Naïve Bayes classifier model generates a F1-score of 0.70, which is better than the other two models. Meanwhile, its recall value of F1-scores for the classifier models (i.e., 0.79) is also better than the other two models' recall results. The precision of Naïve Bayes is 0.64, which can be acceptable for classifying the Twitter data into two classes including damage-related and not related in this case (e.g., Kaur and Kumar 2015). Considering the precision, recall and F1-scores, we find that Naïve Bayes classifier model presents the best performance for identifying the damage-related social media data during Hurricane Matthew.

Table 3. Performance evaluation of the selected classifier models

Classifier models	Precision	Recall	F1-scores
Naïve Bayes	0.64	0.79	0.70
SVM	0.63	0.31	0.35
Decision Tree	0.73	0.27	0.31

CONCLUSION

The proposed machine learning approach uses 2- to 3-gram level TF-IDF vector to detect the features of damage-related tweets in Hurricane Matthew. This research has introduced three classifier models for identifying the damage-related social media data during Hurricane Matthew. We have selected Naïve Bayes model as the most accurate classifier for our case based on the F1 score. An overall precision is 0.64 while the F1-score is 0.70. To improve the performance of the classifier, further studies can focus on the enhancement of the training data sets. Our results provide the future research with a new method to extract the damage-related

information from social media data or the other crowdsourcing data. The training and test data sets can be generalized to the other hurricane or flood disaster events in Florida for the automatic annotation of damage-related social media data. Future studies can compare the regression results of classification results by machine learning approach with real hurricane damages, and the classification results by keyword-filters with real hurricane damages.

REFERENCES

Chapin, F.S. (1975). "Human activity patterns in the city: things people do in time and in space." *Soc. Indic. Res.,* 2 (2), 261–264.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research*, 16, 321-357.

Forman, G. (2003). "An extensive empirical study of feature selection metrics for text classification". *The Journal of Machine Learning Research*, 3, 1289-1305.

Gal-Tzur, A., Grant-Muller, S. M., Kuflik, T., Minkov, E., Nocera, S., and Shoor, I. (2014). "The potential of social media in delivering transport policy goals." *Transport Policy*, 32, 115-123.

Garg, M., and Kumar, M. (2016). "Review on event detection techniques in social multimedia." *Online Information Review*, 40(3), 347-361.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). "Comparing and combining sentiment analysis methods." Proc., *the first ACM conference on Online social networks*, ACM, Boston, MA, 27-38.

Gu, Y., Qian, Z. S., and Chen, F. (2016). "From Twitter to detector: Real-time traffic incident detection using social media data." *Transportation research part C: emerging technologies*, 67, 321-342.

Guan, X., and Chen, C. (2014). "Using social media data to understand and assess disasters." *Natural hazards*, 74(2), 837-850.

Kaur, H. J., and Kumar, R. (2015). "Sentiment analysis from social media in crisis situations." *Proc., 2015 International Conference on Computing, Communication & Automation*, Noida, India, 251-256.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M. (2016). "Rapid assessment of disaster damage using social media activity." Sci. Adv., 2 (e1500779), pp. 1-11.

Mullen, T., and Collier, N. (2004). "Sentiment analysis using support vector machines with diverse information sources." *Proc., the 2004 conference on empirical methods in natural language processing*, Barcelona, Spain.

Ovadia, S. (2009). "Exploring the potential of Twitter as a research tool." *Behavioral & Social Sciences Librarian*, 28(4), 202-205.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). "Thumbs up?: sentiment classification using machine learning techniques." Proc., *the ACL-02 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, Philadelphia, PA, 79-86.

Spence, P. R., Lachlan, K. A., and Rainear, A. M. (2016). "Social media and crisis research: Data collection and directions." *Computers in Human Behavior*, 54, 667-672.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). "Short text classification in twitter to improve information filtering." Proc., *the 33rd international ACM SIGIR conference on Research and development in information retrieval*, Geneva, Switzerland, 841-842.