Predict Ireland's workplace incidents type using Artificial Neural Networks*

Usukhbayar Tsendgombo
CCT college Dublin
Student number: 2022418
2022418@student.cct.ie

Abstract

Workplace incidents refer to any unexpected events that occur in a workplace, which can cause injury, illness, or property damage. In recent years, deep learning has made good progress and has been applied to various applications, such as predictive modeling, image recognition, natural language processing, etc. In order to extract the workplace incidents in Ireland, the paper will discuss the network training of deep artificial neural network on incident cases and compare the ANN results with the other traditional research results. The dataset used for the training is Workplace incidents in Ireland from 2017-2021, and consists of incident type, NACE sector, gender, age group, Employment status and trigger. The result shows that the accuracy 98.94% with 100 epochs of training.

Keywords

Workplace, incidents, deep learning, deep neural networks, artificial neural networks

I. INTRODUCTION

With the amount of data growing at an exponential rate, it is necessary to develop tools that are able to harness that data and extract value from it. The growing need to manage increasingly massive data collections has led to significant interest in developing such big data tools. This steadily growing demand has led to the development of several big data analytics tools in the industry. With the spreading prevalence of Big Data, many advances have recently been made in this field. Frameworks such as Apache Hadoop and Apache Spark have gained a lot of traction over the past decades and have become massively popular, especially in industries. It is becoming increasingly evident that effective big data analysis is key to solving artificial intelligence problems [5].

A vast amount of research has been conducted to identify human and organizational factors that contribute to the occurrence of workplace incidents. Considering the identified factors, the question is how much the workplace incident probability will decrease in the absence of one or more recognized contributory factor. And otherwise, workplace incidents can occur in any type of workplace, including offices, factories, construction sites, and retail stores. They can be caused by a variety of factors, such as unsafe work practices, faulty equipment, inadequate training, or lack of supervision. Also, employers have a responsibility to provide a safe work environment for their employees and to take steps to prevent workplace incidents from occurring. This includes conducting risk assessments, providing adequate training and equipment, implementing safety policies and procedures, and encouraging a culture of safety in the workplace [1].

Fatal incidents: In Ireland, were 53 work-related fatal incidents in 2020. This is an increase from the 47 fatal incidents in 2019, and the highest number of fatal incidents since 56 were recorded in 2015. Of 53 work-related fatal incidents, 23 (43.4%) occurred in Agriculture, Forestry and Fishing alone, while 15 (28.3%) occurred in Construction. For Agriculture, Forestry and Fishing, this follows a similar pattern to recent years with half of all fatal incidents in the five-year period since 2016 occurring in this sector. Construction fatalities, however, have shown considerable variation in recent years, dropping to five in 2017 and showing increases since then. More than half of all fatal incidents in 2020 occurred to self-employed people (28), with 12 fatalities occurring to employees and 13 to non-workers [1].

Non-fatal incidents: Fewer non-fatal incidents were reported in every economic sector except Agriculture, Forestry and Fishing. The greatest decline was in Accommodation and Food Service Activities, which reported 107 nonfatal incidents in 2020, compared with 254 in 2019, a decrease of 57.9%. Large declines in reported incidents were also noted in Arts, Entertainment and Recreation (42.9%), Education (42.1%) and Transportation and Storage (41.0%). Manual handling and falls were the most common triggers in 2020, as they have been in all years since 2016. This shows that these two triggers have consistently been amongst the most prevalent causes of non-fatal incidents in recent years. The part of the body affected in the

greatest number of non-fatal incidents to workers was the back. Back injuries were reported in 20.4% of worker incidents, but only in 4.7% of non-worker incidents. This is because more workers were involved in manual handling incidents caused by lifting or moving heavy objects than non-workers. For non-workers, the part of the body injured most frequently was the head (21.3%). These head injuries most often followed slips or falls. The most common working environment for non-fatal work-related incidents was industrial or maintenance areas for workers (25.5%) and shop or service areas for non-workers (54.9%).[1]

The purpose of the research is to build the workplace incident type prediction model of an artificial neural network by using Apache Spark dataframe. In section II reviews existing approaches on related research of workplace incidents, Apache Spark and Deep learning Artificial neural networks. Section III presents the approach of Apache Spark and Artificial neural network. Experiments include data identification, data preprocessing, model building and setting in Section IV. The conclusion of the research is presented in Section V.

