Research on the Classification of Aviation Safety Reports Based on Text and Knowledge Graph

Wenfang Zhang¹, Hongwei Shi¹, Yang Yang² and Yonghui Luo¹

¹ College of Computer Science, Sichuan University, No.24 South Section 1, Yihuan Road, Chengdu, 610065, China
² Sichuan Airlines Training Centre, No.1 Xinwen Road, West High Tech District, Chengdu, 611731, China
Email: 2019223045206@stu.scu.edu.cn

Abstract. The existing automatic classification methods of aviation safety reports are mainly traditional machine learning classification algorithms, which have the disadvantages of relying on manual feature extraction and selection, unable to consider domain information and dealing with complex dependencies, which limits the improvement of classification accuracy. Therefore, this paper proposes an automatic classification algorithm for aviation safety reports that combines text and knowledge graph. The algorithm uses the knowledge triples with rich relationships as the background knowledge in the aviation safety field into the word2vec word vector training process so that the words can not only learn context information in training but also learn the complex semantics between words in the aviation field information, so that the trained word vectors can express richer semantic features, thereby improving the classification effect. Through a series of comparative experiments, the results show that the algorithm has the highest F1-score value and has a classification accuracy rate of up to 91.4%.

1. Introduction

In recent years, the global civil aviation passenger traffic has increased year by year, which gives aviation safety it brings great challenges. The Aviation Safety Reporting System (ASRS) collects a large number of detailed reports of unsafe incidents or potential safety hazards discovered by frontline employees of civil aviation in their daily work. These reports provide an important basis for eliminating hidden safety hazards, formulating effective corrective measures and macro policies.

The existing research on automatic classification of aviation safety reports at home and abroad mainly focuses on two aspects of classification models and feature extraction: in the selection of classification models [1-4], text classification based on machine learning and clustering methods are mainly used for aviation safety reports train to generate corresponding classification models and clustering models, and then use these models to classify aviation safety reports; in improving text feature extraction methods [5-9], the classification effect is mainly improved by introducing domain dictionaries and considering semantic grammatical information such as word order.

The aviation safety report itself has the characteristics of the complex structure, strong domain, and huge amount of data, which increases the difficulty of automatic classification. The existing research has the following deficiencies:

The method of automatic classification of aviation safety reports based on traditional machine learning requires experts to formulate rules for feature extraction and selection. And it cannot handle complex dependency relationships between words. When using traditional machine learning text classification methods to classify aviation safety reports, effective text feature representation is crucial.
to the classification results, which requires domain experts with rich relevant business knowledge to spend a lot of time designing feature extraction and selection the rule of. In addition, when extracting some text features, it usually needs to use One-hot, TF-IDF[10], and other text representations, which makes the quality of text features also affected by different text representations. In recent years, some scholars have chosen tools such as word2vec[11] to automatically extract text features and selected deep learning models as classifiers. Although the classification effect has been improved to a certain extent, its shortcomings are also obvious. In the feature selection process, it is completely automatic and unsupervised. Extraction, lack of consideration of domain information, especially on highly domain-specific data sets such as aviation safety reports, the classification effect is bound to be limited.

Based on the above deficiencies, this paper proposes an automatic classification algorithm for aviation safety reports that combines text and knowledge graphs. By integrating knowledge graph as background knowledge in aviation safety into word2vec word vector training process to improve the quality of text feature extraction and enhance the features of text features semantic expression ability. Then the convolutional neural network model is used in the selection of the classifier to reduce the dependence on artificial feature selection, to improve the portability of the classification method and the classification accuracy.

2. Improved Word Vector Model

2.1. Kg2vec Model Idea

The knowledge graph[12-13] contains a large number of entities, and there is rich relationship information between entity words. This paper takes a knowledge graph as the domain background knowledge base of machine understanding natural language and proposes a word vector training model, kg2vec model, which combines text and knowledge graph. In this model, the word2vec model and the TransE[14] model are trained together.

