Trouble Of Vespa Mandarinia: Confirming the Buzz about Hornets

Wang Zhiquan[1*]

[1]Business School, Shandong Normal University,
Jinan, Shandong, 250358, China

Abstract. In order to help Washington State interpret the data about Vespa mandarinia provided by the public report, and enable government agencies to adopt corresponding strategies to prioritize correct reports when resources are limited, for further investigation, this article establishes two targeted models: The first unsupervised probability prediction model. First, extract the text information of misjudgment classification in the data set, and carry out preprocessing. The data set is divided into training set and test set according to the ratio of 8:2, and the Latent Dirichlet Allocation model is trained using the misjudgment classification information in the training set. After the model training is completed, this paper makes a probability prediction on the data on the test set, and evaluates the robustness of the model through the accuracy rate on the test set. The second text similarity matching model is based on feature dimensionality reduction and extracting feature keywords as vectors. The TF-IDF algorithm is used to calculate the weight of each feature keyword in the vector to form a standard bag-of-words vector for the correct witnessing of the Vespa mandarinia report. Judge by the similarity of text similarity matching model.

1 Introduction

Vespa mandarinia are native to temperate and tropical East Asia.December 2019 The species first appeared in Washington.Washington State Department of Agriculture said that Vespa mandarinia, as an invasive species, will have a negative impact on the environment, economy and public health in Washington State.

Washington State has set up a helpline and website to collect the sightings of Vespa mandarinia. But most reported sightings mistake other wasps, such as European bumblebee and cicada killer, because they are similar in size, shape and color. This makes it more difficult to interpret the data reported by the public. With limited resources of government agencies, how to give priority to these public reports for further investigation is an urgent problem to be solved. Considering the background information and restricted conditions identified, we need to solve the following problems:

1.Create, analyze, and discuss models that predict the possibility of misclassification using data set files such as spreadsheets of sighting reports published by the Washington Department of Agriculture

2.Use this model to make the reports of priority investigations most likely to be real sightings of Vespa mandarinia

For the first question, this article establishes an unsupervised probability prediction model. First, extract the text information of Negitive ID in the data set and preprocess it. In this paper, the data set is divided into training set and test set according to the ratio of 8:2, and the Latent Dirichlet Allocation model is trained using the misclassified information in the training set. After the model training is completed, this paper makes a probability prediction on the data on the test set, and evaluates the robustness of the model through the accuracy rate on the test set.

Aiming at the second question, this paper performs feature reduction on the result of word segmentation and extracts feature keywords as vectors. The TF-IDF algorithm is used to form a standard bag-of-words vector for correct witnessing of the Vespa mandarinia report. The text similarity matching model is used to judge the accuracy of the new sighting, and then provide a basis for discussing whether the sighting should be investigated first.

However, there are many factors that affect Vespa mandarinia, and all of them cannot be taken into account when assigning the weights of the influencing factors. There are certain errors for very special circumstances.

To solve this problem, we must grasp the dynamic information of Vespa mandarinia and collect Interpret the information and process the data in a timely manner. Realize the accurate processing of data and ensure the rapid and effective application of limited government agency resources. This is also the key of this article.

* Corresponding author: 765808770@qq.com

2 Materials and Methods

2.1 LDA probabilistic topic model

2.1.1 Processing of text data

According to the analysis, this paper extracts the text information of the two attributes of Notes and Lab Comments in the research data. In order to facilitate the establishment of the model, this article preprocesses the extracted text information. Mainly includes the following four steps:

(1) Delete abnormal data: when processing the data, it is found that there are some abnormal data, such as the null value and abnormal situation in the year and the non-string record in the Notes attribute. Since the data provided in the attachment is large enough, we directly delete the abnormal data. After deleting the abnormal data, a total of 4411 records were obtained.

(2) Word segmentation: since the obtained text information of the two attributes of Notes and Lab Comments is connected to the entire sentence, each word is not independent. In order to process the acquired text data into independent features, this paper uses spaces to segment the text information. The following figure shows the results of this article using the discrete graph to determine the position of the word in the text. It can be seen that the two words "Hornet" and "hornet" have a higher frequency in the results. These two words will be treated as two different features, leading to an increase in training time and reducing the accuracy of prediction. Therefore, it is necessary to extract the stems of the words to convert words with the same root but different parts of speech and tense into one word.

