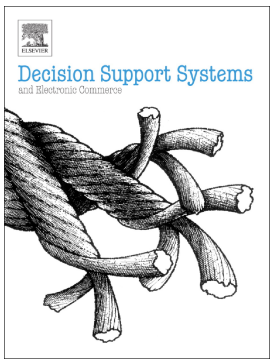
Accepted Manuscript



Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud

Yibo Wang, Wei Xu

PII: S0167-9236(17)30213-0
DOI: doi:10.1016/j.dss.2017.11.001
Reference: DECSUP 12895

To appear in: *Decision Support Systems*

Received date: 8 May 2017
Revised date: 10 November 2017
Accepted date: 12 November 2017

Please cite this article as: Yibo Wang, Wei Xu , Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Decsup(2017), doi:10.1016/j.dss.2017.11.001

Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud

Yibo WANG[a], Wei XU[a, b, *]

[a]*School of Information, Renmin University of China, Beijing, 100872, P.R. China*

[b]*Smart City Research Center, Renmin University of China, Beijing, 100872, P.R. China*

Abstract

Automobile insurance fraud represents a pivotal percentage of property insurance companies' costs and affects the companies' pricing strategies and social economic benefits in the long term. Automobile insurance fraud detection has become critically important for reducing the costs of insurance companies. Previous studies on automobile insurance fraud detection examined various numeric factors, such as the time of the claim and the brand of the insured car. However, the textual information in the claims has rarely been studied to analyze insurance fraud. This paper proposes a novel deep learning model for automobile insurance fraud detection that uses Latent Dirichlet Allocation (LDA)-based text analytics. In our proposed method, LDA is first used to extract the text features hiding in the text descriptions of the accidents appearing in the claims, and deep neural networks then are trained on the data, which include the text features and traditional numeric features for detecting fraudulent claims. Based on the real-world insurance fraud dataset, our experimental results reveal that the proposed text analytics-based framework outperforms a traditional one. Furthermore, the experimental results show that the deep neural networks outperform widely used machine

[*]Corresponding author at: School of Information, Renmin University of China, Beijing, 100872, P.R. China.
Email address:wangyibo90@yeah.net (Y. Wang), weixu@ruc.edu.cn (W. Xu)

learning models, such as random forests and support vector machine. Therefore, our proposed framework that combines deep neural networks and LDA is a suitable potential tool for automobile insurance fraud detection.

Keywords: Insurance Fraud, Fraud Detection, Text Analytics, Topic Modeling, Deep Learning

1. Introduction

As a type of non-life insurance, automobile insurance is a subject of property insurance for motor automobiles. Mainly responsible for various losses due to natural disasters and motor automobile accidents, automobile insurance is a means of transport insurance, including motor automobile damage insurance and motor automobile third-party liability insurance.

Accompanying the rise of the automobile population, automobile insurance has gradually become an important industry related to the development of the global economy and people's livelihoods. With increasing confidence in the optimistic development of the insurance industry, there will be more capital entering the insurance market. For that reason, the competition between insurance companies will be extremely fierce. Therefore, reducing costs and maintaining a lead over the competition are the focuses of the various insurance companies. However, insurance fraud represents a pivotal percentage of insurance company costs. Insurance fraud not only reduces the insurance company profits, resulting in substantive losses, but also affects the insurance company’s pricing strategy and social economic benefits in the long term. For example, according to Insurance Bureau of Canada statistics, the Canadian automobile insurance fraud and related offenses reached 542 million Canadian

Dollars in 2007, and statistics of the U.S. Coalition against Insurance Fraud illuminate that the amount of insurance fraud accounts for 17% to 20% of total insurance company compensation. In China, insurance regulators estimate that the proportion of insurance fraud is approximately 20%, and at the same time, China's automobile insurance claimed 175.09 billion RMB in 2011. Thus, insurance fraud is a worldwide problem, and it will have adverse effects on the state and society.

At present, the risk control and fraud detection ability in automobile insurance companies is still relatively weak [1]. The losses caused by automobile insurance fraud are not only an increase in temporary losses but also a serious impact on the development of insurance companies. Solvency will decrease as the fraudulent claims possess more and more bonus, which are supposed to provide compensation for legal claims and financial support for new business. Therefore, how to correctly identify risk factors and reduce the losses caused by fraudulent claims is a crucial problem that insurance companies urgently need to solve.

Generally, the experience of experts plays a critical role in the judgment of whether a claim is fraudulent [2]. On the one hand, the number of experts is negligible compared to the increasing number of claims. Therefore, the descriptions of cases can hardly be sufficiently extracted, analyzed and judged by the relatively few experts. As a consequence, the average time that experts spend reviewing a case is not sufficient. Moreover, the lack of experience may lead to bias in judgment, and the initiative of experts will cause deviation. Even for the same case, the judgments of different experts may be totally different because their points of view are different. On the other hand, some practitioners and academic researchers have made great efforts in employing data mining algorithms to detect automobile insurance fraud [3].

These data mining algorithms, such as SVM, are mainly modeled on numerical data. However, a wide variety of "fixed" data analysis has been implemented to solve the problems arising from "unfixed" policy holders. Data analysis has been so stressed and emphasized that the accumulated descriptions and experiences have been completely neglected. As far as that is concerned, the performance of data mining algorithms is relatively effective compared to that of experts in recognizing existing fraudulent claims, with machine learning and analysis being more efficient on known fraud types. Nevertheless, data mining algorithms are prone to misjudge in the case of masked numeric data forged by an experienced deceiver. When the data contains text information, the situation will be much better. However, thus far, there have been few research studies incorporating text information into the detection model. Consequently, both data analysis and experience are critical in designing an efficient automobile insurance detection model. Therefore, this paper proposes an automobile insurance fraud detection model based on text mining with Latent Dirichlet Allocation (LDA), combining the precision of data mining methods and the experience of human experts involved in text data. The model not only embodies the numerical attributes, such as time and driving experience, but also gives full consideration to the text description of claims. With the aid of text information and LDA, the model can better detect automobile insurance fraud.

The remainder of the paper is structured as follows. A brief introduction to prior studies is given in section 2. Section 3 explains the technical foundation. The principles of LDA are concisely introduced to support our analysis. Next, section 4 describes how our proposed method works. In section 5, the major variables used in the model are described and employed to examine the proposed method. The result analysis and conclusions are also

summarized in section 5.

