Proceedings of the 55th Hawaii International Conference on System Sciences | 2022

An Innovative Approach to Modeling Aviation Safety Incidents

Donghui Shi
Department of Computer Engineering
School of Electronics and Information Engineering
Anhui Jianzhu University
Hefei, China 230601
sdonghui@gmail.com

Shuai Cao
School of Electronics and Information Engineering
Anhui Jianzhu University
Hefei, China 230601
supercshuai@163.com

Jozef Zurada
Department of Information Systems,
Analytics, and Operations
College of Business
University of Louisville
Louisville, KY 40292
jozef.zurada@louisville.edu

Jian Guan
Department of Information Systems,
Analytics, and Operations
College of Business
University of Louisville
Louisville, KY 40292
jeff.guan@louisville.edu

Abstract

Due to the complexity of aviation safety operations, the number of flight incidents continues to rise. The Aviation Safety Reporting System (ASRS) contains the largest collection of such incidents. Efficient and effective analysis of these incidents remains a challenge. This paper proposes a new approach to analyze aviation safety records using deep learning methods to improve incident classification. The proposed approach, CNN-LSTM, combines the characteristics of convolutional neural network (CNN) and long short-term memory (LSTM) neural network, and a distributed computing method to model aviation safety data. The five machine learning methods Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, Multi-layer Perceptron were used to compare with CNN-LSTM. The results show that CNN-LSTM model can significantly improve the accuracy rates of classification for aviation safety incident reports using Word2Vec. The distributed platform in Spark with clusters can make full use of computing resources when processing textual data from ASRS, reducing time-consumption greatly when compared with machine learning algorithms running on a standalone computer. Timely and accurate identification of causes of reported incidents is important. The results of this study demonstrate a new approach to improve both accuracy and efficiency in incident cause identification.

1. Introduction

Safety is critical to aviation industry. The aviation industry still faces challenges in creating normal and safe operations. As the complexity of aviation operations increases, the number of flight accidents/incidents also increases. As a result, effective retrieval and analysis of aviation safety data to reduce incidents remains a challenge. Aviation incidents can be caused by a variety of factors. Identifying true causes is made more difficult by the fact that many relevant data fields have missing data in the ASRS database. In the last few decades a research stream has emerged that model and predict causes of aviation incidents.

In 1998, Fullwood et al. [1] used Linear Regression model to predict the aviation safety trends in aviation service reports. In 2001, Nazeri and Bloedorn [2] used the association rule method to analyze the ASRS reports, and proved the feasibility of the association rule method in the analysis of aviation safety data. In 2004, Majumdar et al. [3] used the trend analysis method to analyze and predict the unsafe factors that caused aviation incidents. In 2006, Nazeri [4] used an analysis algorithm (an abnormal distribution algorithm) AF (attribute focusing) algorithm to analyze and mine data for difficult-to-find abnormalities.

Koteeswaran et al. [5] proposed an aviation accident prediction method that combines *k*-Nearest Neighbor (*k*-NN) and correlation-based feature selection method. This new method can detect risks by predicting the causes of accidents and improve the aviation management system. The main purpose of analyzing accident data is to explore the causes of accidents and prevent accidents in the future. Rao and Marais [6] proposed a state-based method by defining a grammar describing states and trigger sequences. The result shows that rule-based method can result in better statistics on the cause of the accident. Hegde and Rokseth [7] pointed out that different methods

URI: https://hdl.handle.net/10125/79482
978-0-9981331-5-7

dealing with different information may be combined to have an outstanding prospective in aviation accident. Altay et al. [8] used genetic algorithms and artificial neural networks to predict the age and types of aircrafts as these two factors contribute to accidents. In the prediction of aviation equipment failure, Castilho [9] used the experience of maintenance workers to construct variables, and then used these variables as the input of the Bayesian network, and as a result, they obtained improved prediction results.

In recent years, some research began to use big data including deep learning and data stream methods to predict the risks in aviation industry. Odarchenko et al. [10] pointed out the challenges brought about by the current big data technology in aviation system application and proposed a feasible plan to transform from relational database to non-relational database. Subramanian & Rao [11] used Go-around (GAR) and Missed-approach (MA) data from the Aviation Safety Reporting System (ASRS) incident database and trained Long Short-Term Memory (LSTM) network to predict which categories of incidents are more (or less) likely to occur in the forecast period. This prediction helps to identify the factors that lead to the accident. Incident reporting and investigation are components of safety management. Shi et al. [12,13] applied data stream methods incrementally to build and test classification models for risk factor identification for ASRS. The results demonstrated that data stream method can be a viable approach to automated incident type identification and the use of text-mining and data-streaming technologies can improve safety management systems.