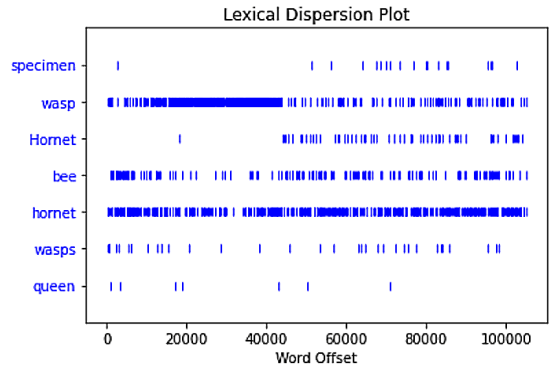

Figure 1. The discrete graph showing the offset of words

(3) Stemming: stemming is the process of grouping together words with the same root but different parts of speech and tense. For example, although "fly" and "flying" have different forms, they should belong to the same word after stemming. This article uses the nltk library in the python programming language to extract the stem of the training text.

(4) De-stop words: De-stop words refers to deleting characters that are useless for prediction in the training text. This paper conducts word frequency statistics on the two attributes of Notes and Lab Comments that are correctly identified as Vespas, and finds that the top 5 words with word frequency occurrences are "There", "2019", "18", "BC", and "Sep". These words are obviously less helpful to the establishment of the prediction model, so the extracted data also needs to be processed to remove the stop words to improve the training, classification speed and classification accuracy.

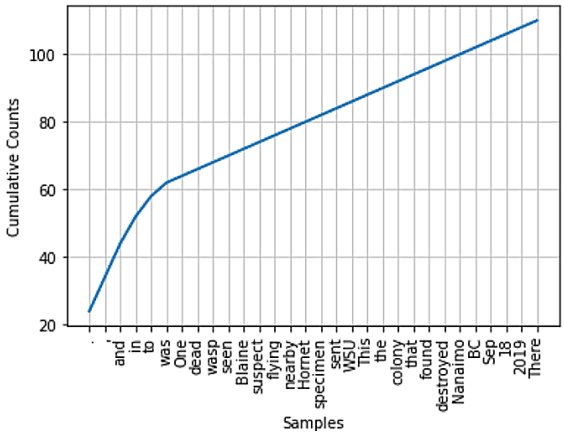

Figure 2. Preliminary word frequency statistics of positive

2.1.2 Latent Dirichlet Allocation Model

Latent Dirichlet Allocation Model (LDA) is a text topic generation model, which includes three layers: word, topic and text. In this paper, the generative model means that every word in the text is obtained by selecting a topic with a certain probability and selecting a word from the topic with a certain probability. LDA is an unsupervised Bayesian model, which can be used to identify hidden topic information in large-scale texts. Each text in the text set is given in probability form. The model establishes a learning model based on the joint probability distribution from text input to classification output on the premise that text feature words are independent of each other through a preset training set. Then, according to this model, the data feature setis input, and the maximum posterior probability outputis obtained.

The core formula of LDA is as follows:

$$P(w|d) = p(w|t) * p(t|d) \quad (1)$$

Among them, each piece of text d represents a word sequence $\langle w_1, w_2, ..., w_n \rangle$ with n words. w_i stands for the i-th word. t stands for topic, which is a topic vector trained by taking text set d as input.

Combined with the knowledge of Bayesian graph model, it can be known that given parameters α and β, the joint probability of topic mixture θ, topic t and word w is:

$$p(\theta, t, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(t_n|\theta)p(w_n, t_n, \beta) \quad (2)$$

In which N represents the number of words in all texts, w_n represents the n th word, t_n represents the

selected topic,and $p(t|\theta)$ represents the probability distribution of the topic t at a given time; $p(w|t)$ represents the probability distribution on the word w when the topic t is given. Parameters α and β are two parameters for LDA model to learn and train by giving you input expectation.

The steps we use LDA probabilistic topic model training prediction are shown below:

(1) Word segmentation and filtering are performed on the text data of the test set.

(2) Assign a digital ID to the word set after word segmentation.

(3) Use the data obtained in (1) and (2) to convert words into sparse vectors containing word ID and word frequency.

