International Journal of Scientific Research in Engineering & Technology
Volume4, Issue5 (September-October 2024), PP: 01-07.
https://www.doi.org/10.59256/ijsreat.20240405001
www.ijsreat.com
ISSN No: 2583-1240

Machine Learning-Based Approach: Enhancing Social Media Platform

Sivakumar Nagarajan
Technical Architect, I & I Software Inc, 2571 Baglyos Circle, Suite B-32, Bethlehem, Pennsylvania, USA.

To Cite this Article: *Sivakumar Nagarajan, "Machine Learning-Based Approach: Enhancing Social Media Platform", International Journal of Scientific Research in Engineering & Technology Volume 04, Issue 05, September-October 2024, PP: 01-07.*

Abstract: *Machine Learning is a technique that aims to learn from data over a task based on certain performance measures. With the flooding of data and increased computational ability, machine learning algorithms have a huge research scope. The task of identifying the correct machine learning algorithm for an application is very challenging without analysing the basic information about the problem such as the domain of the undertaken problem, the features available in the data, etc. The movie industry, a huge part of entertainment industry, has seen a phenomenal growth throughout the globe in recent times. A movie can capture the attention of a viewer and can trigger cognitive and emotional processes in the brain. In this research, the emotional outcome of the viewer was analysed while they watch the movie before its actual release that is, during its preview. Traditionally Functional Magnetic Resonance Imaging device was used to assess human brain activity but proved to be non-feasible and costly so EEG sensors were used to monitor and record the functioning of the brain of volunteers for further analysis. We proposed a model to use the collected data through EEG sensors were analysed using artificial neural network which was Used to find high and low of different brain waves mapping to the emotions depicted in every scene of the movie. Performance measures such as accuracy, precision and recall were calculated to validate our proposed model. Our proposed model resulted in providing assistance to movie makers who could Study the pulse of audience before the actual release and could incorporate changes if necessary.*

Key Word: *Machine Learning, EEG sensors, Magnetic Resonance Imaging,*

I.INTRODUCTION

Data, being the new oil, is present ubiquitously because of multiple sources that generate data. In the recent data-driven age and with the advancements in technology there is a wide scope of analysing the data using Artificial Intelligence (AI) and Machine Learning (ML) algorithms.

Artificial Intelligence

Artificial intelligence is the act of demonstrating and emulating intelligence in terms of computational processes. One of the first and foremost approaches to understand Artificial Intelligence was through Turing test, which was proposed by Alan Turing. For a computer machine to pass a Turing test, the following capabilities are essential:

- Natural Language Processing the ability of a computer to understand communicative language
- Knowledge representation the ability to store information before or during the interrogation
- Automated reasoning the ability to use the stored information and derive conclusions

The main purpose of AI machines is to augment the human capabilities and automate a few processes. The ultimate aim of AI is not to overpower humans but to assist in certain areas where human interventions are not needed. Traditional AI systems do not possess memory, that is, they cannot use any past experiences to predict future decisions. Most AI algorithms learn from the data and with Machine Learning emerging as a prominent field in the recent days, focuses are towards the ML algorithms and the performance metrics of those algorithms on the data. ML is a subset of AI which rely heavily on statistical techniques.

Machine Learning

Machine Learning is a branch of Artificial Intelligence which is gaining a lot of attention among researchers recently, the reason being emergence of big data . The data is available in huge sizes and variety of formats as there are plenty of sources which generate the data. The speed at which the data is generated is also quite high. With these emergence, ML algorithms prove to be very useful for wide range of applications ranging from Finance, Telecom, Healthcare, Retail, Education and other domains.