II. LITERATURE REVIEW

The literature in this area includes research that has been done in the related fields of workplace incidents, Apache Spark and Deep learning Artificial Neural Network. I study this literature with the intention to analyse past research in these areas and consider their methods of research, main results and highlights of conclusions. The literature survey pertaining to this paper has been presented in three parts: (i) Research Related to workplace incidents, (ii) Research Related to Apache Spark, (iii) Research to Related Deep learning Artificial neural network.

A. Research Related to workplace incidents

The authors of [7] presents study modelled incidents and accidents in two companies in the mining and construction sectors in order to identify the most important causes of accidents and develop predictive models. The research into workplace risk is mainly conducted using conventional descriptive statistics, which, however, fail to properly identify cause-effect relationships and are unable to construct models that could predict accidents. The results were compared with a classical statistical technique (logistic regression), revealing the superiority of decision rules, classification trees and Bayesian networks in predicting and identifying the factors underlying accidents and incidents.

The research done in [8] work situations leading to injury or patient and worker injury are emotionally distressing for health care workers. Team interplay may facilitate safe and dynamic practices and help HCWs overcome negative emotions. Organizational support is imperative for individual closure. For safety in health care, employers need to develop strategies for active management of risks, avoiding injuries and providing support after an injury. The method of research is a qualitative design using the critical incident technique. Semi-structured individual interviews were held with 34 HCWs from three regions in Sweden and data were analysed using inductive category development.

An industrial workplace Alerting and Monitoring Platform to prevent workplace injury and accidents presented in [9]. They proposed a platform to improve workplace safety. Their method is the first to generalize the detection of unsafe activities to scenarios that contain multiple people and objects over a period of time. Additionally, to combine PPE detection and unsafe activity detection into a single platform that can be implemented in CCTV cameras to monitor industrial workplaces in real time.

B. Research Related to Apache Spark

The authors of [5] presented a novel framework for the analysis of big data. The proposed combined two widely popular tools, named Apache Spark and Deep learning, under the umbrella of a single structure. By using these highly acclaimed individual tools in coherence with each other, they were able to obtain a model that is capable of conducting large scale big data analysis tasks within short periods of time, with lesser computational complexity and with significantly higher accuracy. This model is an outer structure that allowed us to model all machine learning tasks, such as classification and recommendation, with ease. Eventually, the initial stage, when run through a multi-class classifier

and then combined with the multi-class deep learning stage yielded performances lower than its binary initial stage counterpart. They aim to overcome this challenge in future research.

C. Research to Related Deep learning Artificial neural network

The authors of [10] presents a survey on the application of ANN approaches within the context of big data analytics. In their review, state-of-the-art issues regarding the application of ANN in big data analytics is provided. The paper has discussed the challenges involving ANN in terms of handling big data, and future research opportunities are unveiled. The progress within big data analytics using ANN is described. Research on the applications of ANN in big data analytics is at an early stage and is expected to grow rapidly in the near future. In addition, the paper can be used as a starting point for novice researchers interested in big data analytics using ANN.

Different researchers have not provided a predict model of workplace incidents type by using Artificial Neural Networks. However, I present a detailed study of the prediction model of workplace incident type based on a reliable dataset and source.

