Priority prediction of Asian Hornet sighting report using machine learning methods

Yixin Liu
School of Software Engineering
South China University of Technology
Guangzhou 510006, PR China

Jiaxin Guo
School of Computer Science and Engineering
South China University of Technology
Guangzhou 510006, PR China

Jieyang Dong
School of Mathematics
South China University of Technology
Guangzhou 510006, PR China

Luoqian Jiang
School of Software Engineering
South China University of Technology
Guangzhou 510006, PR China

Haoyuan Ouyang
School of Software Engineering
South China University of Technology
Guangzhou 510006, PR China

Abstract—As infamous invaders to the North American ecosystem, the Asian giant hornet (*Vespa mandarinia*) is devastating not only to native bee colonies, but also to local apiculture. One of the most effective way to combat the harmful species is to locate and destroy their nests. By mobilizing the public to actively report possible sightings of the Asian giant hornet , the government could timely send inspectors to confirm and possibly destroy the nests. However, such confirmation requires lab expertise, where manually checking the reports one by one is extremely consuming of human resources. Further given the limited knowledge of the public about the Asian giant hornet and the randomness of report submission, only few of the numerous reports proved positive, *i.e.* existing nests. How to classify or prioritize the reports efficiently and automatically, so as to determine the dispatch of personnel, is of great significance to the control of the Asian giant hornet. In this paper, we propose a method to predict the priority of sighting reports based on machine learning. We model the problem of optimal prioritization of sighting reports as a problem of classification and prediction. We extracted a variety of rich features in the report: location, time, image(s), and textual description. Based on these characteristics, we propose a classification model based on logistic regression to predict the credibility of a certain report. Furthermore, our model quantifies the impact between reports to get the priority ranking of the reports. Extensive experiments on the public dataset from the WSDA (the Washington State Department of Agriculture) have proved the effectiveness of our method.

Index Terms—Sighting Report, Priority Prediction, Machine Learning

I. INTRODUCTION

Bees are the medium for pollination of plants, and they have made great contributions to human society and ecosystem. Effective pollination by bees can increase the quantity and quality of crops, plant diversity and plant resistance to pests [1]. These crops account for one third of the world's food

Yixin Liu, Luoqian Jiang and Haoyuan Ouyang are with the School of Software Engineering, South China University of Technology. Jiaxin Guo is with the School of Computer Science and Engineering, South China University of Technology. Jieyang Dong is with the School of Mathematics, South China University of Technology. E-mail: {seyixinliu, cs_guojiaxin}@mail.scut.edu.cn, 885032560@qq.com, lqrosie728@163.com, 1228194587@qq.com.

production, which is an immense contribution. In addition, the products produced by bees can also be further processed into daily necessities for cleaning, beauty and other purposes. Therefore, as contributors to the ecosystem and human society, bees are worthy of protection.

However, the bee ecosystem has been severely threatened in recent years [2]. The Asian giant hornet is a kind of bumblebee with bright yellow feet. As an alien species, they could kill an average of 40 other kinds of bees in their lifetime, which has caused great damage to the local bee ecological environment. At the same time, the hornet is also very aggressive when it comes into contact with humans, with venom deadly enough to kill. Therefore, the prevention and control of the Asian giant hornet is of great significance to the protection of local apiculture and public safety.

Locating the nests of the hornets and destroying them is one of the most effective ways to control the Asian giant hornet [3]. Existing methods for identifying their nests are mainly based on remote tracking [4] with image recognition [3]. Although these methods are able to locate nests automatically, they require the equipment of drones and face the problem of great search cost. In actual situations, the government often mobilizes the public to submit sighting reports actively, and then sends inspectors to identify and destroy nests in a targeted manner [5]. This method can greatly reduce the cost of searching by leveraging the power of citizens.

However, given that citizens often lack related knowledge of the Asian giant hornet and the randomness of citizens' report submission, the report collection is actually full of "noise", and only a small part of the reports inspected are finally confirmed as positive. At the same time, the current traditional methods require manual credibility assessment of the reports, which is time-consuming and inefficient [6]. To solve these problems, under the limitation of human and financial resources, the automatic assessment and prioritization of the reports is an important way to improve the efficiency of government resource utilization and the prevention and control of pests.

