

- [0142]** a calculating sub-unit, configured for calculating a false positive rate FPR and a true positive rate TPR based on TP, TN, FN, and FP, wherein their respective calculation formulas are $FPR = FP / (FP + TN)$, $TPR = TP / (TP + FN)$; and
- [0143]** a drawing sub-unit, configured for drawing the ROC curve taking the FPR as the abscissa and the TPR as the ordinate.
- [0144]** Further, the device for training a topic classifier further includes:
- [0145]** a second obtaining module, configured for substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;
- [0146]** a first adjusting module, configured for adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;
- [0147]** a second adjusting module, configured for when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold, returning to the following operation: adjusting the preset AUC threshold until the p is greater than the preset p threshold, and the r is greater than the preset r threshold, and training to get the second topic classifier;
- [0148]** a classifying module, configured for classifying the text data using the second topic classifier.
- [0149]** Further, the collecting unit includes:
- [0150]** a collecting sub-unit, configured for collecting the text data, and segmenting the text data;
- [0151]** a removing sub-unit, configured for removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;
- [0152]** a calculating sub-unit, configured for calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.
- [0153]** Further, the calculating sub-unit includes:
- [0154]** a first calculating sub-unit, configured for calculating the term frequency

etc.) to perform the methods described in the various embodiments of the present disclosure.

[0160] The foregoing description portrays merely some illustrative embodiments of the present disclosure, and are not intended to limit the patentable scope of the present disclosure. Any equivalent structural or flow transformations based on the specification and the drawing of the present disclosure, or any direct or indirect applications of the present disclosure in other related technical fields, shall all fall within the protection scope of the present disclosure.

parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters comprises:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

substituting the first hash table into the logistic regression model, and calculating the optimal model parameters of the logistic regression model using the iterative algorithm, to train and get the logistic regression model containing the optimal model parameters.

4. The method of claim 3, wherein the step of drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier comprises:

substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

drawing the ROC curve based on TP, TN, FN and FP;

calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

otherwise when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirement, and trains to get the first topic classifier.

5. The method of claim 4, further comprising:

substituting the second hash table into the first topic classifier to obtain a probability

that the test sample belongs to a corresponding topic;

adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold, returning to the following operation: adjusting the preset AUC threshold until the p is greater than the preset p threshold, and the r is greater than the preset r threshold, and training to get the second topic classifier;

classifying the text data using the second topic classifier.

6. The method of claim 2, wherein the step of collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set comprises:

collecting the text data, and segmenting the text data;

removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

7. A device for training a topic classifier, comprising:

a memory,

a processor, and

a topic classifier training program stored in the memory and executable on the processor, the topic classifier training program when executed by the processor performing the following operations:

obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

8. The device of claim 7, wherein following operations are further performed when the topic classifier training program executed by the processor:

collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

computing a distribution of the text data on a preset number of topics using a preset topic model based on the first keyword set and the preset number of topics, and clustering the text data based on the distribution of the text data on the topics, to train and get the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target topic classifier based on the manual labeling results on the text data based on the topic models, and using the text data other than the training samples as the test sample.

9. The device of claim 8, wherein following operations are further performed when the topic classifier training program executed by the processor:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

substituting the first hash table into the logistic regression model, and calculating the optimal model parameters of the logistic regression model using the iterative algorithm, to train and get the logistic regression model containing the optimal model parameters.

10. The device of claim 9, wherein following operations are further performed when the topic classifier training program executed by the processor:

substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

drawing the ROC curve based on TP, TN, FN and FP;

calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

otherwise when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirement, and trains to get the first topic classifier.

11. The device of claim 10, wherein following operations are further performed when the topic classifier training program executed by the processor:

substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold, returning to the following operation: adjusting the preset AUC threshold until the p is greater than the preset p threshold, and the r is greater than the preset r threshold, and training to get the second topic classifier;

classifying the text data using the second topic classifier.

12. The device of claim 8, wherein following operations are further performed when the topic classifier training program executed by the processor:

collecting the text data, and segmenting the text data;

removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

13. The device of claim 12, wherein following operations are further performed when the topic classifier training program executed by the processor:

- calculating the term frequency TF and the inverse document frequency IDF of each keyword in the second keyword set;

- calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.