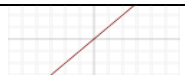
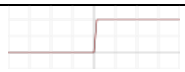
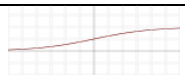
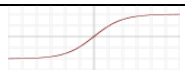
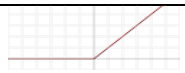
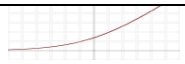
III. THE APPROACH

3.1. Deep artificial neural network

The different types of neural networks in deep learning, such as convolutional neural networks (CNN), recurrent neural networks (RNN), artificial neural networks (ANN), etc. are changing the way we interact with the world. These different types of neural networks are at the core of the deep learning revolution, powering applications like unmanned aerial vehicles, self-driving cars, speech recognition, etc. Artificial neural networks (ANN) is an intelligent system inspired by the human nervous system. ANNs are very good with fitting problems, with enough neurons ANNs can fit any data with arbitrary accuracy.

Artificial Neural Network (ANN) is capable of learning any nonlinear function. Hence, these networks are popularly known as Universal Function Approximators. ANNs have the capacity to learn weights that map any input to the output. One of the main reasons behind universal approximation is the activation function. Activation functions introduce nonlinear properties to the network. This helps the network learn any complex relationship between input and output. Table 1 shows the compares the properties of several activation functions.

Table 1. Table of activation functions

Name	Plot	Range
identity		(-∞,∞)
Binary step		{0,1}
Logistic, sigmoid, soft step		(0,1)
Hyperbolic tangent tanh		(-1,1)
Rectified linear unit ReLu		[0, ∞)
Softplus		(0, ∞)

Neural Network links a set of input nodes xi = ($x1$, $x2$, ... , xn) existing in the input layer with a set of one or more output nodes yj = ($y1$, $y2$, ... , ym) existing in the output layer through an intermediate hidden layer.

Nodes in each layer are activated once they reach the layer threshold value θi. This matching is realized by finding an unknown function h. [2]

$$yj = h(x1,x2,...,xn) \quad (1)$$

$i \in [1,n]$ and $j \in [1,m]$

The research intends to use Keras which is an open source of a high-level Neural Network library, written in Python and runs on Theano and TensorFlow. First, data should be collected and preprocessed into input dataset and desired output dataset. Second, should build and design his network by choosing the type of learning. As well as by fixing the network parameters, for example net input, activation function, number op Epoch, Batch, number of neurons in each layer, etc.

3.1. Apache Spark

Spark is a cluster computing engine of Apache and is purposely designed for fast computing process in the world of Big Data. Spark is Hadoop based efficient computing engine that offers several computing features like interactive queries, stream processing, and many others. In memory cluster, computing offered by Spark enhances the processing speed of the applications [6]. Apache Spark is freely available parallel data processing framework which gains increase in attention in the subject of big data analytics and artificial intelligence. Moreover it feeds the fundamental data in memory while doing computation [4]. Spark is one of the quick and adaptable cluster computing platforms which is made for wide range data processing. It can work using smallest and lowest unit which is known as micro batch process. Spark platform does not use costly disk access as its computation is done on memory which may lead to increased performance in data processing.

Key features are batch/streaming data which is unify the processing of data in batches and real-time streaming, using multiple language: Python, SQL, Scala, Java or R. Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting and runs faster than most data warehouse. Perform Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to down sampling.

Apache Spark is a general-purpose, distributed cluster computing, data processing framework that, like MapReduce in Apache Hadoop, offers powerful abstractions for the processing of a large dataset. The Spark core is complemented by a set of powerful and higher-level libraries: SparkSQL, Spark Streaming, MLlib, GraphX, Packages.

Apache data frames are the collection of distributed data. In data frame, the data is organized in columns and optimized tables. Spark data frames can be constructed from various data sources that include data files, external databases, existing RDDs and Spark data frames. The research used Python and PySpark libraries in Spark DataFrame. On Anaconda, used Pyspark libraries of numpy, scipy and sklearn.

Finally, the spark framework provides the ease of use feature, security and phenomenal speed to Big Data.