The field of credibility evaluation has used multi-factor [7], natural language processing (NLP) and other technologies to evaluate the confidence of the report. There are currently multiple application scenarios, *e.g.* bug report [8] and vehicle data [9] . However, there is little research and application of these technologies in the field of sighting report credibility evaluation.

In this paper, we combine the related technologies of report credibility assessment, and propose a method based on machine learning (ML) to implement the automatic report prioritization. Different from existing methods of identifying hornet nests, we locate them from a new perspective based on the credibility prediction of public sighting reports. We abstracted this problem into a two-classification problem, and extracted a variety of factors in a report to construct rich features. Finally, we built a classification model based on logistic regression to implement automated prediction.

Our contributions are summarized as follows:

- We have proposed an automated prediction method for the confidence level of sighting reports of hornets, which provides a new idea for the identification and precise control of pests.
- We abstract the modeling of the prioritization of sightings reports as a two-classification problem. By utilizing a variety of data from the sighting report, abundant features have been extracted about the credibility of textual notes, sighting location, and attached image(s).
- We build a machine learning framework based on logistic regression algorithms, which can take into account multiple factors of the report and predict the report credibility. At the same time, for the extreme imbalance problem in the sample, we propose a weighted loss function. On the public data set from the Washington Department of Agriculture, a large number of experiments have proved the effectiveness of our method.

II. METHODOLOGY

In this section, we first formalized the problem of priority prediction of sighting reports into a two-category problem. Next, we used multi-factor analysis to construct rich features for the various types of data in the sighting report: location factors, time factors, image factors, and text factors. Finally, we built a machine learning model based on logistic regression to predict the certainty of the report, and further determined the priority of the report based on the confirmation relationship among the reports. We show the overall architecture of the model in Fig. 1.

A. Problem Definition

To maximize the use of limited government resources, it is essential to accurately predict the credibility of an sighting report so as to prioritize the report, and thus efficiently complete the elimination of the Asian Hornet nest. For a sighting report $x^{(i)} \sim u(\cdot)$, where $u(\cdot)$ denotes a certain distribution that the report data is sampled from. Due to the possibility of mistaken, there will be two different categories.

Let the report category label be $y \in \{0, 1\}$, where $y = 1$ denotes that the report confirms the existence of a nest, and $y = 0$ means that the report is a false positive. We denotes the probability that report $x^{(i)}$ belongs to the true positive as $p^{(i)} \in [0, 1]$. Given multiple sighting reports, we can be prioritized according to their confidence probability $p^{(i)}$. For credible reports, we seek to get a higher probability $p^{(i)}$ and vice versa. Formally, we need to solve the following classification optimization problem, that is, to minimize the Cross Entropy in Equation 1:

$$\max \mathbb{E}_{x^{(i)} \sim u(\cdot)} \left[y^{(i)} p^{(i)} + \left(1 - y^{(i)}\right) \left(1 - p^{(i)}\right) \right] \quad (1)$$

To solve the above optimization problem, we seek to fit a function $p^{(i)} = f(\cdot | x^{(i)}; \theta)$ to predict the confidence probability of the report, so as to further prioritize reports. In the following section, we first construct features from multiple dimensions of the report, and then build a machine learning predictor based on these features to learn an optimal function.

B. Features Extraction

Location Feature. For the location factor, we get the probability feature of reporting the hive found at a certain position by constructing the prior distribution probability feature of the Asian Hornet in space based on the mechanism of bee migration. We first briefly explain the habits of the Asian Hornet. The hornet will only be seen by the public from April to December of a year since its will hibernate in winter. During migration season, the average distance between the migration destination and the original nest is about 30km. After arriving at the destination, the bees move around the nest with a maximum radius of 8km. Based on these habits, given the probability distribution of the initial hive in space, we aim to construct the observation probabilities of bees everywhere in the space.