(4) Bring the data obtained in (2) and (3) into the trained LDA model to extract hidden information.

(5) Classification according to the probability of each test set text on each topic.

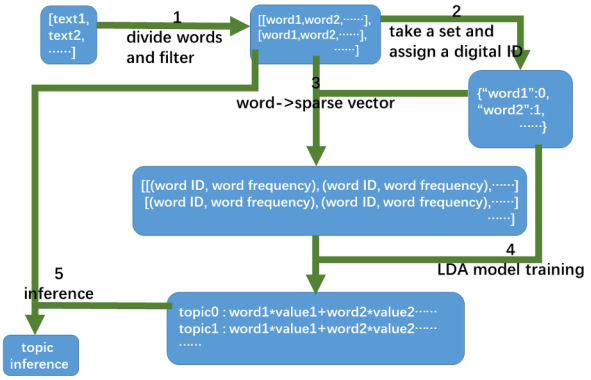

Figure 3. The steps of Latent Dirichlet Allocation

2.2 Text similarity matching model

2.2.1 Feature dimension reduction of text keywords

$$MI(l_i,c_i)=\log\frac{p(l_i\cap c_i)}{P(l_i)P(c_i)}=\log\frac{p(l_i|c_i)}{P(l_i)} \quad (3)$$

Among them, represents the probability that texts satisfying word and category appear in the training text data set. represents the probability of the text containing the word in the training text data set. represents the probability that the text of category appears in the training text data set.

We use the mutual information algorithm to calculate the features of each word in each text in the data set, and get the quantitative results, which are arranged in the order from large to small.

At the same time, according to the Markov hypothesis, the words in the article are often dependent, and many words appear in the form of phrases. Therefore, we select the top 10% phrases with more than 2 characteristic words as the representative of the original text data, such as "hornet specimen", "yellow heads", "black thorax" and "Brown Striped abdomens" and so on. We will extract the feature keywords to form a word bag, such as ["hornet specimen","yellow head", "black thorax", "Brown Striped abdomens"].

2.2.2 Weight calculation of feature keywords

After dimensionality reduction of original text features, a part of feature keywords are selected, but the importance of each feature keyword is different, and the more important phrases account for more weight in similarity comparison.

Therefore, we need to evaluate the importance of each phrase and give weight. TF-IDF is a classical algorithm to evaluate the importance of phrases, so we use this algorithm to calculate the weight of keywords after feature dimensionality reduction. The calculation formula of TF-IDF algorithm is as follows:

$$w=tf\times\log\frac{N}{df+1} \quad (4)$$

Among them, w is the weight result of a certain feature keyword we want to calculate. df represents the number of texts in the training set that contain the feature keyword. tf represents the ratio of the number of occurrences of the characteristic keyword in a text to the total number of words in the text. N refers to the number of texts in the training text data set.

We use mutual information algorithm and TF-IDF algorithm to get some feature keywords and weights, as shown in Table 4. It can be seen that the weight of "hornet specimen" feature keywords is relatively large, which is 0.12. In the four feature keywords, the weight of "Brown Striped abdomens" is the smallest, which is 0.034. All feature keywords and weights constitute the standard bag vector.

Table 1. Part of feature keywords and weights

Feature keywords	weight
hornet specimen	0.12
yellow head	0.08
black thorax	0.076
brown striped abdomens	0.034

2.2.3 Text Similarity Matching Model

Each feature keyword in the bag of words obtained in the previous step is multiplied with the calculated weight, and we get the standard bag of words vector $X=(x_1,x_2,x_3,...,x_n)$.

In order to measure the similarity between the matching text $Y=(y_1,y_2,y_3,...,y_n)$ and the standard bag-of-words vector, we established a text similarity matching model, using cosine similarity for text similarity calculation, the specific calculation formula is as follows:

$$\cos\theta = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad (5)$$

Among them, n represents the dimension of the vector. x_i represents the i -dimensional keyword of the standard bag of words vector. y_i represents the i -dimensional keyword of the standard bag of words vector.

In order to test the rationality of the text similarity matching model, we do similarity matching on three text records in the test set, and the matching results are shown in the following table. The results show that text1 is the most similar to the standard bag vector, which is in line with the actual situation.