2.2. Algorithm of Kg2vec Model

The basic idea of the kg2vec model is to add knowledge triples information with rich relationships during the training of word2vec word vectors so that the related head entity words and tail entity words are more similar to some extent. To integrate the knowledge triples instance information into the CBOW model of word2vec, we construct a triples instance \((h,r,w_i)\), where \(r\) represents a variety of different relationships associated with the word \(w_i\). Our goal is to use the TransE knowledge representation model to project the entities and relationships into a low-dimensional vector space, representing the semantic connections between the entities and relationships. For any triples \((h,r,w_i)\) fact information, there is \(h+r \approx w_i\), such as \(china + capital \approx beijing\). The triples information is used as the supervision data to make the word \(w_i\) can learn the rich semantic relationship in the knowledge graph. According to this goal, we define the objective function of the model as shown in the following equation (1).

\[
L = \sum_{i=1}^{\|V\|} \sum_{r \in R_{w_i}} \log p(w_i \mid h + r)
\]

(1)

Where \(\|V\|\) is the size of the vocabulary dictionary; \(R_{w_i}\) represents the collection of relations containing words \(w_i\); \(h\) is the words in the triples \((h,r,w_i)\) that have a relationship \(r\) with the word \(w_i\); \(p(w_i \mid h + r)\) is the probability that the known target words \(w_i\) have relations \(r\) with words and the predicted target words \(w_i\) are. The probability is calculated as following equation (2).

\[
p(w_i \mid h + r) = \frac{\exp(e_{h+r}^T \cdot \theta_{w_i})}{\sum_j^{\|V\|} \exp(e_{h+r}^T \cdot \theta_{w_j})}
\]

(2)

Where \(e_{h+r}\) represents the linear addition of vector \(e_h\) and vector \(e_r\), that is, \(e_{h+r} = e_h + e_r\); \(\theta_{w_i}\) represents the auxiliary parameter for the classification.

Taking the rich relationship information of triples instance in the knowledge graph as supervision, the word2ve's CBOW model training process is added to enable the training output word vector to

---

The text is a continuation of the previous discussion on incorporating knowledge graphs into word vector models to improve classification accuracy. The paper proposes an algorithm, Kg2vec, that integrates knowledge graph triples as supervision during the training of word vectors. This approach enhances the model's ability to capture semantic relationships and improve classification performance, especially in domain-specific tasks like aviation safety reports. The Kg2vec model is seen as a step forward in integrating knowledge bases with machine learning models for natural language processing tasks.
learn the complex semantic relationship between words. We define the objective function of the 
kg2vec model as shown in the following equation (3):

\[ L = \sum_{i=1}^{|C|} \log p(w_i \mid \text{context}(w_i)) + \gamma \sum_{r \in R_u} \log p(w_i \mid h + r) \]  

(3)

Where the left half \( \sum_{i=1}^{|C|} \log p(w_i \mid \text{context}(w_i)) \) is the objective function of word2vec's CBOW model; the right half \( \sum_{r \in R_u} \log p(w_i \mid h + r) \) is the knowledge vector representation model of the knowledge graph; \( \gamma \) is the weighted parameter used to balance the contribution rate of the two models; \( C \) is the training corpus; \(|C|\) is the size of the corpus.

2.3. Generation Process of Kg2vec Model
The kg2vec model uses the method of negative sampling for optimization and approximate calculation of the softmax function during training. The knowledge triples samples are divided into positive samples and negative samples, and the triples instances that are not in the knowledge graph are regarded as negative samples, that is, when the triplet sample \((h, r, w_i)\) is a fact, the corresponding \(w_i\) is a positive sample, and the other words are negative samples. For the relationship \(r\) between the given target word \(w_i\) and the word \(h\), the training objective function of the knowledge representation model is as following equation (4):

\[
1 \sum_{i=1}^{k} \sum_{r \in R_u \cup \text{NEG}_u} \left[ \delta(w_i \mid u) \cdot \log \sigma(e_{h+r}^T \cdot \theta_u) + \left(1 - \delta(w_i \mid u)\right) \cdot \log \left(1 - \sigma(e_{h+r}^T \cdot \theta_u)\right) \right]
\]  

(4)

Using the random gradient derivation of the word2vec, we can get the gradient update of parameters \(\theta_u, e_h, e_r\) as shown in equations (5), (6) and (7).

\[
\theta_u = \theta_u + \alpha \left( \delta(w_i \mid u) - \sigma(e_{h+r}^T \cdot \theta_u) \right) \cdot e_{h+r}
\]  

(5)

\[
e_h = e_h + \alpha \sum_{w \in \{w_i\} \cup \text{NEG}_u} \left( \delta(w_i \mid u) - \sigma(e_{h+r}^T \cdot \theta_u) \right) \cdot \theta_u
\]  

(6)

\[
e_r = e_r + \alpha \sum_{w \in \{w_i\} \cup \text{NEG}_u} \left( \delta(w_i \mid u) - \sigma(e_{h+r}^T \cdot \theta_u) \right) \cdot \theta_u
\]  

(7)

Where \(\alpha\) is the learning rate.