Given the initial distribution $P_{x,y}^{(0)}$ of the nest in space, we assume that the distance of the new location of nest from the old one follows a Gaussian distribution with an average value of 30 km. We quantify the relation between probability values before and after the migration as follow:

$$P_{\text{nest}}^{(t+1)}(\lambda, \phi) = \sum_{(\lambda', \phi')} \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-(d-30)^2}{2\sigma^2}} P_{\text{nest}}^{(t)}(\lambda', \phi') \quad (2)$$

where (λ, ϕ) and (λ', ϕ') represent the latitude and longitude of two different coordinates in space, $P_{\text{nest}}^{(t)}(\lambda, \phi)$ denotes the probability of a nest existing at position (λ, ϕ) in month t, σ is a shape parameter for Gaussian distribution, which measures the difference of distances of the newly selected location from the old one, and thus reflects the migration precision. The larger the σ, the wider the range over which the queen builds her new nest.

In non-migratory months, we can spot hornets away from near their nest, as its possible location is definite. The further away from the nest, the lower the probability of spotting hornets, and *vice versa*. Formally, according to the nest distribution as defined by Equation (2), we can obtain the
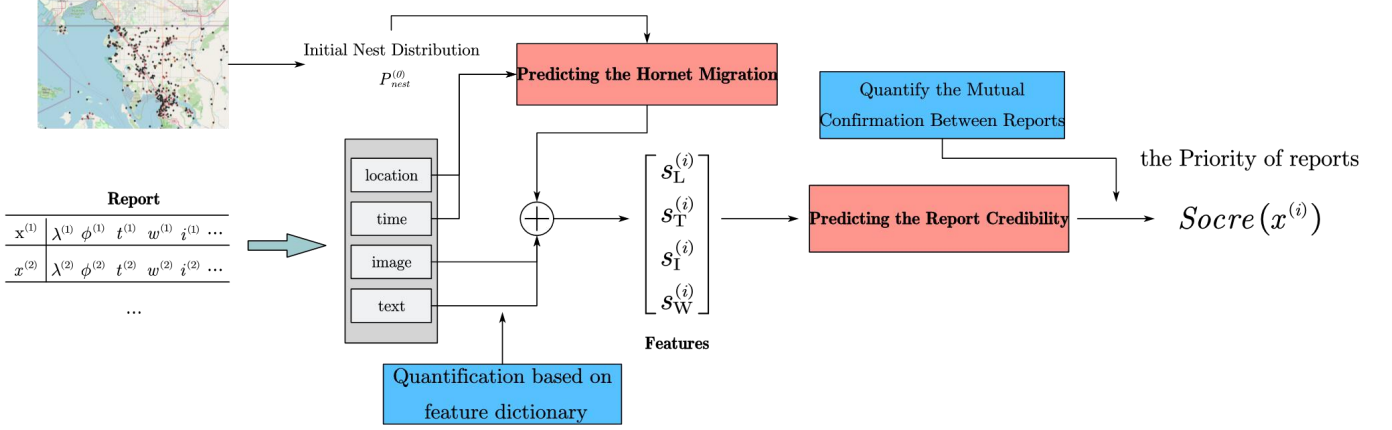
Fig. 1. Given a sighting report, we first extract four features: locational, temporal, graphical, and textual. After constructing the feature vector of the report, we use a logistic regression model to predict and classify the credibility of the report. Finally, we quantitatively consider the interaction between the reports and determine the priority of the reports.

probability distribution of spotting hornets in the whole space as:

$$P_{\text{observe}}^{(t)}(\lambda', \phi') = \sum_{(\lambda, \phi)} e^{-\beta_1 d} P_{\text{nest}}^{(t)}(\lambda, \phi) \tag{3}$$

where $P_{\text{observe}}^{(t)}(\lambda', \phi')$ denotes the probability of finding hornets at position (λ', ϕ') in month t and β_1 is an impact factor that determines the strength of the influence of distance on the probability of spotting a hornet.

Time Feature. For the time factor, we considered the migration habit of the hornet. Since they are more prevalent amid April-December, the public are more likely to spot them during this time, and the report is more likely to be positive. In December-March, the only surviving individual queens hibernate underground, so it is unlikely to spot a hornet above the ground. The period from April-December each year is hence defined as the active period, denoted by $\mathbb{T}$. The temporal feature of each report was characterized as s_{T} by the following rules:

$$s_{\text{T}} = \begin{cases} c_T, & T \in \mathbb{T} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where c_T is the number of positive reports in month T over the historical data.