2. Related Work

2.1 Insurance fraud detection

In recent years, insurance fraud detection has garnered large amounts of attention because a range of fraudulent methods have brought great loss to insurance companies and society as a whole. Insurance fraud detection is a branch of financial fraud detection. It mainly includes automobile insurance fraud detection and healthcare insurance fraud detection [1]. Automobile insurance fraud detection has attracted more attention than other financial fraud detection [4]. Canadian automobile insurance companies suffered a loss of more than half a billion Canadian Dollars because of automobile insurance fraud in 2007, while approximately one-fifth of insurance company compensation has been paid out to fraud in America every year. Moreover, according to reports, approximately 21%-36% of automobile insurance claims involve factors of suspected fraud, but only less than 3% of them are prosecuted [3]. According to Ngai et al. [4], approximately one-third of data mining applications are related to automobile insurance fraud detection in the field of financial fraud monitoring. To conquer the problem of insurance fraud, researchers have invested great effort into finding effective fraud indicators and methods.

2.2 Feature engineering in automobile insurance fraud detection

Fraud indicators play a critical role in insurance fraud detection. Appropriate indicators definitively make it possible for detection methods and algorithms to maximize the effectiveness of detection. To examine the indicators systematically, some scholars have sorted the indicators into several groups according to their meanings, while other scholars

have implemented special treatments on the indicators.

Weisberg and Derrig discussed a number of fraud elements related to deliberate build-ups to find appropriate ways to specify the responsibility of insured drivers and identify suspicious claims [5]. They put forward that conditions unrelated to accidents and misrepresentation of materials should raise attention in the specification of fraud indicators. Moreover, Brockett et al. used up to 65 fraud indicators and divided them into 6 categories: characteristics of the accident, the claimant, the insured, the injury, the treatment and the lost wages [6]. Similarly, Artís et al. sorted the fraud indicators into three groups: the accident information, the insured driver information and the automobile information [7]. They suggested that occurrence time and location have an important influence on the existence of fraud. According to their research, accidents that occurred at night, on the weekend or in non-urban areas have a relatively large possibility of being fraudulent. Meanwhile, Wilson grouped the fraud indicators into another three categories: General Indicators, Automobile Indicators and Ownership Indicators [8]. Each category has its own unique features for assessing the claims.

Šubelj et al. put forward that insurance claims data can be represented by networks, in which vertices denote entities, such as drivers or automobiles, and edges denote relations between entities [9]. They divide every attribute into either intrinsic attributes or relational attributes and let the relational ones show the relations between entities. Finally, they detect fraudulent claims by employing the Iterative Assessment Algorithm (IAA), which uses the two types of attributes to give a suspicion score to each participant of the accident. Karamizadeh and Zolfagharifar sought to find important fraud factors by employing

clustering algorithms [10]. According to their study, surplus commitment and physical commitment are considered to be the most important fraud indicators in detecting fraudulent claims. Not only the indicators themselves but also the number of indicators can have an impact on the detection. Weisberg and Derrig illustrate that there exists a strong relationship between the number of fraud indicators and the suspicion rating [11]. More fraud indicators amplify the percentage of strong suspicion of fraudulent claims.

These studies conduct an in-depth analysis of numerical and categorical fraud indicators, including the time, location, automobile, policy and relevant personnel information, which can be derived from the standard structured insurance documents and claims. Nonetheless, the lack of expert experience makes the fraud detection methods prone to being cheated on by skillful deceivers in terms of standard structured data, because standard structured data is relatively more likely to counterfeit. Therefore, expert experience, usually reflected in text data, should be taken into consideration in insurance fraud detection.

2.3 Automobile insurance fraud detection models

Great efforts have been made to seek appropriate models for automobile insurance fraud detection. In early studies, insurance fraud detection is conducted primarily through manual audits. For example, Weisberg and Derrig recruited and trained experienced claims adjusters to specify the responsibility of insured drivers and identify suspicious claims [5]. Each adjuster needed extensive training and audited only 50 files in a two-week period. Unfortunately, the proliferation of automobile insurance has greatly increased, and manual auditing cannot effectively address the consequent rapid rise in the number of claims. Moreover, Artís et al. proved that auditing technology is prone to missing some fraudulent

claims, and the proportion of missing ones is relatively large [7]. Accordingly, a great number of detection methods and systems using computers have been proposed. First, traditional statistical methods, such as linear regression and discrete choice models, have been developed for insurance fraud detection. For example, Weisberg and Derrig employed a multiple linear regression model to select features of different types of fraud [11]. Artis et al. offered a logistic regression model to detect insurance fraud claims [7]. However, as mentioned by Viaene et al., the predetermined functional form and restrictive model assumptions of statistical methods limit their usefulness [12].

With the development of artificial intelligent theory, such as neural networks, data mining methods have been widely used for modeling and detection purposes and have obtained good detection performance. For example, in terms of using neural networks, He et al. used a multi-layer perceptron network to classify medical general practitioners into four classes [13]. To avoid over-fitting derived in the training stage and improve the generalization of the model, He et al. employed a weight decay term in the error function. Brockett et al. applied Knhonen's self-organizing feature map to classify automobile bodily injury claims by the degree of fraud suspicion [6]. The results showed that the proposed method outperformed both an insurance adjuster's fraud assessment and an insurance investigator's fraud assessment with respect to consistency and reliability. Moreover, as for decision trees, Gepp et al. compared the performance of decision tree and survival analysis in a US-based real-life automotive insurance fraud dataset [14]. Support vector machine (SVM) is also applied for insurance fraud detection. Sundarkumar et al. put forward a one-class SVM with undersampling to analyze an automobile insurance fraud dataset and a credit card customer

churn dataset [15].

Recently, some hybrid detection models have been developed that integrate intelligent techniques with many traditional statistical methods, such as various logit models, and some emerging intelligent methods, such as Bayesian networks, to improve prediction accuracy. Viaene et al. offered a Bayesian learning neural network for insurance fraud detection [12]. The neural network with one hidden layer of three neurons is developed by a practical Bayesian learning approach with automatic relevance determination. Moreover, Bermudez et al. developed a Bayesian dichotomous model with an asymmetric link accompanied by Markov Chain Monte Carlo and the Gibbs sampling method [16].

Moreover, some theories are applied to the detection models. Xu et al. applied the neural networks ensemble method with a random rough subspace to insurance fraud detection [17]. A set of reductions was generated to avoid data inconsistencies by using rough set reduction. They found that the proposed method outperforms single neural network classifiers. At the same time, Tao et al. applied fuzzy thought to a SVM model with the help of dual membership, which calculates the relativity between a claim and the two-sample mean vector. Moreover, Sundarkumar and Ravi conducted experiments on a hybrid method that combines a one-class SVM and k reverse nearest neighborhood to find fraudulent cases in insurance and banking [18].

To find an appropriate method for insurance fraud detection, some researchers have compared several algorithms. For instance, Hassan and Abraham give a brief introduction to decision trees, SVM and neural networks in their implementation for insurance fraud detection [19]. The result shows that the model using decision trees yielded the best

performance. Nian et al. put forward an unsupervised spectral ranking method for anomalies (SRA) by deriving a ranking vector to detect insurance fraud [20].