Although data stream methods were verified to be better or comparable to the traditional machine learning methods [12,13] and some researchers just began to use deep learning methods to analyze safety reports [14], two problems still remain to be solved, i.e., the prediction accuracy rates need to be improved and run time of the algorithms need to be reduced as natural language processing tasks required in processing incident reports and subsequent modeling can be very resource intensive. To the best of our knowledge, there is very little or no literature on distributed framework for processing ASRS data sets.

In order to solve the two problems, we will explore the use deep learning methods and distributed platform to process the textual data from ASRS, and construct models to classify the incidents. The paper is organized as follows. Section 2 provides the data description. Section 3 presents the methods used in this paper, including feature selection methods, neural networks based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), and distributed Clusters and Spark architecture. Section 4 discusses the results of the experiments. Finally, Section 5 concludes the paper.

2. Data Description

The Aviation Safety Reporting System (ASRS) is provided by the U.S. National Aviation Safety Data Analysis Center. It includes many confidential aviation incident reports, and the reports were collected from the aircraft crew, flight attendants, maintenance personnel, and air traffic control personnel. Each record in the reports consists of structured numeric and text fields such as the date and hour of the incident, type of aircraft, personnel, etc. as well as unstructured text data, i.e., the description of incidents entered by flight and ground personnel. These narratives provide valuable information that help to determine the cause of incidents. Therefore, in computer simulation we only use unstructured textual data.

The data set in this study contains 158,070 incident records in which incident types were manually classified by human experts reading the reports. There are 97,481 incidents attributable to human factors, accounting for about 61.67% of the total number of incidents. There are 31,796 incidents caused by aircraft-related factors, accounting for 20.12% of the incidents. Others factors such as weather, ambiguity, and company policy occurred in relatively small numbers. One can see that human factors are the cause of more incidents than any other factors combined. Therefore, in this study, identifying human factors is our main objective, and the other factors are categorized as non-human factors. We will classify presence of human factors and nonhuman factors for incidents reports. Figure 1 shows the proportion of incidents types.

3. Methods

In the study, we used Sklearn running on a standalone computer and Spark with clusters. PySpark is an interface for Apache Spark in Python and it is used in the distributed environment in the study. Sklearn (scikit-learn) is a library for machine learning algorithms in Python. Figure 2 shows that the architecture of identification model of human factors in aviation incidents. In the figure, Hadoop Distributed File System (HDFS) is designed to reliably store very large files across machines in a large cluster.

3.1 Feature selection methods

Feature engineering is a process of identifying relevant input variable in the original data. The process generally consists of three parts: feature processing, feature selection and dimensionality reduction. Feature processing includes a series of steps such as data selection and cleaning. In text mining, it mainly refers to removing special characters, removing stop words, and case conversion. This step was carried out with the Python natural language toolkit NLTK (Natural Language Toolkit). The vector space model was used to convert text into a vector. In the study two methods, Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec were used to extract the structured information from textual data.