Image Feature. For the image factor, we considered the number of images attached to a sighting report for the reason that report with graphical proof improves the credibility of the report. In the process of manual verification, the existence of image information is also an important factor to consider for deciding the sending of personnel. We therefore define s_{I} to represent the integrity of image information:

$$s_{\text{I}} = \begin{cases} n_{\text{I}}, & \text{image(s) provided} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where n_{I} denotes the number of image(s) attached to the report.

Text Feature. For the textual factor, we considered the contribution of the length of the text and the keywords contained in the text to the credibility of the report. As for the length of the text, we believe that the length of the text is proportional to its credibility because it contains more evidence. In addition, we extracted two types of text keywords. If the note involves more features closer to *V. mandarinia* than other species, the report is less likely to be mistaken. We constructed a hornet feature dictionary (Table I) from a glossary extracted from attached data, related *Wikipedia* webpages, and literatures. For each report, we are then able to calculate the word frequencies of all the words it contains with respect to this dictionary and obtain its word frequency feature.

TABLE I
THE DICTIONARY OF KEY CHARACTERISTICS OF *V. mandarinia*

	Asian giant hornet	Other confusing hornets
Nest Location	"Underground","forests", "burrows","roots","trunks", $\cdots$	"Time limbs", "house eaves", "exposed", "lawns", $\cdots$
Body Appearance	"Yellow heads", "black thorax", "striped abdomens", "giant", $\cdots$	"Small", "black and white colored", $\cdots$

Combining these two points, we define s_{W} as the text feature to evaluate a report:

$$s_{\text{W}} = \frac{1}{n_{\text{W}}} \sum_{i=1}^{n_{\text{W}}} (q_i - k_{\text{i}}) + \beta_2 \log(n_{\text{W}} + 1) \tag{6}$$

C. Priority Prediction

Classification model. Logistic regression is a generalized linear model, which conducts mapping from any real number to a probability values. For a report $x^{(i)}$ which is characterized by the feature vector $\mathbf{s}$, the logistic function can be used to construct its probability p.

$$p\left(x^{(i)}\right) = \frac{1}{1 + e^{-\theta^T \phi\left(x^{(i)}\right)}} \tag{7}$$

where $\phi(\cdot)$ is the feature mapping for reports (yielding $\mathbf{s}^{(i)}$), and θ is the parameter of the model to learn. To learn the

model, we seek to minimize a loss function called *binary cross-entropy function* denoted as $H(\cdot)$. However, there are only few positive report among a large number of reports, which means that dataset is usually extremely unbalanced. To tackle this problem, we use the weighted binary cross-entropy function, where, for the correctly classified fraction, *viz.* when $y = 0$, if our model predicts it as misclassification, we scale up its loss by a degree of τ to ensure that our model would not predict all samples as misclassifications:

$$H(p, y) = \tau y \log(p) + (1 - y) \log(1 - p) \quad (8)$$

where τ is the weighting coefficient of the sample status, set as the reciprocal of the percentage of non-misclassified reports. On this basis, we can obtain the overall loss function $J(\cdot)$ of the given dataset $\mathbf{x}$:

$$\min J(\theta) = \frac{1}{n} \sum_{i=1}^{n} H\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \frac{\beta_3}{2} \|\theta\|_2^2 \quad (9)$$

where $h_\theta(\cdot)$ is the logistic regression function with parameter θ, and β_3 is the balancing coefficient for regular terms.

We update the parameter of model using stochastic gradient descent method. Let η be the step of parameter update, we train our model to minimize the loss $J(\theta)$.

$$\theta' \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}. \quad (10)$$

Priority determination. In the previous section, we predict the credibility of report based on four dimensions of the report content. However, to predict the priority of report, we have to consider their connection. Intuitively, if there are many reports submitted in a certain area at the same time, it highly indicates that this area is likely to have nest hidden. Therefore, based on the forecast confidence of a single report, we seek to further consider the mutual confirmation relationship between the reports, so as to achieve more accurate report priority ranking.