3. Improved Text Classification Algorithm
The improved text classification algorithm in this paper consists of three parts: text representation, neural network, and classification evaluation. Among them, the kg2vec word vector extracts semantic features, and the convolution operation layer extracts n-gram syntax features. The overall flow chart of the algorithm is shown in Figure 1.

4. Experimental and Analysis

4.1. Experimental Data
To verify the effectiveness of the above algorithm, the paper downloaded about 51394 real aviation safety reports data from January 2010 to December 2018 from the ASRS online database for the experiment. The training set and the test set are divided in the proportion of 8: 2, as shown in Table 1.
The last row in the table is the valid triples knowledge instances obtained from the aviation safety knowledge graph, totaling 96,019, which are used as the domain background knowledge base of this experiment.

| label            | Datasets size | Train  | Test  |
|------------------|---------------|--------|-------|
| Human Factors    | 12092         | 9674   | 2418  |
| Aircraft         | 9695          | 7756   | 1939  |
| Procedure        | 6471          | 5177   | 1294  |
| Ambiguous        | 6259          | 5007   | 1252  |
| Company Policy   | 2262          | 1810   | 452   |
| Weather          | 1975          | 1580   | 395   |
| Environment      | 1152          | 922    | 230   |
| ATC Equipment    | 1052          | 842    | 210   |
| knowledge triples instances |         |        | 96019 |

4.2. Experimental Design

To verify the effectiveness of the above algorithm, two groups of comparative experiments are designed in this paper.

The first group of experiments is to verify the effectiveness of the kg2vec word vector model in improving the classification accuracy of aviation safety reports. One hot, word2vec, and kg2vec word vectors are used as the text feature representation of aviation safety report respectively, and the same convolution neural network model is input to classify. Then, the validity of the kg2vec word vector to improve the accuracy of the aviation safety report classification is judged by comparing the accuracy of classification results.

The second set of comparative experiments is to verify the effectiveness of the convolutional neural network-based aviation safety report classification algorithm designed in this paper. The kg2vec+CNN method designed in this paper is compared with the classical machine learning-based classification method Support Vector Machine (SVM)[15] and Naive Bayes (NB)[16] algorithm.

4.3. Experimental Results

In this paper, accuracy and F1-score are used to evaluate the classification results. Table 2 is the experimental results with accuracy as the evaluation index. Figure 2 is the experimental results with the F1-score as the evaluation index, F1-score is the average of the model precision rate and recall rate.

| algorithms        | one-hot+CNN | word2vec+CNN | kg2vec+CNN | NB    | SVM   |
|-------------------|-------------|--------------|------------|-------|-------|
| Human Factors     | 0.771       | 0.853        | 0.914      | 0.797 | 0.876 |
| Aircraft          | 0.667       | 0.781        | 0.831      | 0.699 | 0.824 |
| Procedure         | 0.652       | 0.745        | 0.762      | 0.685 | 0.706 |
| Ambiguous         | 0.539       | 0.657        | 0.731      | 0.546 | 0.654 |
| Company Policy    | 0.565       | 0.542        | 0.589      | 0.462 | 0.518 |
| Weather           | 0.508       | 0.621        | 0.638      | 0.475 | 0.647 |
| Environment       | 0.436       | 0.580        | 0.564      | 0.411 | 0.568 |
| ATC Equipment     | 0.406       | 0.412        | 0.457      | 0.369 | 0.465 |
| Average Accuracy  | 0.568       | 0.649        | 0.686      | 0.556 | 0.652 |

The experimental results show that:

1. The classification accuracy of the method kg2vec+CNN is up to 91.4%, and the average classification accuracy is 3.6% higher than the word2vec+CNN method and 11.7% higher than the One-hot+CNN method. This shows that the improved kg2vec word vector feature representation method proposed in this paper enriches the text feature representation and helps to improve the accuracy of text classification in aviation safety reports.

2. Methods one-hot + CNN, word2vec + CNN, and kg2vec + CNN all use a convolutional neural network as the classifier, which belongs to deep learning classification model; NB and SVM belong to traditional machine learning classification algorithm. According to the experimental results in Table 2,
the average classification accuracy of NB method is the lowest; the average accuracy and F1-score of SVM method are second only to kg2vec + CNN method, but the overall performance of small data sets is the best; while kg2vec + CNN method is better for big data classification, and the average classification accuracy and F1-score are the highest.