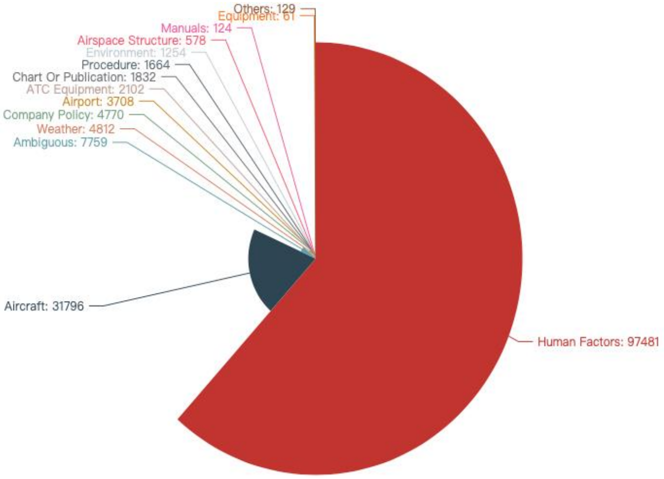

Figure 1. The proportion of incidents types

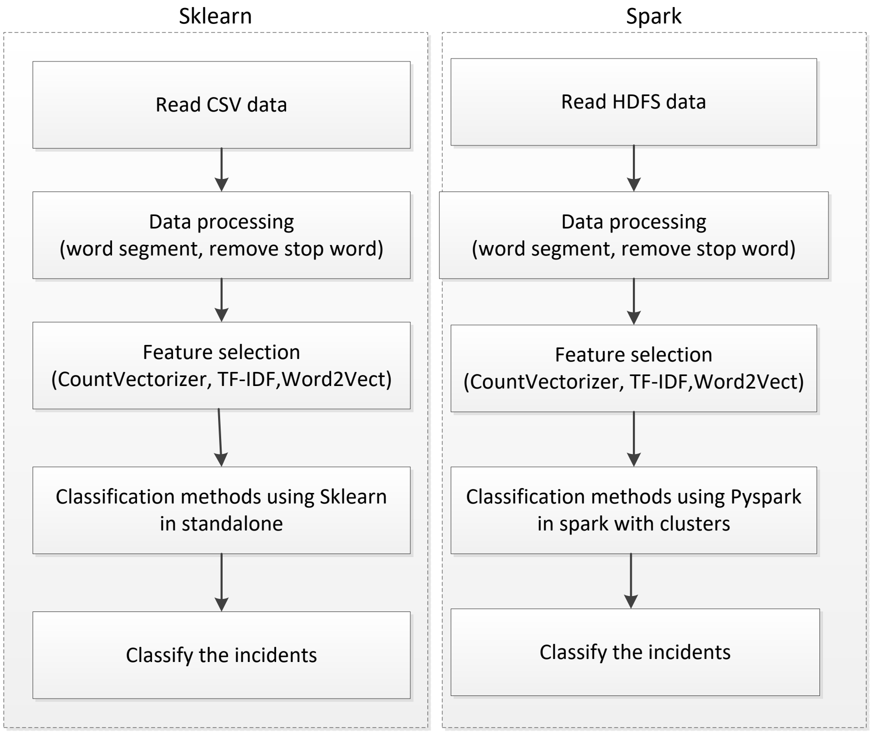
Sklearn
Spark
Read CSV data
Read HDFS data
Data processing
(word segment, remove stop word)
Data processing
(word segment, remove stop word)
Feature selection
(CountVectorizer, TF-IDF,Word2Vect)
Feature selection
(CountVectorizer, TF-IDF,Word2Vect)
Classification methods using Sklearn
in standalone
Classification methods using Pyspark
in spark with clusters
Classify the incidents
Classify the incidents

Figure 2. The architecture of identification of human factors in aviation incidents

3.2 CNN-LSTM

Neural networks based on Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) have their distinct advantages in classification tasks [15]. Convolutional neural networks can be used in mining the potential semantic information of textual data. The multi-convolution kernel performs convolution operations on the word vectors of the text. LSTM networks are well-suited to classifying, processing and making predictions based on time series data. In the field of text processing, CNN-LSTM neural network [16], are known to produce good results. In the study, we built a CNN-LSTM neural network model to classify aviation incidents. The structure of the model is shown in Figure 3.

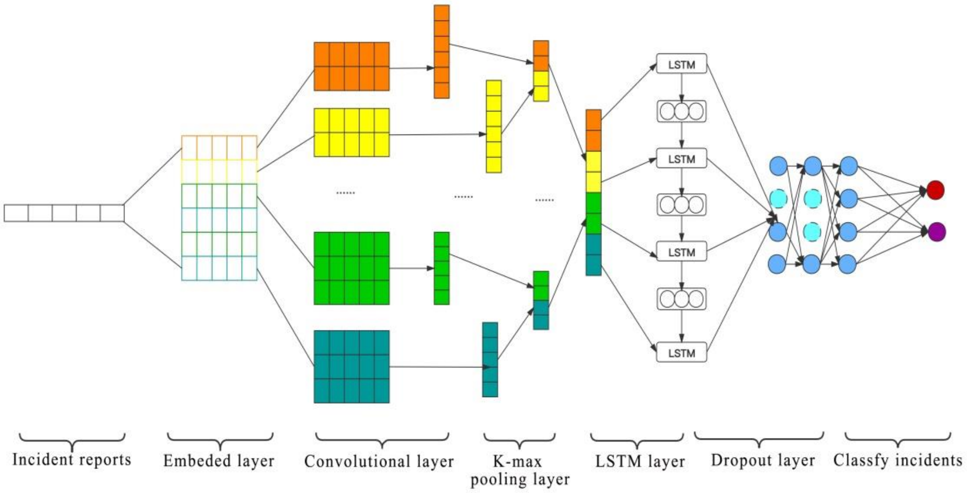

Figure 3. Text classification of CNN-LSTM model

CNN is widely used in image data, time series data processing and other fields. The network structure has the characteristics of non-full connection and parameter sharing. Compared with the fully connected network, the network complexity and the number of weights in CNN are greatly reduced. The core of CNN consists of the following parts, input and output layer, convolutional layer, pooling layer and fully connected layer.

LSTM neural network is an extension of RNN, which solves the problem of long-term dependence, especially when dealing with text data. It can predict the probability of the next word through the semantic context information of the text. The cell state in the network model is the core of the LSTM network, which is somewhat similar to a conveyor belt. Figure 3 is the LSTM neural network mechanism. LSTM uses the structure of gates to select information, and gates are usually composed of sigmoid functions. Since the result of the sigmoid output value is between 0 and 1, then 0 and 1 can be used to indicate two states, 0 means fail, and 1 means pass. There are three types of gate states: input gate, output gate and forget gate.