However, the aforementioned algorithms can only use the explicitly existing attributes of the data set for automobile insurance fraud detection. Nevertheless, there are more attributes hidden in the data set in other forms of the data, for example, hidden attributes which needs multiple and superimposed non-linear and/or linear transformation. Unfortunately, such hidden attributes can hardly be extracted and utilized by traditional data mining methods. Therefore, this paper introduces the deep learning method to take full advantage of the data, including explicitly existing attributes and hidden attributes. Deep learning can use several hidden layers to mine the hidden attributes contained in the data to better describe the features of automobile insurance fraud.

2.4 The contribution of our work

As aforementioned, data mining methods provide powerful tools for exploring the factors and methods of automobile insurance fraud detection. On one hand, however, previous practitioners and academic researchers have focused on the improvement of detection accuracy by coming up with powerful numeric and categorical features and algorithms, ignoring the validity of textual information in the insurance claims. On the other hand, the efficiency of manual analysis is far less than that of the data mining methods. Therefore, this paper proposes an LDA-based automobile insurance fraud detection method using deep learning, combining the experience of humans and the efficiency of artificial intelligence. In our method, LDA and deep learning technology are complementary. Topic model LDA is responsible for the analysis of text data by extracting the textual features hidden in the text

description of the claims. Deep learning is employed to seek high-quality attributes. Furthermore, the output of LDA can provide inspiration for the exploration of deep learning [21].

3. A text analytics framework

As aforementioned, we found that combining the experience of human experts and data mining methods can make full use of the advantages of both sides. First, the professional knowledge of experts makes it possible to avoid overfitting in data mining methods on the existing data. Second, data mining methods can help to avoid biases in the judgment of experts arising from their subjective attributes. Finally, data mining technology is more efficient than human experts in the detection of automobile insurance fraud. Therefore, we propose a detection method that utilizes the text description of the accident from human experts, using LDA and deep learning technology, to improve the effectiveness of automobile insurance fraud detection. In this paper, the use of LDA and deep learning technology can improve the performance of our detection model. Deep learning technology employs the distribution of topics generated by LDA. At the same time, with the help of deep learning, the topic model can achieve in-depth expansion. The use of these two methods can allow our model to better depict the fraudulent behavior of automobile insurance fraud, thus providing better support for the detection of fraud. The framework is outlined in Figure 1.

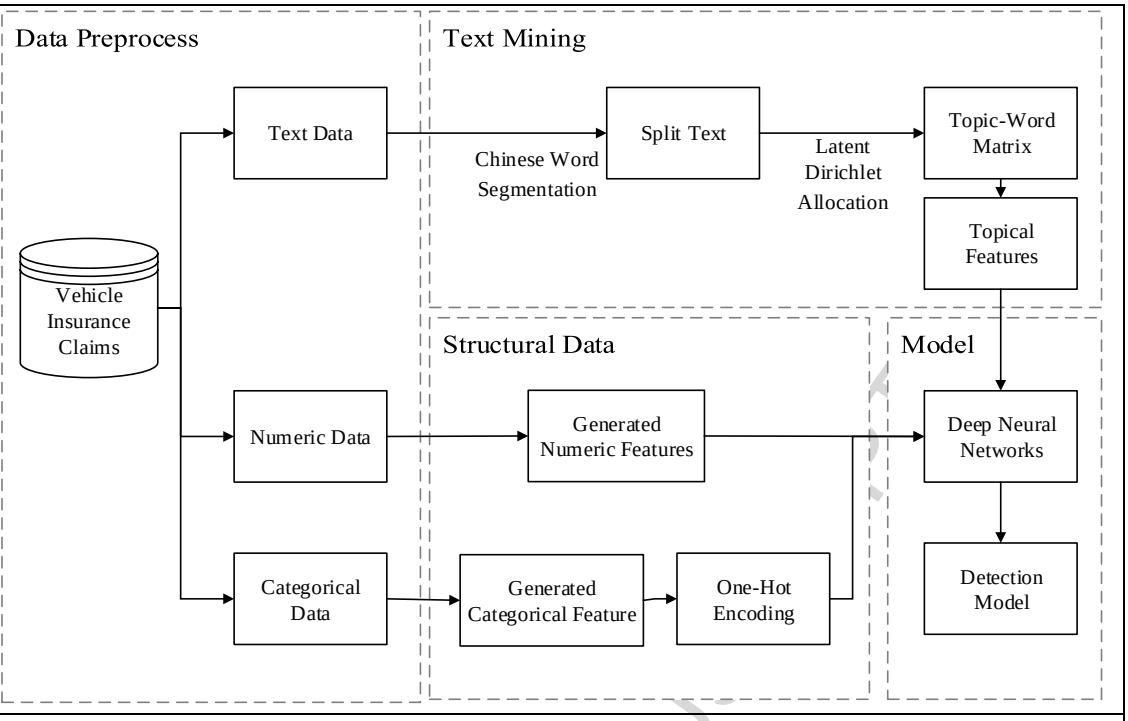


Figure 1. A Text Mining and DNN Framework for Insurance Fraud Detection

3.1. Data preprocess

Automobile insurance fraud data can be divided into structured data and unstructured data. Structured data include numeric data and categorical data, such as the number of past claims. Unstructured data refer to text data, such as the description of the accident. In the process of data preprocessing, the three types of data are cleaned according to their characteristics to facilitate the data processing.

3.2. Text mining

The experience of human experts exists in the text description of the accident, which cannot be understood by computers directly. Therefore, text mining is introduced to help extract high-value information that is buried in the text description.

Text mining is a relatively new technology in the field of insurance fraud detection. Therefore, some text mining methods accompanied by other technical methods will be

introduced to support our research.

Text mining can be described as a technical method that extracts useful information from unstructured data (text information) [22, 23]. Because of the complexity of text data, a variety of theories and technologies have been applied in the implementation of text mining [24]. Among the methods and theories, natural language processing (NLP) [25] and probability theory [24] are most widely used. While text mining is a branch of data mining, because of different research objects, text mining has its own difficulties: unstructured data problems. Due to the characteristics of unstructured text data, it is difficult for algorithms based on computers to understand the semantics of the text. Moreover, most traditional data mining technologies are not appropriate for text mining because of the uncertainty of text data in the form of diversification. Accordingly, some methods and changes need to be introduced to process text data. In this paper, we employ Chinese word segmentation and LDA to solve the problem of incompatibility between text data and data mining algorithms. Chinese word segmentation is used to preprocess the text data, and the topic model LDA is used to extract topics [26], which contain the experience of human experts in the processed text data.