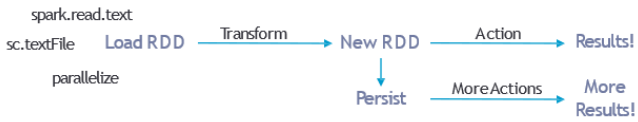


Figure 1. Spark programming model

IV. EXPERIMENT

Data for this analysis were collected from data.gov.mn published by Health and Safety Authority. The Health and Safety Authority (HSA) is the national body in Ireland with responsibility for occupational health and safety. The Workplace incidents in Ireland from 2017-2021, and consists of incident type, NACE sector, gender, age group, year, Employment status and trigger.

4.1. Datasets Identification

Ten factors were selected as inputs (independent variables) in this investigation as shown in Table 2.

Table 2. Input parameters

Input 1	Age group
Input 2	Gender
Input 3	NACE sector
Input 4	Incident type
Input 5	Employment status
Input 6	Year
Input 7	Trigger

The first input represents the age group, which is seven groups of aged victims. The second input represents the gender of the victims. The third input presents a pan-European classification system that groups organizations according to their business activities, which is nineteen types of economic area classification. This search facility, provided by the Central Statistics Office, allows users to find NACE codes for the various NACE categories. The fourth input depicts the four types of incidents; fatality, non-fatal injury, non-work related and dangerous occurrence. The fifth input presents five types of employments status, which is employee, family worker, non-worker, trainee, employment status unknown and self-employed. The sixth input is the year of incidents that happened. The tenth input represents the trigger of incidents, which is forty-three reported triggers.[3]

4.2. Data Preprocessing

First, imported the required libraries which is pyspark, SparkSession, pandas, numpy, seaborn, keras sequential, dense. By Spark SQL, read a csv dataset file into Spark DataFrame from Hadoop /user1 by SparSession. The inferred schema can be visualized using the printSchema() method. Figure 2 shows the first view of spark Dataframe.

```
root
 |-- Age: string (nullable = true)
 |-- Gender  : string (nullable = true)
 |-- NACE Sector: string (nullable = true)
 |-- Incident Type: string (nullable = true)
 |-- Employment Status  : string (nullable = true)
 |-- Year: integer (nullable = true)
 |-- County: string (nullable = true)
 |-- Trigger: string (nullable = true)
```

Figure 2. First view of Spark Dataframe

The types of variables are all string except variable of "Year". The shape of the data was 45787 rows and 8 columns. Check for duplicates, method:

"print('rows:{0}'.format(sdf.count()))" (2)

The total number of duplicates is 26504 that are exact copies of each other. To drop these rows by using the ".dropDuplicates(...)" method. Second, checked a missing observation of the DataFrame. Check the missing observations, method:

"Dataframe.select(*(sum(col(c).isNull().cast("int")).alias(c) for c in Dataframe.columns)).show()" (3)

A total number of the missing observations was 15. To drop any rows that have missing data is "dataframe.dropna()" method. After this, checked one more time the missing observations. Eventually, the final data shape of DataFrame was 26489 rows and 8 columns.

Before training the neural networks, the Data was not suitable for a basis descriptive statistics and Artificial neural networks because of indexes were all a string type. Therefore, the type variables of DataFrame were replaced index of string to integer by using "Dataframe.na.repalce()" method. All variables replaced by number (Gender: ['Male'=0, 'Female'=1,'Gender unknown'=2], etc.). Following this, converted type of variables string to integer by using "df.withColumn("col",col("col").cast('int')) method. Finally, all variables of DataFrame were able to perform descriptive statistics and Artificial neural network (ANN).

4.3. Model Building and Setting

This investigation is based on the model having the seven inputs and output variable (Incident type).

The neural networks proposed for this research are multilayer feedforward neural networks operating under supervised learning; they consist of three layers including one input layer, one hidden layer and one output layer.

The input layer, the hidden layer, and the output layer are fully interconnected. The weights of the connections and the biases are initialized randomly by the system, and adjusted through the learning process.