3.3 Clusters and Spark architecture

Clusters are formed when independent computers are connected over a network to solve larger computing tasks. Clustering has high scalability and reliability. The disadvantage is that the communication time cost increases with the increase of computing nodes. However, when processing large-scale data, the running time of the algorithm model is longer, and the proportion of the communication time in the total running time gradually decreases, so the communication time is insignificant.

In recent years, with the application of big data technology, Spark is often regarded as the first choice for a big data computing platform. At present, Spark's functions have covered a wide range of computing fields, such as machine learning, streaming/real-time computing, and graphics processing. Its advantages include: fast speed, memory-based computing; and ease of use. Spark also provides rich interfaces and supports many programming languages including Java, Python, and Scala.

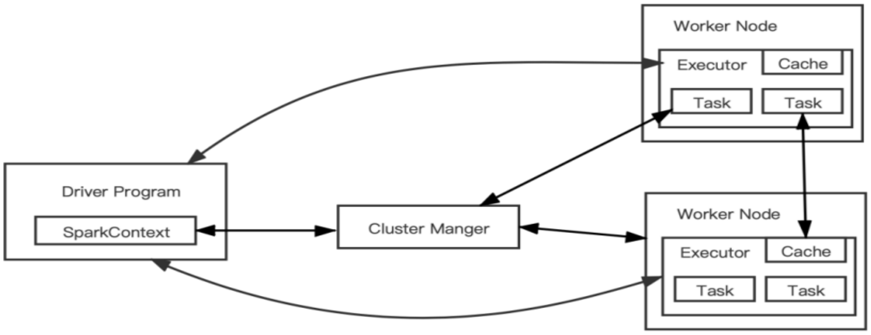

Figure 4. Spark architecture

Figure 4 shows the overall architecture of Spark. The user codes for data processing through Driver Program, and creates a SparkContext object by running the main() function, through which the interaction between the user and the cluster is realized. The Cluster Manager in the middle of the figure is specifically used to manage resource scheduling. It now supports Local, Standalone and Yarn modes. Cluster Manager will start Executor while allocating computing resources. In Executor, each computing unit is called Task, and each computing node in the cluster is called Worker Node. The start of the thread pool is also completed by Executor. The main task of the thread pool is to manage the running status of the Task. The Executor will eventually report the running status of the Task to the Driver.

4. The experiment design and results from computer simulation

In the distributed environments, we use one master and 1, 2, 4 and 6 slave nodes. The operating system is selected as Oracle Linux Server release 7.4, the Hadoop is 2.6.0 version, and the Spark version is 2.2.0. CPU of the master node and slave nodes is Intel Xeon E5-2683 v4, the memory of the master node is 64G and the memory of slave node is 16G.

The classification performance and running time of the classification algorithms for Sklearn in Spark in standalone mode and in Spark with clusters in different sample sizes are shown in Figures 5 through 13, where LR, NB, RF, SVM, MLP and CNN-LSTM represent Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and CNN-LSTM combined model; feature selection methods are TF-IDF and Word2Vec. Four groups of data (1000, 10,000, 100,000, 150,000) were randomly generated.

The classification accuracy rates of CNN-LSTM are affected by the parameter settings. Through experiments, the model has the best results when the dimensions of the word vector selected are 256; the number of convolution kernels is set to 128; the number of CNN hidden layer nodes is 128; the numbers of LSTM hidden layer nodes are set to128; the function is selected as cross-entropy; the optimization function selected is Adam (a replacement optimization algorithm for stochastic gradient descent). In order to enhance the generalization ability of the model and prevent over-fitting of the data, a Dropout layer is added between the LSTM and the fully connected layer. When the value of Dropout is set to 50%, the accuracy rates of the models are the highest.

Figures 5 and 6 compare the performance of classification algorithms based on TF-IDF and Word2Vec representation in Sklearn, Spark with single node and Spark with clusters (4 slave nodes). The experimental sample data used the data set with 100,000 as the data size is moderate.