3.2.1 Chinese word segmentation

The text descriptions of the accidents appear in the form of long sentences in different structures. Nevertheless, the key information of the description is distributed across several words in one sentence. Therefore, we should focus on analyzing a few keywords instead of the entire sentence. Chinese word segmentation refers to the segmentation of a sequence of Chinese characters into short strings of words [27]. It is a process of recombining a sequence of words according to certain criteria. Different from English and various other languages,

there is no space as a natural delimiter in Chinese [28]. Although there also exists the problem of division of phrases in English, Chinese word segmentation is more complicated than English at the word level. Different segmentation may lead to different meanings in Chinese. During the stage of Chinese word segmentation, the text descriptions in the form of Chinese sentences are split into sequences of Chinese words, which are convenient for the next processing steps according to the specific meaning of the text. At the same time, stop words can be abandoned during this stage to prevent impact on the analysis. After Chinese word segmentation and stop word removal, what remains are words that are more related to our research.

3.2.2 LDA

In natural language processing, the topic model Latent Dirichlet Allocation is an important generative model that can be used to identify hidden topic information in large-scale document collections or corpuses [29]. It uses the Bag of Words approach, which treats each document as a word-frequency vector, transforming the text data into numeric data that is compatible with data mining algorithms [30]. Each document can be viewed as a representation of a probability distribution of topics, and each topic can be viewed as a representation of a probability distribution of many words. Moreover, because of the non-correlation between the components of the random vector in the Dirichlet distribution, the candidate topics are independent of each other [31]. To achieve the above objectives, the LDA uses a joint distribution to compute the conditional distribution of the hidden variable under a given observable variable. The observable variable is a set of words, and the latent variable is the topics.

In our proposed method, the words arising from the text description can be viewed as documents for LDA to extract topics regarding automobile insurance behaviors. Because the text description is obtained from the automobile insurance claims, the topics extracted by LDA are about different aspects of the claims, such as responsibility, accident, etc. The documents are composed of these automobile insurance-related topics, whereas the topics are formed by words by distribution. It should be noted that documents and words are explicitly illustrated and that topics are implicitly expressed. Generally, topics are represented by several words that have a higher probability distribution in the corresponding topics. After the processing of LDA, the distribution of each claim to every topic is affiliated with the data set.

3.3. Structural data

A numeric feature is a type of feature that can be naturally handled by algorithms, while a categorical feature needs the transformation entitled one-hot encoding. In one-hot encoding, a categorical feature with M states corresponds to M bits, such as ‘001’ for one state of a categorical feature that has three states in total. In this manner, the model can address categorical features.

3.4. Deep learning model

All categorical, numeric and topical features are transmitted into the input layer of the deep neural network (DNN) to start the training process. Then, the input layer maps the features to the first hidden layer, and the process continues. Each hidden layer includes a number of nodes for processing the input data of the layer and transporting the result to the next layer. The activation function of each layer can add nonlinear mapping to the mapping process to guarantee that the abstraction ability of the DNN is more effective. At the same time, to avoid

the gradient vanish of errors in the process of back propagation, this paper employed ReLU instead of Sigmoid as the activation function [32]. After the iterative process of hyperparameter optimization, the DNN model outputs detection results and determines whether a claim is fraudulent.

4. Empirical Analysis

4.1 Data description

The data used in this paper are real-world data derived from an automobile insurance company, and the fraud label is confirmed by the insurance company's professional department. We ultimately obtain 37082 available items in the dataset, and each item represents an automobile insurance claim. Overall, there are 415 fraudulent claims and 36667 non-fraudulent claims. In the dataset, the ratio of fraudulent claims to legitimate claims is close to 88:1, which represents imbalanced data. Imbalanced data may greatly affect the performance of classification algorithms. Therefore, sampling methods are employed to solve the data imbalance problem. Because there is a large difference in the amount of data between the classes of claims, we both undersample legitimate claims (majority class) and oversample fraudulent claims (minority class) to balance the dataset [33][34]. We use SMOTE to oversample fraudulent claims and randomly undersample legitimate claims to get the same amount of data from the majority class to form a balanced dataset. Finally, the dataset contains 1660 legitimate claims and 1660 fraudulent claims.

Each claim consists of 16 attributes and 1 fraudulent label that indicate whether the claim is a fraudulent claim. The attributes can be divided into 10 categorical attributes, 5 numeric attributes and 1 text attribute. A description of the data is provided in Table 1, and summary

statistics of the numeric data are listed in Table 2.

Table 1. Attributes of the Data

No.	Attributes	Description
1	ReportTiming	The time of the claim
2	TimeToStart	Number of days after the effective date of the insurance
3	TimeToEnd	Number of days before the expiration date of the insurance
4	Color	The color of the insured car
5	Type	The type of the insured car
6	ReportReason	The reason in the claim
7	Brand	The brand of the insured car
8	LicensePlateNumber	License plate number of the insured car
9	CheckReason	The reason checked
10	Reporter	Who reported the accident
11	Driver	The driver of the car
12	LossType	The loss type of the insured car in the accident
13	AccidentType	The type of the accident
14	HistoryTimes	The number of past claims of the insured car
15	Region	The region the accident occurred
16	Description	The text description of the accident from human experts
17	Fraud	Fraudulent or not

Table 2. Summary Statistics of the Numeric Data

Attributes	Range	Mean	Standard Deviation
ReportTiming	[0, 23]	12.068	3.315
TimeToStart	[0, 643]	200.467	116.009
TimeToEnd	[0, 365]	163.923	116.149
HistoryTimes	[0, 18]	1.737	1.913

The types of insured cars include sedans, passenger vehicles, sport utility vehicles, sports cars, vans and others. Most of the insured cars are sedan, accounting for 59.9%. Passenger vehicles account for 10.1%, ranking second. The reported and checked reasons for accidents contain collision, fire, stolen, scratch, natural disasters and others. Among the claims, 86.1%

accidents are caused by collision, and scratching accounts for 9.3%. In terms of the regions in which the accidents occurred, urban regions represent 41.9%, rural regions represent 33.9%, and the remaining regions are others.

Categorical data and numeric data can be used to generate features of the corresponding type. Text data is used to extract the in-depth topic features that contain expert experience. The extracted topical features will be presented in detail later.

4.2 Topical feature extraction

The text description includes a professional description of the scene of the accident by a human expert. In this paper, LDA is applied to extract the topic hidden in the text description. LDA can be used to obtain the distribution of the keywords for each topic and the distribution of the topics in each description.

In this paper, we employ perplexity to determine the appropriate number of topics. Perplexity measures the probability distribution's ability of prediction. An appropriate probability distribution has a relatively low perplexity. The formula is as follows.

$$\text{perplexity} = \exp\left(-\frac{\sum_1^D \log \mathrm{p}(W)}{N}\right) \quad (1)$$

Additionally, because we use 10-fold cross-validation to improve the model's ability to resist overfitting, LDA is carried out 10 times with cross-validation. Each time, LDA is built on text from 9-fold training data, and the text from the remaining test data is processed by the model built. In the experiment, we found that the results of the 10 LDA models are very similar. The perplexities of 10 LDA models reach the lowest points under the same condition that the number of topics is 5. The reason is that the description in data is completely about the scene of vehicle accidents, which causes the word distribution and topic distribution in

each fold to be very close. Therefore, the appropriate number of topics is 5 in each 10-fold cross-validation experiment.