Machine learning is used to optimise performance metrics using past examples which is called as training data. First, a model is built using the training performance metrics. The model can be predictive, which are used to predict output for test data (future sample data); or can be descriptive, which are used to describe the data in order to obtain knowledge (finding patterns or association rules) out of the data; or can be both. In the present scenario, ML can be broadly categorised into:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

This type of learning is a predictive approach of machine learning. The model is first built using training data (past examples). Since the main goal of this type of learning is prediction, if the new instances are following similar pattern to the past data, then predictions can be made accurately. The supervised learning is further classified into two segments:

Regression

This is a type of predictive modelling where continuous values are predicted. For example: predicting the salary of an employee based on their years of experience. A model can be built (also called as fitting a regression line/curve) and trained using regression techniques by finding the suitable regression coefficients. Then the model can be used to predict the values for new instances. The regression model can be evaluated using performance metrics such as R-squared score , Root Mean Squared Error (RMSE) values, etc. Regression is broadly classified into two types Linear regression, Polynomial regression. If the data is spread in a linear fashion, then linear regression is used. If the data is spread in a non-linear fashion, polynomial regression is used.

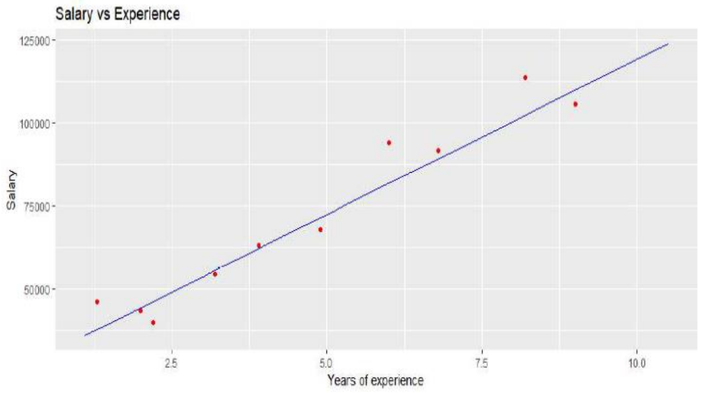


Figure 1 Linear Regression showing Salary (in $) Vs Years of Experience

The blue line in Figure 1 indicates the regression line that was fitted for the data and the red points are the actual values of salaries which are available. The difference between the predicted value and the actual value are called as the error values; and with these error values, the performance metrics such as R-squared or RMSE values are calculated.

Classification

This is a type of predictive approach, where the class labels are known beforehand; so the model knows what to predict. For example: based on a digital marketing for a product, whether an email is a spam mail or not is a classification problem. In these type of problems, the class labels Spam or Not Spam will be available along with the past samples. A model can be built and trained on the available samples and this model can then be used to predict for new instances whether the user will buy that product or not. The classification model can be evaluated by the performance metrics such as accuracy, precision, recall and F-Score

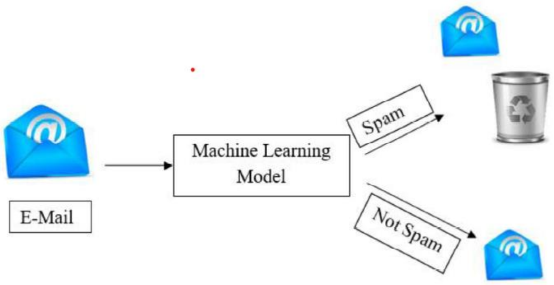


Figure 2 E-Mail spam classification

Figure 2 shows an example of e-mail spam classification. A machine learning (classification) model is built with the available data which provides information on whether an e-mail is a spam mail or not. This model is then used to predict a new e-mail as a spam mail or a normal mail.

II.UNSUPERVISED LEARNING

This type of learning searches for patterns available in the data or finds association in a transactional data. There are no target variables or class labels available to predict. Unsupervised learning is further classified into three types:

Clustering:

This is a type of unsupervised learning which forms/creates clusters from the existing data. To form the clusters, there are lot of techniques such as K-Means, Hierarchical clustering algorithms for example, data containing the details about customers visiting a mall; if the age of customers and the amount that they spend are known, then customers can be clustered, i.e., similar customers can be grouped and be present in a cluster.