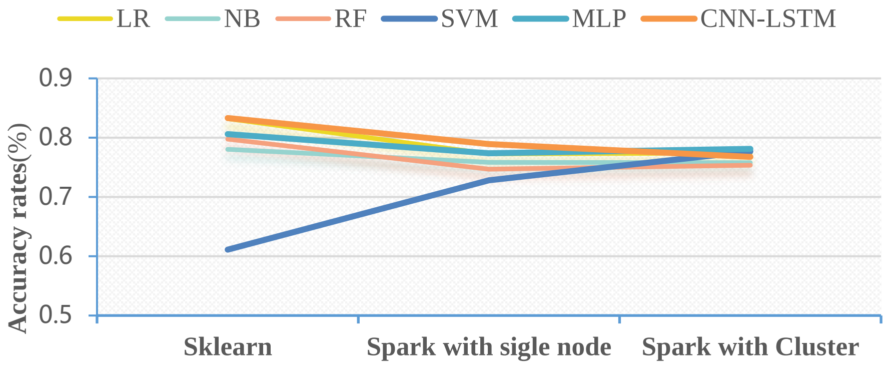

Figure 5. The results of the models based on TF-IDF

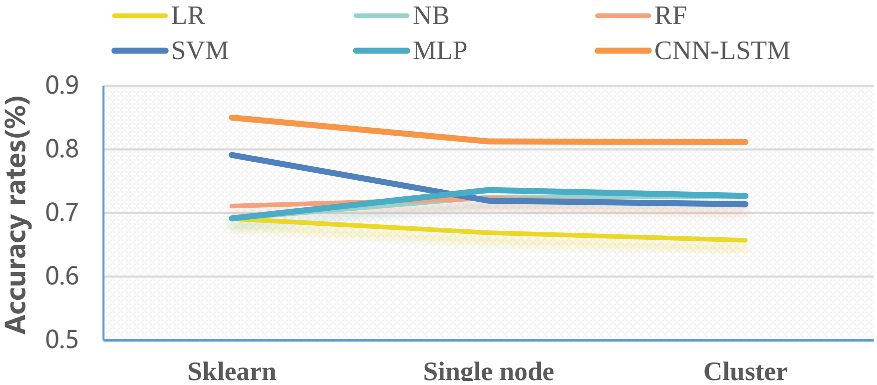

Figure 6. The results of the models based on Word2Vec

In Figure 5, one can see that when the features are represented by TF-IDF, the accuracy rates of the methods using Sklearn except for SVM are higher than those using Spark with single node and Spark with cluster. In Figure 6, when the features are represented by Word2Vec, the accuracy rates of the classification algorithms using Sklearn are better than those using Spark with single node and Spark with clusters. A possible reason for the better performance of Sklearn may be the different processing strategy for Sklearn and Pyspark. The accuracy rates for the same classification algorithms using the Sklearn are slightly higher than that of Pyspark. Figures 5 and 6 show that the different feature selection methods have an impact on the accuracy rates of the classification model. On the whole, the accuracy rates for TF-IDF are slightly higher than those for Word2vec.

In order to compare the performance of the classification algorithms based on data sets with different sizes, the whole data set is divided into four groups: 1000, 10,000, 100,000, and 150,000. Figures 7, 8, 9, and 10 show the results for TF-IDF and Word2Vec for the Sklearn and Spark using the six slave nodes on four different data sets.

Figures 7 and 8 show the results using the TF-IDF and Word2Vec in the Sklearn, and Figures 9 and 10 show the results using the TF-IDF and Word2Vec in Spark clusters. For Sklearn, as the size of sample data increases, the overall accuracy rates of the models show an upward trend. The accuracy rates of the logistic regression model in Figure 6 in the four sample data sets are 0.785, 0.8195, 0.832, and 0.8113, respectively. When the sample size is from 1,000 to 10,000, the accuracy rates of the four classification algorithms NB, SVM, MLP and CNN-LSTM declined. For example, the accuracy rate of the SVM decreases, and the accuracy rate dropped from 0.845 to 0.613.

Figures 9 and 10 show the results for TF-IDF and Word2Vec in Spark using the six slave nodes on different numbers of data sets. As the size of data increases, the accuracy rates of the classification models using TF-IDF and Word2Vec in Spark with clusters increase as well.

It can be seen from Figure 7 to Figure 10 that the CNN-LSTM model using Word2Vec has the best performance. Figure 11 shows the comparison of the accuracy rates of the CNN-LSTM model for Word2Vec. In the figure, we use 4 slave nodes in the distributed environment.

