International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

EARTHQUAKE HAPPENS OR NOT USING DATASCIENCE TECHNIQUE

SALMAN.S , SURYA.MR, SYED SHAMEER.S

FINAL YEAR CSE - DHAANISH AHMED COLLEGE OF ENGINEERING

ABSTRACT

With the increasing popularity of online social networks, spammers find these medias easily accessible to trap users in malicious activities by posting spam messages. Here, Twitter platform is focused to analyze and perform the detection of spam tweets. To avoid the activities of spammers, Google Safe Browsing and Twitter's Bot Maker tools detect and block spam tweets. Even though these tools can block malicious links they cannot protect the user from spammers' activities in real-time as early as possible. Some of the approaches are only using user-based features while others are based on tweet based features only. There is no comprehensive solution which can enhance and consolidate tweet's text information along with the user based features. For solving this issue, a framework is proposed which considers the user based and tweet based features for classifying the tweets. The advantage of using tweet text feature is that the spam tweets can be identified and detected even if the spammer creates and login with a new account, which was not possible only with the user and tweet based features. Estimation and evaluation of the entire process is performed using four different machine learning algorithms namely - SVM, KNN, ANN, RF and Naive Bayes. With the help of Neural Networks, an accuracy of 91.65% can be achieved which surpass the existing solution approximately by 18%.

INTRODUCTION

1.1 DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Natural Language Processing (NLP):

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts.

OBJECTIVES

The goal is to develop a machine learning model for Earth Quake Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

LITERATURE SURVEY

General

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

Loan default trends have been long studied from a socio-economic stand point. Most economics surveys believe in empirical modeling of these complex systems in order to be able to predict the loan default rate for a particular individual. The use of machine learning for such tasks is a trend which it is observing now. Some of the survey's to understand the past and present perspective of loan approval or not.

Review of Literature Survey

Title : A Review of Application of Data Mining in Earthquake Prediction

Author : G.V. Otari, Dr. R.V. Kulkarni

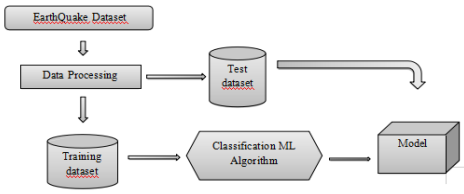
Year : 2012

SYSTEM ANALYSIS

3.1 Existing System:

They proposed two methods and named them as picking target window prediction PTWP and multitarget regression (MTR) tasks to determine earthquake source parameters. The Ncheck algorithm improves window prediction selection to reduce false alarms in multistation waveforms that have noise, and MTR with hard-shared orthogonal are proven to improve earthquake parameter determination performance. Our system can provide reliable earthquake parameters. The three stations with three-component seismogram traces are represented in red, green, and blue components to form pixels in a row in one frame in a 10 s window. The sampling rate at each station varied between 20 and 25 Hz and was then normalized to 20 Hz. They used a band pass filter to minimize noise and normalize each stream by dividing its absolute peak amplitude. The data set has high noise for 506 seismic events and has a peak SNR of less than 50 dB.

PROPOSED SYSTEM



Earthquakes are a natural disaster that can cause a lot of damage to both lives and properties. The machine learning is applied to every field where the dataset can be used to learn patterns

and then from that pattern the prediction can be done. Our objective is to build a machine learning model that uses the past earthquake related dataset the data is pre-processed by using variable identification that is finding the dependent and independent variables after that the data is used to train the model by using machine learning libraries. Different algorithms are used to compare the model and Earthquakes are a natural disaster that can cause a lot of damage to both lives and properties. The machine learning is applied to every field where the dataset can be used to learn patterns and then from that pattern the prediction can be done. Our objective is to build a machine learning model that uses the past earthquake related dataset the data is pre-processed by using variable identification that is finding the dependent and independent variables after that the data is used to train the model by using machine learning libraries. Different algorithms are used to compare the model and the performance metrics are calculated .

PROJECT REQUIREMENTS
General:

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system.The following are the requirements that are to be discussed.

1. Functional requirements

2.Non-Functional requirements

3.Environment requirements

A. Hardware requirements

B. software requirements

FUNCTIONAL REQUIREMENTS:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

NON-FUNCTIONAL REQUIREMENTS:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

Environmental Requirements:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

2. Hardware requirements:

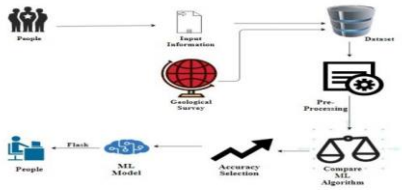
Processor : Pentium IV/III

Hard disk : minimum 80 GB

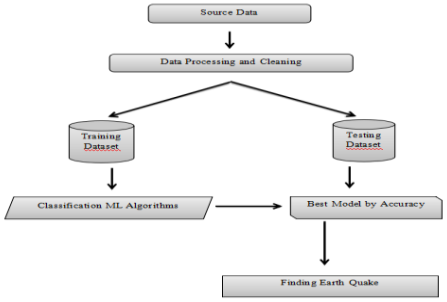
RAM : minimum 2 GB

evaluated.