Some representative words of topics are shown in Table 3 according to our training of LDA. It shows some of the keywords whose distribution probabilities are more than 0.05 in the LDA model.

Table 3. Key Words of Five Topics

Topic	Key Words
1	Scene, Driver, Third-party, Liability
2	Report, Treatment, Compensation
3	Policeman, Advise, Witness
4	Drive, Right, Collision
5	Front, Back, Glass, No-injury, Impaired

First, topic 1 contains a bag of words that describes liabilities of accidents, including the driver of the car and the third party. Some insurance swindlers fabricate the liability after the accident to defraud the premium. Second, topic 2 describes how to address the accident. Different people employ different methods in the face of the accident. Some people will report the accident to the police for help, and others will accept compensation from a legitimate private party with pleasure. Topic 3 is the description of the scene of the accident, including whether there are witnesses and whether there was police intervention. Additionally, driving behavior is mainly distributed in the words derived from topic 4. Traditionally, experts describe suspicious driving behavior based on their own experience. Finally, topic 5 is about the damage due to the accident, including damage to the automobile and personal injury.

The distribution of the 5 topics in the text description will be affiliated with the corresponding claim. These topical features will help the abstraction process of the DNN extract the experience of human experts.

4.3 Experimental analysis

4.3.1 Hyperparameter Optimization

To make algorithms achieve better results, we need to set hyperparameters. The initial values of hyperparameters use the optimal values in theory or practice and are adjusted continuously by the accuracy of the experimental results.

First, the number of trees, maximum depth of trees and number of attributes used by each tree need to be tuned in random forests (RF). We employ grid search to perform hyperparameter optimization. According to the central limit theorem, the more trees there are in RF, the better the result and the longer the training time will be. However, when the result is optimized to some extent, the effect of the increase in the number of trees will disappear. The number of trees is sampled from the range of [10, 200] by a step of 5. At the same time, an appropriate maximum depth of a tree can effectively avoid over-fitting and under-fitting problems. We evaluate the maximum depth between 2 and 10. Moreover, the number of attributes used by each tree is generally set as the root of the total number of attributes, and we evaluate the parameter from 2 to 17. Experimental results prove that RF with 100 trees whose maximum depth is set as 3 and 4 attributes used by each tree achieve the best performance.

Second, a Gaussian Kernel is employed as the kernel of SVM. Grid search is implemented to tune the gamma and penalty factor C. We first set the range of C and gamma as $[2^{-8}, 2^{8}]$ and then narrow the range of C to [0.1, 5] by a step of 0.1 and the range of gamma to [0.02, 1] by a step of 0.02. The experimental results show that the configuration of C=1, gamma=0.1 can make SVM perform best on the dataset.

Third, to determine the architecture of the DNN, we start with 3 hidden layers to tune the number of hidden layers, adding one hidden layer each time. Grid search is applied to help build the optimal architecture [35]. While adding hidden layers, we change the number of nodes in each layer to make the model perform best. As the number of hidden layers increases, the number of nodes that need to be optimized increases rapidly, so we tune the number of nodes in each hidden layer in the range of [1,20] by a step of 5 and then perform small range optimization, for example, in a range of [5, 10] by a step of 1. Larger numbers of hidden layers may lead to better results but at the expense of longer computation time [36]. When the number of hidden layers is greater than 7, the performance of the model does not increase as the number of hidden layers increases, but the amount of computation continues increasing. Finally, the DNN uses 7 hidden layers, and the numbers of nodes in each hidden layer are 8, 8, 10, 7, 8, 6 and 4. Meanwhile, to avoid the gradient vanish, DNN uses ReLU as the activation function. Dropout is employed to avoid overfitting. The dropout probability is 0.2. Moreover, the learning rate starts at 0.5 and decreases as the epoch increases.

It is worth noting that we have conducted a comparative experiment on the impact of LDA. The optimal hyperparameters of RF and SVM are not affected by LDA. However, the addition of LDA results changes the architecture of DNN. For the experiment containing LDA analysis, the optimal architecture of DNN (as far as we know) is as described above. For the experiment without LDA analysis, the optimal structure of DNN (as far as we know) is 7 hidden layers, and the numbers of nodes in each hidden layer are 6, 7, 9, 4, 5, 4, and 4.

In addition, 10-fold cross-validation is used to ensure the validity of the fraud detection models. The data with textual description are randomly divided into 10 folds to reduce the

likelihood of model overfitting.

4.3.2 Experimental results

The output of LDA applied on the text description of the accidents is affiliated with the automobile insurance fraud data. Then, we use the DNN algorithm to abstract the data and build the detection model. The output of LDA can guide the DNN in the process of attribute abstraction. Therefore, the combination of LDA and DNN allows our model to perform better when characterizing fraudulent behaviors. For comparison, we also evaluate some well-known classifiers, such as RF and SVM, on the same data set. These two algorithms are representative algorithms in traditional data mining algorithms. To compare the performance of traditional machine learning methods, Fernández-Delgado et al. evaluated 179 classifiers derived from 17 families using 121 data sets from the UCI database [37]. SVM and RF outperform the other classifiers with accuracy rates of 94.1% and 92.3%. To some extent, RF and SVM represent the peak of the traditional machine learning field. More specifically, RF and SVM are the leaders in fraud detection among traditional data mining algorithms. Fortuny et al. showed that SVM was a more appropriate algorithm than naïve Bayes in fraud detection [38]. Meanwhile, Bhattacharyya et al. found that RF and SVM yielded a better result than logistic regression [39]. Moreover, Whitrow et al. compared the performance of 7 algorithms, including RF, logistic regression, SVM, naïve Bayes, quadratic discriminant analysis, CART and k-nearest neighbors, in fraud detection. The experimental results showed that RF and SVM performed better than other algorithms in fraud detection [40]. Therefore, RF and SVM are selected in this paper because of their outstanding performance in fraud detection.

To ensure the validity of the three algorithms, we used 10-fold cross-validation in all

experiments.

The experimental results are shown in Table 4. The TP Rate of DNN is 8.7% higher than that of RF and 22.8% higher than that of SVM, which means that DNN is stronger in ability to identify fraudulent claims in actual fraud claims. The possible reason is that, compared to RF and SVM, DNN can better understand and abstract the expert experience extracted by LDA. From the perspective of misjudgment, which is measured by FP Rate, DNN performs better than do RF and SVM. Meanwhile, in terms of accuracy, DNN yields a better performance, with a score of 0.914, than do RF and SVM. At the same time, the precision of DNN is higher than that of SVM and RF. Moreover, the F1 of DNN is 15.1% higher than that of SVM and 9.9% higher than that of RF. The performance of DNN indicates that DNN outperforms the other two algorithms in terms of the data we used.