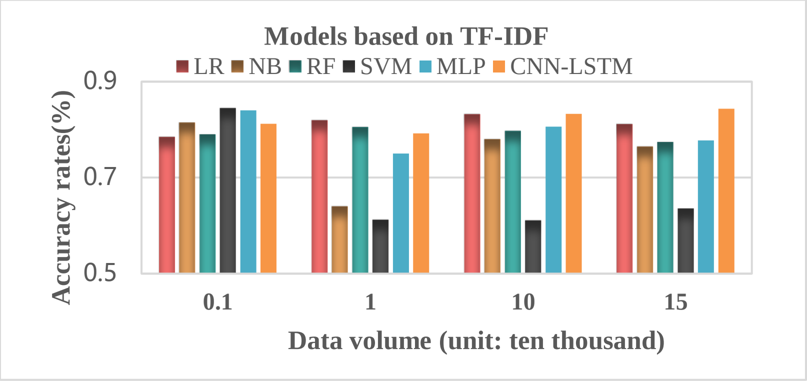

Figure 7. The accuracy rates using the TF-IDF in Sklearn

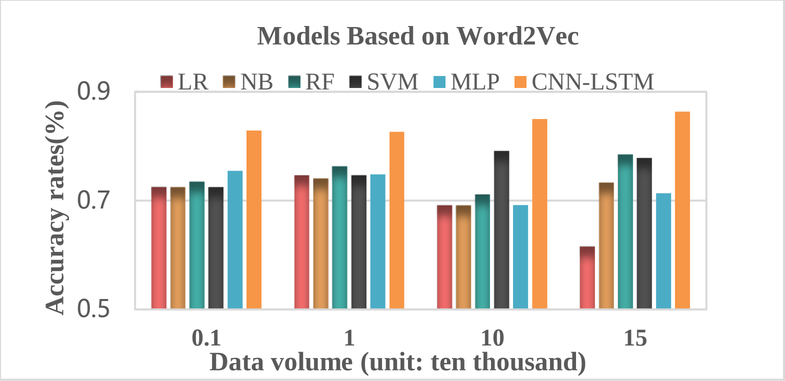

Figure 8. The accuracy rates of using Word2Vec in Sklearn

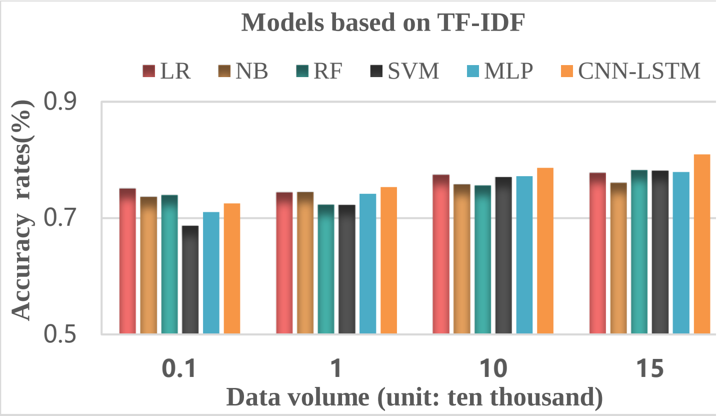

Figure 9. The accuracy rates of using TF-IDF in Spark with clusters

It can be seen from Figure 11 that as the size of data increases, the accuracy rates curve of the CNN-LSTM for Sklearn is always above the curve for Spark with single node and Spark with 4 slave nodes. The reason for this phenomenon could be due to the fact that Sklearn and Pyspark have different data processing strategies. From the figure, one can see that the accuracy rate curve for the Spark single-node almost overlaps those for Spark with 4 nodes.

Figure 12 and Figure 13 compares the running time of the models for the Sklearn and the Spark clusters with different nodes. In the two figures, the data set with150,000 were used. The experimental environment includes Sklearn, Spark with single node, Spark with 2 slave nodes, Spark with 4 slave nodes, and Spark with 6 slave nodes.

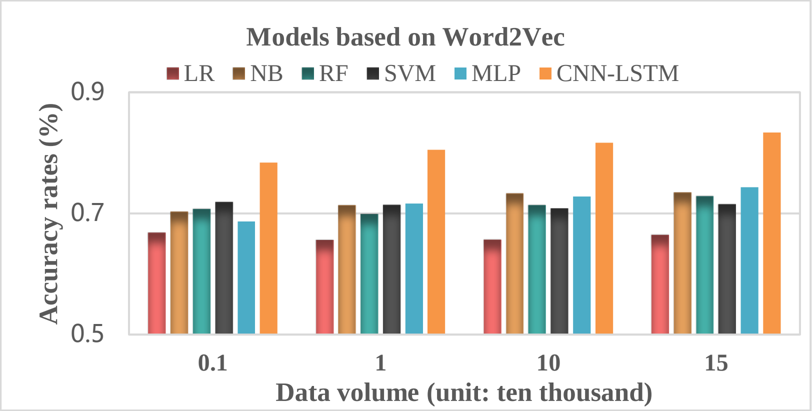

Figure 10. The accuracy rates of using Word2vec in Spark with clusters

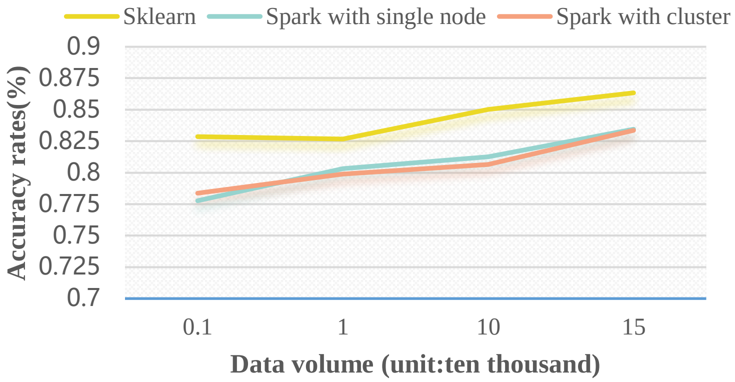

Figure 11. The accuracy rates of using Word2vec in Sklearn, Spark with single node and spark with 4 nodes