The creation of the model starts from install Scipy (including numpy), Keras and backend (Tensorflow). There are six input variables and one output variable which is incident type. The learning model to map rows of input variables (X) to an output (y). Then, split X, y variables into Training and Testing parts and each of variables' test size is configuration 0.10. The data stored in a 2D array where the first dimension is rows and the second dimension is columns, e.g. After this, specify the number of neurons or nodes in the layer as the first argument, and specify the activation function the activation argument. In this model, used the rectified linear unit activation function referred to as "relu" on the first two layers and the Sigmoid function in the output layer. Because, the model's output neuron produces a probability between 0 and 1, and the neuron with highest probability choose as the predicted class.

In the stage of compile, I used cross entropy as the loss argument for a binary classifaction problems and is defined in Keras as "binary_crossentropy". Because as mentioned before, where the target variable has two possible values. The optimizer defined as the efficient stochastic gradient descent algorithm "adam". The figure 3 shows a summary of the model.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_3 (Dense)              (None, 12)                84

 dense_4 (Dense)              (None, 8)                 104

 dense_5 (Dense)              (None, 1)                 9

=================================================================
Total params: 197
Trainable params: 197
Non-trainable params: 0
_________________________________________________________________
```

Figure 3. Model summary

Training occurs over epochs and each epoch is split into batches. Fit the model with 100 epochs and batch size is 10. Finally, the input layer, the hidden layer, and the output layer are fully interconnected.

As a result, the model's accuracy score is 98.94%.

V. CONCLUSION

In this paper, I have presented a work place incident type prediction model of artificial neural network by using Apache Spark dataframe. The accuracy of model is 98.94%. The dataset was collected from data.gov.mn published by Health and Safety Authority which is The Workplace incidents in Ireland from 2017-2021, and consists of incident type, NACE sector, gender, age group, year, Employment status and trigger. The paper has provided the type of workplace incidents, performance of Apache Spark and application of ANN.

VI. REFERENCES

[1] Health and Safety authority, Annual Review of Workplace Injuries, Illnesses and Fatalities 2019-2020, 2021, pp.7.

[2] Rohaifa Khaldi, Abdellatif El Afia, Raddouane Chiheb and Rdouan Faizi, "Artificial Neural Network Based Approach for Blood Demand Forecasting: Fez Transfusion Blood Center Case Study" in Mohammed V University, Rabat, Morocco, pp. 2.

[3] Health and safety authority, "Guidance on the Safety, Health and Welfare at Work (Reporting of Accidents and Dangerous Occurrences) Regulations 2016"

[4] A. Brahmane, B. Krishna, "Big data classification using deep learning and apache spark architecture", *Neural Computing and Applications*, 2021.

[5] A. Gupta, H. Thakur, R. Shrivastava, P. Kumar, S. Nag, "A Big Data Analysis Framework Using Apache Spark and Deep Learning", 2017.

[6] "What is Apache Spark? Spark tutorial guide for beginner", janbakstraining.com, 2018.

[7] T. Rivas, M.Paz, J.E. Martin, J. Matías, J. García, J.Taboada "Explaining and predicting workplace accidents using data-mining techniques", *Reliability Engineering & Sytem Safety,* 2011.

[8] N. Emma, W. Charlotte, R. Alex, K. Susanne, "Health care workers' experiences of workplace incidents that posed a risk of patient and worker injury: a critical incident technique analysis", 2021.

[9] S. Adhikesaven, "An Industrial Workplace Alerting and Monitoring Platform to Prevent Workplace Injury and Accidents", 2022.

[10] H. Chiroma, U.A. Abdullahi, S.M. Abdulhamid, A.A. Alarood, L.A. Gabralla, N. Rana, L. Shuib, I. Hashem, D. Gbenga, A. Abubakar, A. Zeki, T. Herawan "Progress on Artificial Neural Networks for Big Data Analytics: A Survey*", Federal College of Education (Technical), Gombe, Nigeria,* 2018.