Table 4. The Performances of SVM, RF and DNN with LDA

	TP Rate	FP Rate	Accuracy	Precision	F1
SVM	0.682	0.108	0.787	0.863	0.762
RF	0.823	0.198	0.812	0.806	0.814
DNN	0.910	0.082	0.914	0.917	0.913

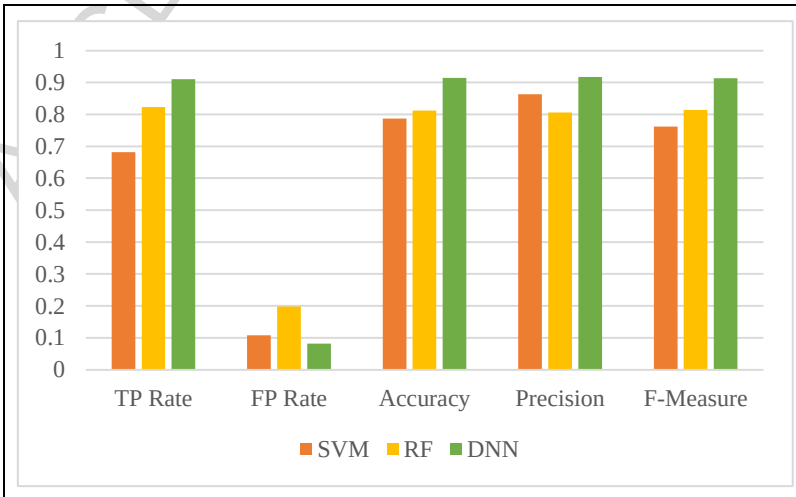


Figure 2. The Performances of Three Algorithms with LDA

For further explanation, a t-test is conducted on 10 results of 10-fold cross-validation, and

the result shows that the accuracy of DNN (mean=0.914) is significantly higher than that of RF (mean=0.812) (t=33.4004, p<0.0001) and SVM (mean=0.787) (t=35.3993, p<0.0001). Moreover, the difference between the precisions of DNN (mean=0.917) and RF (mean=0.806) (t=33.1232, p<0.0001) and between those of DNN and SVM (mean=0.863) (t=14.9733, p<0.0001), is considered to be extremely statistically significant. Therefore, we believe that the performance of DNN on data sets with the help of LDA outperformed RF and SVM.

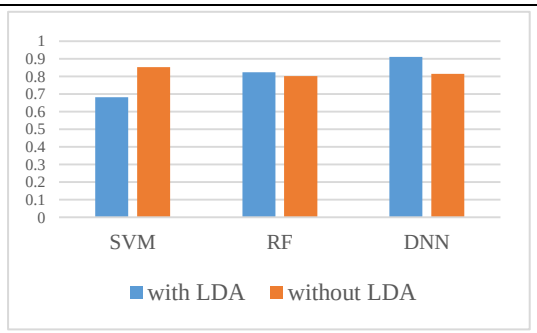
4.3.3 Comparative experiments

To show the contribution of LDA in our detection model, we remove the LDA output in the data set. The three algorithms, DNN, RF and SVM, are applied for modeling. The results are shown in Table 5. First, in terms of TP Rate, LDA analysis helps DNN and RF perform better. Meanwhile, the TP Rate of DNN is 3.9% lower than that of SVM without LDA analysis, but LDA helps DNN surpasses SVM by 22.8% in TP Rate. Second, the FP Rates of all three algorithms decrease with the help of LDA. Moreover, the accuracy and precision of RF and DNN also increase after LDA analysis is added to our proposed methods.

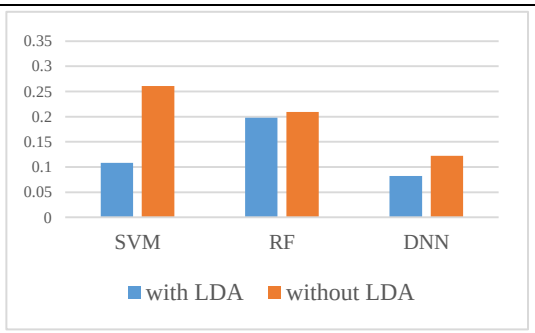
Through the analysis of comparative experimental results, we found that the improvement of LDA on models is obvious, especially for DNN. However, the addition of LDA analysis reduces the TP Rate, accuracy and F1 of SVM.

Table 5. The Performances of SVM, RF and DNN without LDA

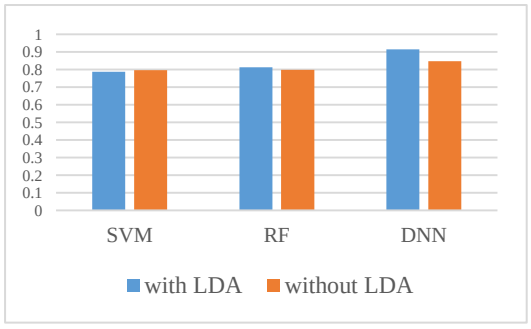
	TP Rate	FP Rate	Accuracy	Precision	F1
SVM	0.853	0.261	0.796	0.765	0.807
RF	0.802	0.209	0.797	0.793	0.798
DNN	0.814	0.122	0.846	0.869	0.841



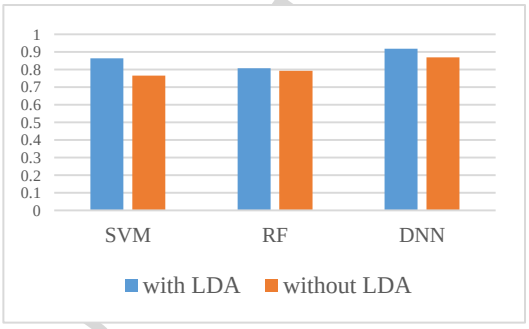
a. TP Rate



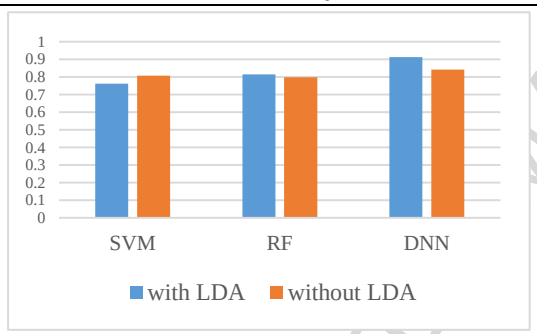
b. FP Rate



c. Accuracy



d. Precision



e. F1

Figure 3. The Comparative Evaluation criterion: with LDA versus without LDA

The t-test indicates that LDA significantly improves the performance of DNN. The accuracy of DNN with LDA (mean=0.914) is significantly higher than the accuracy of DNN without LDA (mean=0.846) ($t=13.7877$, $p<0.0001$). Meanwhile, there is a significant difference between the precision of DNN with LDA (mean=0.917) and the precision of DNN without LDA (mean=0.869) ($t= 11.9275$, $p<0.0001$).