Specifically, the predicted possibility p in Equation 7 only considers one report and does not consider its inter-corroboration with reports that are similar to it in space and time. If the reports are in the same migration cycle (which means that the reports are between April of a certain year and March of the next year), within which the hornets do not migrate, we are interested in whether the reports lead to the same nest; if different, we consider whether the nests that the reports lead to are related. For this, we defined the mutual influence factor $F_{i,j}$ between reports x_i, x_j as:

$$F_{i,j} = \begin{cases} e^{-\lambda d_{ij}}, & \text{if } t_i = t_j \\ \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(d_{ij} - 30\Delta t_{i,j})^2}, & \text{otherwise} \end{cases} \quad (11)$$

where $\Delta t_{i,j}$ denotes the difference between the migration cycles of i-th and j-th report. Based on $F_{i,j}$, we construct the priority evaluation Z of a report:

$$Z_i = \sum_{j=1}^{n} F_{i,j} p_j \quad (12)$$

where Z_i denotes the priority of i-th report and n denotes the total number of reports involved in ranking. According to Equations (11) and (7), the priority Z_i of each report can be obtained. The greater the Z-value of a report, the more likely it is to be positive, and thus the sooner it should be investigated.

III. EXPERIMENTS

In the experiment, we used the Asian Hornet sighting report dataset[1] published by the WSDA (Washington State Department of Agriculture) to verify our model. This dataset includes the sighting report ranging from 2019 to 2021, which details of sighting report and the label after reviewed by the official. We use accuracy metric to evaluate the classification model and conduct an ablation experiment to demonstrate the effectiveness of our proposed module.

A. Implementation Details

Data Cleaning. Among the 4,400 samples, the main informational features of each report were the coordinates, date, text, (possible) images, and officially given identification label. We excluded the following parts of the data: 15 reports with invalid date, 15 with "unprocessed" label, and 56 prior to 2019. These parts account for only a small fraction, and contribute little to our later study. After cleaning, we got 4,355 pieces of data. There are three categories of result labels in the cleansed data: positive, negative, and unverified. The number of sample categories is highly unbalanced, with only 14 sightings verified as positive cases, accounting for only 0.3%, and the remaining half of each case being negative and unverified, respectively.

Parameter Setting. To determine the initial distribution $P_{x,y}^{(0)}$ in Equation 2, we use the confirmed sighting report locations of the data set from March 2019 to April 2020 as the initial nest probability distribution of Washington State. The probability value of each positive reported location is set to 1, and the other locations are all set to 0. The parameter β_1 in Equation 3 is set to 0.57, and the σ in Equation 2 is set to 10.

B. Results of classification prediction

Note that the category of report in this dataset are extremely unbalanced. We conduct an ablation experiments to prove the effect of the weighted loss coefficient in Equation 8. To this end, we take the parameter τ as a model parameter, and obtain the training and test sets by 5-fold split (we will ensure the inclusion of samples of both statuses in each set). Then, we performed grid search for $\tau \in \{\tau_0 - \Delta\tau, \tau_0, \tau_0 + \Delta\tau\}$, with a 5-fold cross-validation for each parameter (still ensuring both inclusions). Since the reports were mostly negative, the naïve use of average accuracy would not evaluate well the prediction of our model, so we took the average accuracy on positive reports as the metric. After searching for an optimal parameter setting, a prediction accuracy of 75% is achieved on only few samples. Considering the prediction difficulty due to the extreme imbalance of the samples (with
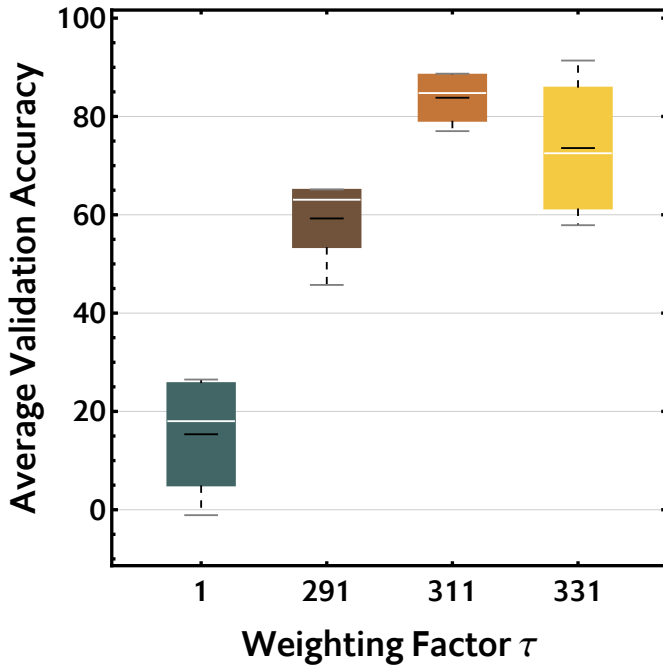
Fig. 2. Under different scaling parameter τ settings, the cross-validation accuracy on report samples of imbalanced statuses.