14. A computer-readable storage medium, wherein a topic classifier training program is stored in the computer-readable storage medium, the topic classifier training program when executed by the processor performing the following operations:

- obtaining a training sample and a test sample, wherein the training sample is obtained by manually labeling after a corresponding topic model having been trained based on text data;

- extracting features of the training sample and of the test sample respectively using a preset algorithm, computing optimal model parameters of a logistic regression model by an iterative algorithm based on the features of the training sample, to train and get a logistic regression model containing the optimal model parameters; and

- drawing a ROC curve of receiver operating characteristic based on the features of the test sample and the logistic regression model containing the optimal model parameters, and evaluating the logistic regression model containing the optimal model parameters based on the area AUC under the ROC curve, to train and get a first topic classifier.

15. The computer-readable storage medium of claim 14, wherein following operations are further performed when the topic classifier training program executed by the processor:

- collecting the text data, and preprocessing the text data to obtain a corresponding first keyword set;

- computing a distribution of the text data on a preset number of topics using a preset topic model based on the first keyword set and the preset number of topics, and clustering the text data based on the distribution of the text data on the topics, to train and get the corresponding topic models of the text data; and

selecting from among the text data the training samples that correspond to a target topic classifier based on the manual labeling results on the text data based on the topic models, and using the text data other than the training samples as the test sample.

16. The computer-readable storage medium of claim 15, wherein following operations are further performed when the topic classifier training program executed by the processor:

extracting the features of the training sample and of the test sample respectively using a preset algorithm, and correspondingly establishing a first hash table and a second hash table;

substituting the first hash table into the logistic regression model, and calculating the optimal model parameters of the logistic regression model using the iterative algorithm, to train and get the logistic regression model containing the optimal model parameters.

17. The computer-readable storage medium of claim 16, wherein following operations are further performed when the topic classifier training program executed by the processor:

substituting the second hash table into the logistic regression model containing the optimal model parameters to obtain true positive TP, true negative TN, false negative FN, and false positive FP;

drawing the ROC curve based on TP, TN, FN and FP;

calculating the area AUC under the ROC curve, and evaluating the logistic regression model containing the optimal model parameters based on the AUC value;

when the AUC value is less than or equal to a preset AUC threshold, determining that the logistic regression model containing the optimal model parameters does not meet the requirement, and returning to the following operation: computing optimal model parameters of the logistic regression model using the iterative algorithm so as to train and get the logistic regression model containing the optimal model parameters;

otherwise when the AUC value is greater than the preset AUC threshold, determining that the logistic regression model containing the optimal model parameters meets the requirement, and trains to get the first topic classifier.

18. The computer-readable storage medium of claim 17, wherein following operations

are further performed when the topic classifier training program executed by the processor:

- substituting the second hash table into the first topic classifier to obtain a probability that the test sample belongs to a corresponding topic;

- adjusting the preset AUC threshold, and calculating a precision rate p and a recall rate r based on TP, FP, and FN;

- when the p is less than or equal to a preset p threshold, or the r is less than or equal to a preset r threshold, returning to the following operation: adjusting the preset AUC threshold until the p is greater than the preset p threshold, and the r is greater than the preset r threshold, and training to get the second topic classifier;

- classifying the text data using the second topic classifier.

19. The computer-readable storage medium of claim 15, wherein following operations are further performed when the topic classifier training program executed by the processor:

- collecting the text data, and segmenting the text data;

- removing stop words in the text data after the segmentation based on a preset stop word list, to obtain a second keyword set;

- calculating a term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than a preset threshold of TF-IDF, to obtain the corresponding first keyword set.

20. The computer-readable storage medium of claim 19, wherein following operations are further performed when the topic classifier training program executed by the processor:

- calculating the term frequency TF and the inverse document frequency IDF of each keyword in the second keyword set;

- calculating the term frequency-inverse document frequency TF-IDF value of each keyword in the second keyword set, and removing the keyword whose TF-IDF value is lower than the preset threshold of TF-IDF, to obtain the corresponding first keyword set.