The comparative experiment illustrates that the method proposed in this paper can improve the accuracy of automobile insurance fraud detection. First, this paper tests RF and SVM on a

data set that has no LDA topical features. To explain the importance of text mining in fraud detection, LDA was employed to extract the topics to allow the algorithm to analyze text data, and experimental results show that text mining is critical to improving the accuracy of fraud detection. Additionally, the accuracy of DNN with LDA increased by 6.8%, the precision increased by 4.8%, the Recall/TP Rate increased by 9.6%, and the F1 value increased by 7.2%. In addition, to further expand the advantages of LDA, this paper tests DNN on the data extended with topical features, and RF and SVM are employed as contrasts. DNN makes a great contribution to the good performance of the model. The F1 value of DNN is 9.9% higher than that of RF, which is dominant in two traditional classification algorithms. The multi-layer structure enables DNN to better extract features from the data, while LDA provides a better raw material for this type of abstraction.

For the processing of text information, LDA and the deep learning algorithm are complementary. LDA can help and guide the abstraction level and aspects of analysis in deep learning, while deep learning can analyze and abstract the topics arising via LDA. Because the topics extracted by LDA are latent [41], we need to dig deeply into the topics for more effective information. The advantage of deep learning in feature abstraction and deep mining makes it the best choice in such a situation. Comparative tests are conducted to illustrate the fitness and effectiveness of LDA and the deep learning algorithm. On the one hand, the performance of the deep learning algorithm with LDA topical features is better than that without LDA topical features; on the other hand, for automobile insurance claim data with LDA topical features, deep learning outperforms RF and SVM. Additionally, the combination of LDA and DNN provides more features, which rectifies the incorrectly classified claims by

traditional classification methods. The reason is that the topic extraction process of LDA and the abstraction process of DNN can provide more powerful topical features, which cannot be provided by traditional methods. These extra-high-quality topics are critical to the detection of insurance fraud. DNN can deeply mine these features and details, and therefore the proposed method can yield better performance.

4.4 Further discussion

For further understanding and explanation of our proposed LDA-based deep learning model, we compare and analyze the performance of the three algorithms on the claims. To illustrate the effectiveness of DNN with LDA, the analysis focuses on the claims D_x that RF/SVM misjudges to be legal but DNN correctly detects. Through analysis, we find that the descriptions in these claims are less distributed on Topic 1 and Topic 3 but more distributed on Topic 4 and Topic 5 after being processed by LDA. In the claims D_x, the average probabilities of Topic 1 and Topic 3 are below the average level of the training set, while average probabilities of Topic 4 and Topic 5 are above the average level of the training set. To illustrate the problem more effectively, relevant attributes of a real piece of claim d_x are extracted as below.

Table 6. Relevant Attributes of Claim d_x

Attributes	Value	Average level in D_x	Average level in the training set
Topic1	0.176	0.161	0.201
Topic2	0.201	0.195	0.201
Topic3	0.116	0.151	0.205
Topic4	0.324	0.217	0.180
Topic5	0.234	0.276	0.212

Claim d_x is typical data in D_x. RF and SVM failed to detect data d_x, while DNN with LDA succeeded in catching it. The probability that claim d_x belongs to Topic 1 and Topic 3 is lower

than the average level in the training set, while the probability that claim d_x belongs to Topic 4 and Topic 5 is higher than the average level in the training set. Therefore, DNN with LDA can better detect the claims in which the description is more prone to describe driving behavior and damage of accidents and avoid describing the liabilities of the accidents and scenes of accidents.

Because of the multi-layer nonlinear structure, DNN can deeply understand and use data. Traditional data mining algorithms, such as RF and SVM, can perform linear and nonlinear transformation of attributes at a shallow level, while DNN can perform such transformations at multiple deep levels. For example, Topic 1 contains descriptions of liability and causes of accidents, corresponding to the numerical attributes ReportReason and CheckReason. When these attributes enter the nodes of the hidden layers of DNN, the nonlinear transformation and activation functions compare and process the attributes and pass the results to the next hidden layer. Deeper hidden layers take the results for further transformation, utilizing the analysis of the shallow layer by the model. Moreover, benefiting from ReLU to reduce the risk of gradient vanish, DNN can contain more hidden layers, which results in deeper transformation and mapping. Therefore, the possibility of generating effective new attributes is greatly improved, which can hardly be done by shallow-layer data mining algorithms. As a result, DNN can better abstract and utilize data.

5. Conclusions and Future Work

This paper makes several contributions to the detection of automobile insurance fraud. First, this paper introduces text mining methods to resolve the text description where the experience of human experts is hidden. The experimental results confirm that the text mining method is important for the analysis of fraudulent behaviors. Second, this paper proposes an LDA- and deep learning-based automobile insurance fraud detection model. The experimental results show that our proposed method is effective. The complement of LDA and the deep learning

method makes it possible for the model to characterize the behavior of automobile insurance fraud. Additionally, the topic extraction process of LDA and the abstraction process of DNN can provide more effective topical features, which cannot be supplied by traditional methods.

The model proposed in this paper can be improved in several aspects. First, this method can be verified on more data sets. Different data can be used to extract different topics so that the model can be tuned to accommodate more situations. Second, in the analysis process of the text description derived from human experts, different types of subjective ideas are inevitably involved, which causes adverse effects on the results. Therefore, future work will focus on the elimination of individual characteristics hidden in the text descriptions from human experts.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 71301163, 71771212), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075, 15YJA630068), Hebei Social Science Fund (HB13GL021), Fundamental Research Funds for the Central Universities, and Research Funds of Renmin University of China (No. 15XNLQ08).

References

[1] A. Abdallah, M.A. Maarof, A. Zainal, Fraud detection system: a survey, Journal of Network and Computer Applications 68 (2016) 90-113.

[2] C.A. Knapp, M.C. Knapp, The effects of experience and explicit fraud risk assessment in detecting fraud with analytical procedures, Accounting Organizations & Society 26 (1) (2001) 25-37.

[3] S. Viaene, R.A. Derrig, B. Baesens, G. Dedene, A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, Journal of Risk & Insurance 69 (3) (2002) 373-421.

[4] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, Decision Support Systems 50 (3) (2011) 559-569.

[5] H.I. Weisberg, R.A. Derrig, Fraud and automobile insurance: a report on bodily injury liability claims in Massachusetts, Journal of Insurance Regulation 9 (4) (1991) 497-541.

[6] P.L. Brockett, X. Xia, R.A. Derrig, Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud, Journal of Risk & Insurance 65 (2) (1998) 245-274.