SYSTEM DESIGN
System Architecture:



Workflow Diagram:



MODULE DESCRIPTION

12.1 LIST OF MODULES

- Data Pre-processing
- Data Analysis of Visualization
- Comparing Algorithm with prediction in the form of best accuracy result
- Deployment Using Flask

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

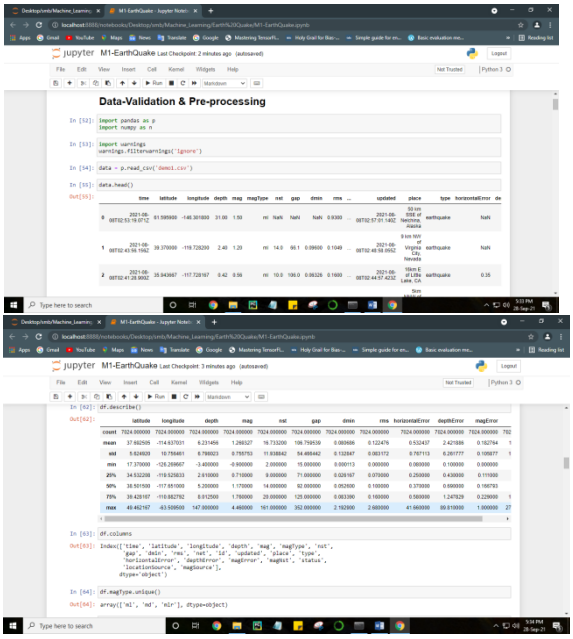
- ✓ User forgot to fill in a field.
- ✓ Data was lost while transferring manually from a legacy database.
- ✓ There was a programming error.
- ✓ Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.



ALGORITHM AND TECHNIQUES

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

Naive Bayes algorithm:

- The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a

given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.
- Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms.Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.
- Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location.
- Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score is will be find out. This application can help to find the Prediction of Earth Quake.

REFERENCE

REFERENCE [1] S. Widiyantoro et al., "Implications for megathrust earthquakes and tsunamis from seismic gaps south of Java Indonesia," Sci. Rep., vol. 10, no. 1, pp. 1–11, Dec. 2020, doi: 10.1038/s41598-020-72142-z. [2] R. E. Abercrombie, M. Antolik, K. Felzer, and G. Ekström, "The 1994 Java tsunami earthquake: Slip over a subducting seamount," J. Geophys. Res., Solid Earth, vol. 106, no. B4, pp. 6595–6607, Apr. 2001, doi: 10.1029/2000jb900403. [3] Z. E. Ross, Y. Yue, M. Meier, E. Hauksson, and T. H. Heaton, "PhaseLink: A deep learning approach to seismic phase association," J. Geophys. Res., Solid Earth, vol. 124, no. 1, pp. 856–869, Jan. 2019, doi: 10.1029/2018JB016674. [4] E. L. Olson and R. M. Allen, "The deterministic nature of earthquake rupture," Nature, vol. 438, no. 7065, pp. 212–215, Nov. 2005, doi: 10.1038/nature04214. [5] O. M. Saad, K. Inoue, A. Shalaby, L. Samy, and M. S. Sayed, "Automatic arrival time detection for earthquakes based on stacked denoising autoencoder," IEEE Geosci.=10.1029/2018GL077870. [7] W. Zhu and G. C. Beroza, "PhaseNet: A deep-neural-network-based seismic arrival-time picking method," Geophys. J. Int., vol. 216, pp. 261–273, Oct. 2018, doi: 10.1093/gji/ggy423. [8] E. Pardo, C. Garfias, and N. Malpica, "Seismic phase picking using convolutional networks," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 9, pp. 7086–7092, Sep. 2019, doi: 10.1109/TGRS.2019.2911402. [9] O. M. Saad and Y. Chen, "Earthquake detection

and P-wave arrival time picking using capsule neural network," IEEE Trans. Geosci. Remote Sens., early access, Sep. 28, 2020, doi:11, no. 1, pp. 1–12, Dec. 2020, doi: 10.1038/s41467-020- 17591-w. [11] A. Lomax, A. Michelini, and D. Jozinovi´c, "An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network," Seismolog. Res. Lett., vol. 90, no. 2A, pp. 517–529, Mar. 2019, doi: 10.1785/0220180311. [12] O. M. Saad, A. G. Hafez, and M. S. Soliman, "Deep learning approach for earthquake parameters classification in earthquake early warning system," IEEE Geosci. Remote Sens. Lett., early access, Jun. 9, 2020, doi: 10.1109/lgrs.2020.2998580. [13] M. Kriegerowski, G. M. Petersen, H. Vasyura-Bathke, and M. Ohrnberger, "A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms," Seismolog. Res. Lett., vol. 90, no. 2A, pp. 510–516, Mar. 2019, doi: 10.1785/0220180320. [14] D. Jozinovi´c, A. Lomax, I. Štajduhar, and A. Michelini, "Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network," 2020, arXiv:2002.06893. [Online]. Available: https://arxiv.org/abs/2002.06893 [15] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, [16] M. P. A. van den Ende and J. P. Ampuero, "Automated seismic source characterization using deep graph neural networks," Geophys. Res. Lett., vol. 47, no. 17, pp. 1–11, Sep. 2020, doi: 10.1029/2020GL088690. [17] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 2, pp. 497–504, https://arxiv.org/abs/2006.01332 [19] A. Howard et al., "Searching for mobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140. [20] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Jun. 2019, pp. 10691–10700. [21] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, arXiv:1706.05098. [Online]. Available: 10.1142/S012906571950014X. [23] W. Hu, L. Xiao, and J. Pennington, "Provable benefit of orthogonal initialization in optimizing deep linear networks," 2020, arXiv:2001.05992. [Online]. Available: https://arxiv.org/abs/2001.05992 [24] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," in Proc.2020, doi: 10.1038/s41598-020-58908-5. [26] A. D. Nugraha et al., "Hypocenter relocation along the Sunda arc in Indonesia, using a 3D seismic-velocity model," Seismolog. Res. Lett., vol. 89, no. 2A, pp. 603–612, Mar. 2018, doi: 10.1785/0220170107.