Therefore, to a certain extent, it shows that our text similarity matching model can be used to match the text to check whether the eyewitness event is the event of correctly identifying the Vespa mandarinia.

Table 2. Similarity matching results of part of text records in the test set

Text		Similarity value
Text1	Hornet specimen sent to WSU	0.77
Text2	Dead hornet in light	0.42
Text3	doorbell cam image	0.24

3 Results & Discussion

3.1 LDA probabilistic topic model

The LDA probabilistic topic model divides the test set data into four types of topics. After each text is entered into the model, the probability distribution of the text topic will be obtained, that is, the probability of the text on the four topics. The table shows part of the test data including Text1, Text2, Text3 and Text4, and the predictions on the LDA model. Among them, 0.73 means that the probability that Text1 is Topic1 is 0.73. Statistics found that the probability of all texts in the test set on Topic1 is greater than that on the other three topics, accounting for 86%. Therefore, when the probability of making a mistake does not exceed 0.14, this article believes that Topic1 is the most suitable text topic that other bees mistakenly believe that Asian Vespa mandarinia, that is, Topic1 can represent the Negtive ID category;

Table 3. Partial results of mistaken classification predicts

	Topic1	Topic2	Topic3	Topic4
Text1	0.73	0.12	0.05	0.1
Text2	0.8	0.1	0.05	0.05
Text3	0.61	0.02	0.22	0.15
Text4	0.3	0.2	0.1	0.4

In this article, it is defined that the accuracy of the input text on the Topic1 topic representing Negative ID is greater than 0.75, which means that the input text is correctly classified under the Negative ID category. In this paper, the accuracy of the LDA model on the test set is used to evaluate the robustness of the model. The specific calculation of accuracy is as follows:

$$p = \frac{TP}{TP + FP} \quad (6)$$

Among them, p , TP and FP are the possibility of misclassification, the number of correctly classified texts and the number of misclassified texts, respectively.

Through the above formula, we calculate that the probability of misclassification predicted by the LDA probabilistic topic model in this article is 0.67. Therefore, for the second question, the conclusion of this article is that when the probability of making a mistake does not exceed 0.14, the LDA probabilistic topic model predicts the probability of misjudgment classification is 67%, which proves the robustness of the model to a certain extent.

3.2 Text similarity matching model

The text similarity matching model judges whether the sighting report is optimal through the similarity value, which may be a report for correctly identifying the Vespa mandarinia. The greater the similarity of the text similarity matching model, the more likely the sighting is to correctly identify the Asian Vespa mandarinia. Therefore, the sighting should be investigated first.

4 Conclusions

This article mainly adopts the method of information transformation, using data visualization and other technologies to reflect a large number of and diversified data in the form of graphs or tables, so that the data can provide information in a more intuitive form. This article will understand the biological characteristics of the Vespa mandarinia into the model's influencing factors. The model is more realistic and focuses on solving local problems in accordance with local conditions.

The model for predicting the probability of misclassification in this paper has been trained and tested many times. Integrate and modularize the large amount of data actively provided by the masses, so as to make the priority investigation most likely to be the report of positive sightings. On this basis, the targeted allocation of resources will effectively solve the problems caused by the Vespa mandarinia to Washington State as soon as possible.

The model in this article is intended to be helpful to the Washington State Department of Agriculture, to contribute to the local solution to the Vespa mandarinia problem, and to provide references and solutions to places where biological invasion problems exist in the

world. Biological invasion is a worldwide problem, but with our efforts, it will definitely be overcome.

References

1. Zhou Xia, Zhang linyan, Ye Wanhui. ecological space theory and its application in biological invasion research [J]. advances in earth science, 2002(04):588-594.
2. Wang jiamiao. research on topic model for text semantic analysis application [D]. Hefei university of technology, 2020.
3. Jia Longjia. Research on Feature Weighting Algorithm and Text Representation Strategy in Text Classification [D]. Northeast Normal University, 2016.
4. Liu Aiqin, Xiaoning Ma. Construction of automatic short text classification system based on probabilistic topic model [J]. National journal of library science, 2020,29(06):102-112.
5. Chen Hongshu. Research on Patent Text Mining Method and Application Based on Topic Model [D]. Beijing Institute of Technology, 2015.