[7] M. Artís, M. Ayuso, M. Guillén, Detection of automobile insurance fraud with discrete choice models and misclassified claims, Journal of Risk & Insurance 69 (3) (2002) 325-340.

[8] J.H. Wilson, An analytical approach to detecting insurance fraud using logistic regression, Journal of Finance & Accountancy 85 (150) (2009) 1-15.

[9] L. Šubelj, Š. Furlan, M. Bajec, An expert system for detecting automobile insurance fraud using social network analysis, Expert Systems with Applications 38 (1) (2011) 1039-1052.

[10] F. Karamizadeh, S.A. Zolfagharifar, Using the clustering algorithms and rule-based of data mining to identify affecting factors in the profit and loss of third party insurance insurance company auto, Indian Journal of Science & Technology 9 (7) (2016) 1-9.

[11] H.I. Weisberg, R.A. Derrig, Quantitative methods for detecting fraudulent automobile

bodily injury claims, Risques 35 (1998) 75-101.

[12] S. Viaene, G. Dedene, R.A. Derrig, Auto claim fraud detection using Bayesian learning neural networks, Expert Systems with Applications 29 (3) (2005) 653-666.

[13] H. He, J. Wang, W. Graco, S. Hawkins, Application of neural networks to detection of medical fraud, Expert Systems with Applications 13 (4) (1997) 329-336.

[14] A. Gepp, J.H. Wilson, K. Kumar, S. Bhattacharya, A comparative analysis of decision trees vis-à-vis other computational data mining techniques in automotive insurance fraud detection, Journal of Data Science 10 (3) (2012) 537-561.

[15] G.G. Sundarkumar, V. Ravi, V. Siddeshwar, One-class support vector machine based undersampling: application to churn prediction and insurance fraud detection, Proceedings of the 31st IEEE International Conference on Computational Intelligence and Computing Research, 2015. pp. 1-7.

[16] L. Bermúdez, J.M. Pérez, M. Ayuso, E. Gómez, F.J. Vázquez, A Bayesian dichotomous model with asymmetric link for fraud in insurance, Insurance Mathematics & Economics 42 (2) (2008) 779-786.

[17] W. Xu, S. Wang, D. Zhang, B. Yang, Random rough subspace based neural network ensemble for insurance fraud detection, Proceedings of the 31st IEEE Fourth International Joint Conference on Computational Sciences and Optimization, 2011. pp. 1276-1280.

[18] G.G. Sundarkumar, V. Ravi, A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, Engineering Applications of Artificial Intelligence 37 (2015) 368-377.

[19] A.K.I. Hassan, A. Abraham, Modeling insurance fraud detection using imbalanced data

classification, Advances in Nature and Biologically Inspired Computing (2016) 117-127.

[20] K. Nian, H. Zhang, A. Tayal, T. Coleman, Y. Li, Auto insurance fraud detection using unsupervised spectral ranking for anomaly, The Journal of Finance and Data Science 2 (1) (2016) 58-75.

[21] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research 11 (3) (2010) 625-660.

[22] R. Feldman, I. Dagan, Knowledge discovery in textual databases (KDT), Proceedings of the First International Conference on Knowledge Discovery from Databases, 1995. pp. 112-117.

[23] A. H. Tan, Text mining: The state of the art and the challenges, Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, 1999. pp. 65-70.

[24] A. Hotho, A. Nürnberger, G. Paaß, A brief survey of text mining, LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 20 (2005) 19-62.

[25] B. Furlan, V. Batanović, B. Nikolić, Semantic similarity of short texts in languages with a deficient natural language processing support, Decision Support Systems 55 (3) (2013) 710-719.

[26] R.Y. Lau, C. Li, S.S. Liao, Social analytics: learning fuzzy product ontologies for aspect-oriented sentiment analysis, Decision Support Systems 65 (2014) 80-94.

[27] F. Wu, Y. Huang, Y. Song, S. Liu, Towards building a high-quality microblog-specific Chinese sentiment lexicon, Decision Support Systems 87 (2016) 39-49.

[28] Y. Wang, W. Xu, H. Jiang, Using text mining and clustering to group research proposals for research project selection, Proceedings of the 48th Hawaii International Conference on System Sciences, 2015. pp. 1256-1263.

[29] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993-1022.

[30] H. Yuan, R.Y. Lau, W. Xu, The determinants of crowdfunding success: a semantic text analytics approach, Decision Support Systems 91 (2016) 67-76.

[31] D. Blei, J. Lafferty, Correlated topic models, Advances in Neural Information Processing Systems 18 (2006) 147-154.

[32] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2010. pp. 315-323.

[33] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations Newsletter 6 (1) (2004) 20-29.

[34] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263-1284.

[35] T. Domhan, J.T. Springenberg, F. Hutter, Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves, Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015. pp. 3460-3468.

[36] J. Evermann, J.-R. Rehse, P. Fettke, Predicting process behaviour using deep learning, Decision Support Systems 100 (2017) 129-140.

[37] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15 (1) (2014) 3133-3181.

[38] E.J.d. Fortuny, M. Stankova, J. Moeyersoms, B. Minnaert, F. Provost, D. Martens, Corporate residence fraud detection, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. pp. 1650-1659.

[39] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study, Decision Support Systems 50 (3) (2011) 602-613.

[40] C. Whitrow, D.J. Hand, P. Juszczak, D. Weston, N.M. Adams, Transaction aggregation as a strategy for credit card fraud detection, Data Mining and Knowledge Discovery 18 (1) (2009) 30-55.

[41] T. Wang, Y. Cai, H.-f. Leung, R.Y. Lau, Q. Li, H. Min, Product aspect extraction supervised with online domain knowledge, Knowledge-Based Systems 71 (2014) 86-100.

Biographical Note

Mr. Wang is a Ph.D candidate at School of Information, Renmin University of China. He got his master and bachelor degree in Information Systems at School of Information, Renmin University of China. Her research interests include big data analytics, business intelligence and decision support systems. He has published several research papers in international journals and conferences, such as HICSS.

Dr. Xu is an associate professor at School of Information, Renmin University of China. He is a research fellow at Department of Information Systems, City University of Hong Kong. He got his bachelor and master degree in Mathematics at Xi'an Jiaotong University and doctor degree in Management Science at Chinese Academy of Sciences. His research interests include big data analytics, business intelligence and decision support systems. He has published over 90 research papers in international journals and conferences, such as Annals of Operations Research, Decision Support Systems, European Journal of Operational Research, IEEE Trans. Systems, Man and Cybernetics, International Journal of Production Economics, and Production and Operations Management.

Highlights

1 A novel deep learning methodology is proposed for automobile insurance fraud detection;

2 LDA-based text analytics is developed to extract the features in the text description of the accidents;

3 Text features and traditional numeric features are explored for the detection of fraudulent claims;

4 Our research contributes to advance the computational method for analyzing and detecting insurance fraud.