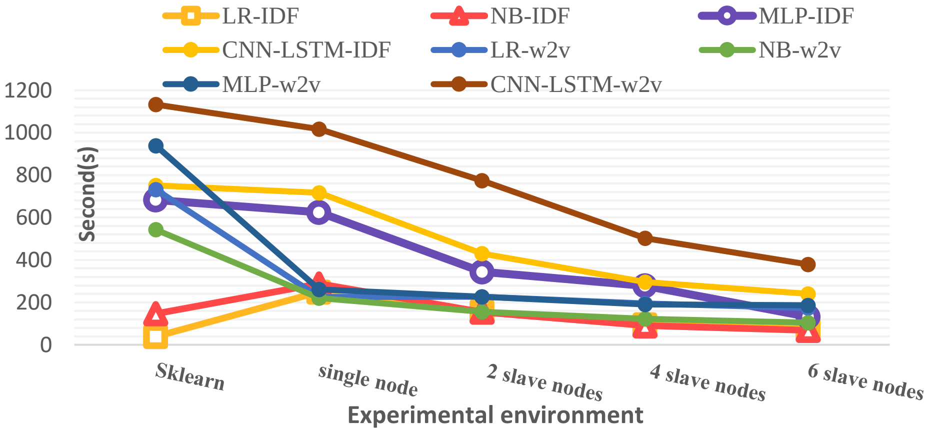

Figure 12. The running time of the 4 models (for IF-IDF and Word2vec) with the different experimental environment

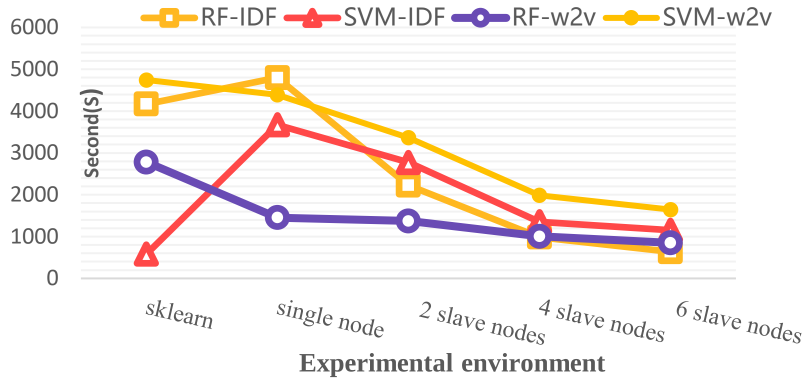

Figure 13. The running time of the 2 models (for IF-IDF and Word2vec) with the different experimental environment

In Figures 12 and 13, LR-IDF represents a logistic regression model using TF-IDF, LR-w2v represents a logistic regression model using Word2Vec, and so on. It can be seen from the figures that the overall running time of RF and SVM is higher than that of the other four models. The running time of different classification models for Sklearn and Spark clusters are different. The running time of the LR-IDF and SVM-IDF models in the Sklearn environment is lower than that of the Spark with clusters. In the Spark with clusters, the running time of the models decreases gradually with the number of cluster nodes, especially between 4 nodes and 6 nodes, which is approximately a smooth straight line. On the one hand, the running times of the models using Word2Vec are higher than those using the TF-IDF model. For example, the running time of CNN-LSTM-w2v is 938.91s, 260.66s, 226.6s, 190.62s, and 186.43s, for Sklearn, Spark with single node, Spark with 2 slave nodes, Spark with 4 slave nodes, Spark with 6 slave nodes, while the running time of CNN-LSTM-IDF is 1132.59s, 1016.67s, 774.14s, 502.36s, and 378.76s.

5. Conclusion

In the paper, we classified the human factors and the nonhuman factors from aviation incidents using LR, NB, RF, SVM, MLP and CNN-LSTM in standalone and distributed environment with data sets of sizes 1,000, 10,000, 100,000, and 150,000. Two feature selection methods TF-IDF and Word2Vec were used to extract relevant incident type data from the aviation incident reports. Then six models were tested to assess their potential in classifying the incidents in Sklearn, Spark with single node, and Spark with clusters. Overall accuracy rates and running time are used to measure the performance of these models.