(3) The F1-score of kg2vec + CNN was the highest (0.837), followed by SVM and one-hot + CNN (0.596). This shows that among the five classification methods designed, the model quality of this method (kg2vec + CNN) is the highest.

![Figure 1. Algorithm overall flow.](image)

**Figure 1.** Algorithm overall flow.

![Figure 2. Comparison of the classification F1-score of the five algorithms.](image)

**Figure 2.** Comparison of the classification F1-score of the five algorithms.

To sum up, kg2vec + CNN method has the highest accuracy and better model quality among the five designed methods and is suitable for text classification of big data sets. At the same time, the classification accuracy of the aviation safety reports is affected by the size of the sample data set and the characteristics of input data. Considering the reality that the ASRS has a very large number of aviation safety reports, the kg2vec + CNN classification scheme designed in this paper is applicable to solve the problem of a large number of the aviation safety reports classification.

5. Conclusion

Given the strong field-specific characteristics of aviation safety reports, this paper proposes a kg2vec word vector training model that integrates text and knowledge graph. It innovatively integrates aviation safety knowledge graph as background knowledge in aviation safety into word2vec's word vector training process, so that word vectors can make full use of domain information in the learning process to learn more rich semantic features, which helps to improve the classification accuracy of aviation safety reports. Then designed an algorithm that is more suitable for large-scale aviation safety report automatic classification, namely kg2vec+CNN. Train the unstructured text data word vector of the aviation safety reports through the kg2vec model, and use the obtained word vector with rich semantic features as the input text feature matrix of CNN. After CNN's convolution kernel pooling,
text with n-gram syntax features is obtained. The feature vector finally inputs to the softmax classifier for classification, and the classification result of the aviation safety reports is obtained. The experimental results show that the kg2vec+CNN classification algorithm designed in this paper has the highest F1-score value and has a classification accuracy of up to 91.4%. This proves that the kg2vec+CNN classification algorithm designed in this paper applies to large-scale aviation safety report classification.

6. References

[1] Oza N, Castle J P and Stutz J 2009 Classification of Aeronautics System Health and Safety Documents IEEE Transactions on Systems Man & Cybernetics Part C 39 670-80
[2] Ahmed, Mohammad Salim et al 2010 Multi-label ASRS Dataset Classification Using Semi Supervised Subspace Clustering Proc. Conf.on Intelligent Data Understanding, CIDU 2010 (California: Mountain View)
[3] Chris A, Waldemar K, and Piotr M 2012 Application of diffusion maps to identify human factors of self-reported anomalies in aviation Work 41 188-97
[4] Chunbao Yang 2008 Research on Automatic Classification of Aviation Safety Reports Based on Domain Dictionary Diss
[5] Wolfe S 2007 Wordplay: An Examination of Semantic Approaches to Classify Safety Reports American Institute of Aeronautics and Astronautics Conference pp 1-13
[6] Switzer J, Khan L and Muhaya F B 2011 Subjectivity classification and analysis of the ASRS corpus Proc. Int. Conf.on Information Reuse and Integration pp 160-5
[7] Tanguy L, et al 2016 Natural language processing for aviation safety reports: From classification to interactive analysis Computers in Industry 78 80-95
[8] Xia F, Jicun T and Zhihui L 2009 A Text Categorization Method Based on Local Document Frequency Proc. Int. Conf.on Fuzzy Systems and Knowledge Discovery
[9] Mengna L 2019 Analysis of Influencing Factors and Risk Prediction of Aviation Safety Accident Report Based on Text Mining Diss
[10] Spärck J K 2004 A statistical interpretation of term specificity and its application in retrieval Journal of Documentation vol 60 pp 493-502
[11] Mikolov T, et al 2013 Efficient Estimation of Word Representations in Vector Space Computer Science
[12] HengQi H, Juan Y, Xiao L and YunJiang X 2019 Review on Knowledge Graphs Computer systems applications 28 1-12.
[13] Ruiqi L, Jing W and Xin Z 2019 Knowledge Graph Standardization White Paper china electronics standardization institute p 185
[14] Bordes A, et al 2013 Translating embeddings for modeling multi-relational data Proc. conf.on Advances in Neural Information Processing Systems pp 2787-95
[15] Chen Shan 2012 Research of Support Vector Machine in Text Classification Future Computer, Communication, Control and Automation Springer Berlin Heidelberg
[16] Xiaoli G, Wei W and Huojun Z 2018 Several improved naive Bayes classifier models. Electronics World, 555 21 42-3