Fig. 3. The priority rankings of the 7 positive samples in the validation set with different ranking models. For these positive reports, the higher the priority predicted by the model, the better. Noted that we use the sort position of the report among 14 report in testing set to indicate its priority.

positive/negative $\approx 1/400$), the model performs well enough. Also, comparing results with and without setting the weighting parameter ($\tau = 1$ for the original binary cross-entropy function), we find that the weighting correction parameter contributes much to improved performance. With the best parameter settings ($\tau = 311$), our classification model can achieve the average accuracy of 83.5%. However, compared with recent related works (*e.g.* [3] achieved detection rate of 93%), there is still room for improvement in our method.

C. Results of priority prediction

To prove that introducing the mutual influences among reports improves the priority sorting, we performed ablation experiments to demonstrate the effectiveness of our prioritization module by comparing to the baseline method that directly uses the probability computed using Equation 7 to sort the priority. We construct a small and balanced dataset that includes 28 reports, half of which are positive. Specifically, the dataset was split as 1:1, with 14 for training and 14 for testing. The results in Fig. 3 show that the priority of positive sample predicted by our method is much higher compared to the baseline, which means that our method can better identify high-priority reports. This demonstrates the effectiveness of the purposed module.

IV. CONCLUSION

In this paper, we proposed the priority prediction of Asian Hornet sighting reports based on Machine Learning, which solved the problem that prioritizing Asian Hornet sighting reports efficiently and automatically under the limitation of resources. We formalized the problem of priority prediction

of sighting reports into a two-category problem. To characterize a sighting report, we construct rich features from four dimensions: location, time, image and text. Then, we purpose a machine learning model based on logistic regression to predict the credibility of the report, and determined their priority based on the relationship among the reports. Extensive results on the public dataset demonstrate the effectiveness of the proposed method.

REFERENCES

[1] S. A. Steffan, P. S. Dharampal, B. N. Danforth, H. R. Gaines-Day, Y. Takizawa, and Y. Chikaraishi, "Omnivory in bees: Elevated trophic positions among all major bee families," *The American Naturalist*, vol. 194, no. 3, pp. 414–421, 2019.

[2] Q. Rome, A. Perrard, F. Muller, C. Villemant *et al.*, "Monitoring and control modalities of a honeybee predator, the yellow-legged hornet vespa velutina nigrithorax (hymenoptera: Vespidae)," *Aliens*, vol. 31, no. 31, pp. 7–15, 2011.

[3] T. Shams and P. Desbarats, "Detection of asian hornet's nest on drone acquired flir and color images using deep learning methods," in *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2020, pp. 1–6.

[4] S. Kim, C. Ju, J. Kim, and H. I. Son, "A tracking method for the invasive asian hornet: A brief review and experiments," *IEEE Access*, vol. 7, pp. 176 998–177 008, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2958153

[5] M. Pusceddu, I. Floris, R. Mannu, A. Cocco, and A. Satta, "Using verified citizen science as a tool for monitoring the european hornet (vespa crabro) in the island of sardinia (italy)," *NeoBiota*, vol. 50, p. 97, 2019.

[6] W. Zhiquan, "Trouble of vespa mandarinia: Confirming the buzz about hornets," in *E3S Web of Conferences*, vol. 245. EDP Sciences, 2021, p. 02044.

[7] D. W. Gjertson, "A multi-factor analysis of kidney regraft outcomes." *Clinical transplants*, pp. 335–349, 2002.

[8] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer, "Web credibility: Features exploration and credibility prediction," in *European conference on information retrieval*. Springer, 2013, pp. 557–568.

[9] R. W. Thomas and J. M. Vidal, "Assessing the credibility of vehicle data reported by anonymous sources," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEM-CON)*. IEEE, 2017, pp. 384–388.