Our results show that the accuracy rates of the models in Sklearn are higher than those in the Spark with clusters. CNN-LSTM using Word2Vec is the best in classifying these incidents. Generally, as the number of samples increases, the overall accuracy rates increases in Spark with clusters using TF-IDF and Word2Vec. The accuracy rates of some models will fluctuate with data sizes, and the rates of CNN-LSTM models always perform better when data size increases. It shows that CNN-LSTM has a better stability and generalization ability. The accuracy rates of the CNN-LSTM model are affected by the word vector's dimension, the number of convolution kernels, the number of hidden layer nodes and other parameters. The optimal dropout parameters are selected through comparison experiments.

In addition, the models using the TF-IDF consume less time compared with the models using the Word2Vec, and LR-IDF and NB-IDF consume relatively less time, while RF-IDF and SVM-IDF consume more time. Although in Sklearn and Spark with single node, CNN-LSTM model consumes more time, the models in Spark with 2, 4, 6 slave nodes consume less time. When processing a small data set, the models in Sklearn in standalone mode have obvious advantages, and the models take less time compared with the models in the Spark with clusters. Due to time-consuming communication between data partitions, the models in Spark with single node and Spark with clusters will take more time. As data size increases, the running time of the models in Spark with clusters will decrease, which is preferable when processing aviation incidents with large amount of text data.

The models presented in this paper can automatically classify the cause of an incident as either caused by human factors or caused by nonhuman factors, without manual and time-consuming involvement of human experts. Using variable reduction one can also find, in the textual data describing the incidents, the major factors that influence the prediction. This may help to find the causes of incidents and reduce the occurrence rates of incidents. The research described in this research has practical implication. An accurate prediction model can help identify the true cause of incidents. Incidents occur at a higher frequency. When incidents are reported, the causes of the incidents are not known. It is important that we identify accurately and timely the cause of each reported incident. We feel our study shows an effective alternative to improve both the accuracy and efficiency of the incident cause identification process.

References

[1] Fullwood, R.R., Hall, R.E., Martinez-Guridi, G., Uryasev, S., & Sampath, S.G. (1998). Relating aviation service difficulty reports to accident data for safety trend prediction. *Reliability Engineering & System Safety*, 60(1), 83-87.

[2] Nazeri Z, Bloedorn E. (2004) Exploiting Available Domain Knowledge to Improve Mining Aviation Safety and Network Security Data. *Proceedings of the ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies: 24 September 2004; Pisa*.

[3] Majumdar A., Ochieng, W.Y., & Nalder, P. (2004) Trend analysis of controller-caused airspace incidents in New Zealand, 1994-2002. *Transportation research record*, 1888(1): 22-33.

[4] Nazeri, Z. (2010). Data mining of air traffic control operational errors. *Conference on Data Mining (DMIN'06)* 321-324

[5] Koteeswaran, S., Malarvizhi, N., Kannan, E., Sasikala, S., & Geetha, S. (2019). Data mining application on aviation accident data for predicting topmost causes for accidents. *Cluster Computing*, 22(5), 11379-11399.

[6] Rao, A.H., & Marais, K. (2020). A state-based approach to modeling general aviation accidents. *Reliability Engineering & System Safety*, Volume:193, Article Number:106670.

[7] Hegde, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment–A review. *Safety science*, Volume:122, Article Number 104492.

[8] Altay, A., Ozkan, O., & Kayakutlu, G. (2014). Prediction of aircraft failure times using artificial neural networks and genetic algorithms. *Journal of Aircraft*, 51(1), 47-53.

[9] Castilho, I.X. (2015) Fault prediction in aircraft tires using Bayesian Networks. Diss. *Master's thesis*, Instituto Superior Técnico, Portugal.

[10] Odarchenko, R., Hassan, Z., & Zaman, A. (2019). Use of Big Data in Aviation: New Opportunities, Use Cases, and Solutions. In *Automated Systems in the Aviation and Aerospace Industries* (pp. 436-452). IGI Global.

[11] Subramanian, S.V. & Rao, A.H. (2018). Deep-learning based Time Series Forecasting of Go-around Incidents in the National Airspace System. 2018 *AIAA Modeling and Simulation Technologies Conference*. Kissimmee, FL, USA, January 2018.

[12] Shi, D., Zurada, J., & Guan, J. (2017). Identification of human factors in aviation incidents using a data stream approach. Proceedings of the *50th Hawaii International Conference on System Sciences*, pp. 1073-1082.

[13] Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A data-mining approach to identification of risk factors in safety management systems. *Journal of Management Information Systems*, 34(4), 1054-1081.

[14] Zhong, B., X Pan, Love, P., Sun, J., & Tao, C. (2020). Hazard analysis: a deep learning and text mining framework for accident prevention. *Advanced Engineering Informatics*, Volume 46, Article Number 101152.

[15] Islam, M.Z., Islam, M.M. & Asraf A. (2020) A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in medicine unlocked*, Volume:20 Article number: 100412.

[16] Rehman, A.U., Malik, A.K., Raza, B. et al. (2019) A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimed Tools Appl*, 78, 26597–26613.