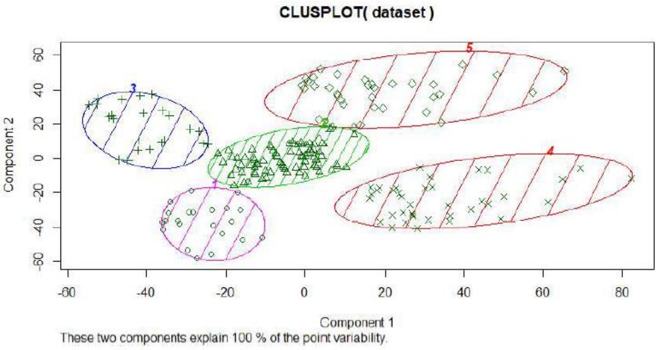


Figure 3 Clustering customers based on their age and spending amount

Association Rule Learning:

This type of unsupervised learning looks for association in transactional data. Transactional data is a type of data which contains transaction details of products such as billing in a supermarket. These data will contain list of products purchased which provides an option to search for associations among those lists.

Algorithms such as Apriori and Eclat provide the associations formed between products which are proved as efficient algorithms. A classic example of association rule mining is the Market Basket Analysis. This is used to provide the associations between various products.

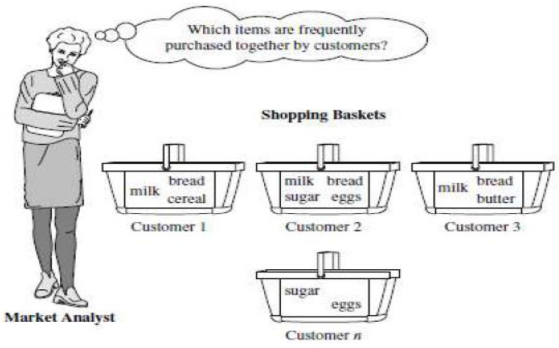


Figure 4 Market Basket Analysis

Figure 4 shows the Market Basket Analysis or various customers. Association rules are formed by mining the basket data. From the Figure 4, an association rule can be derived such that: whenever milk is purchased, bread is also purchased.

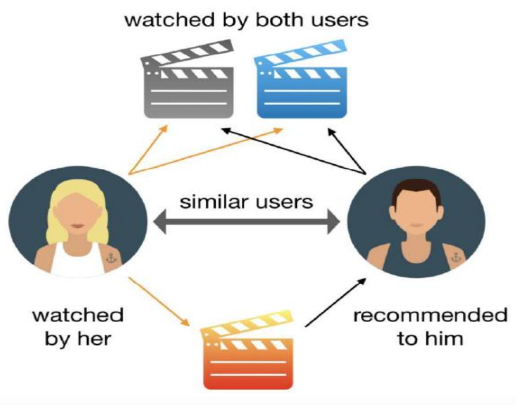


Figure 5 Movie Recommendation System

IV. REINFORCEMENT LEARNING

This type of learning focuses on solving problems that requires decision to be made sequentially. There are various possible solutions to a problem and the model will return a state. The state is either rewarded or given a penalty based on the output. The model learning is incremental and the best solution is selected based on the maximum reward. Figure 6 shows the general reinforcement learning framework.

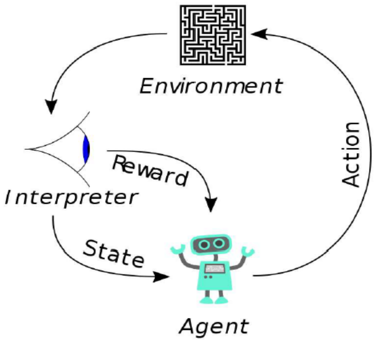


Figure 6 Reinforcement Learning

Machine learning framework involves the following phases as shown in Figure7 below.

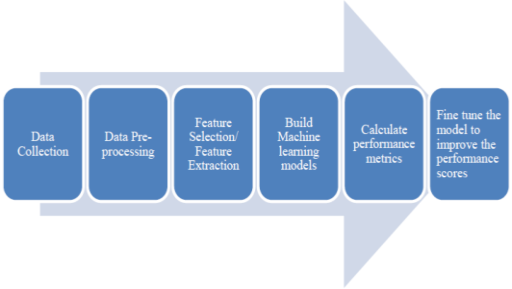

Data Collection
Data Pre-processing
Feature Selection/ Feature Extraction
Build Machine learning models
Calculate performance metrics
Fine tune the model to improve the performance scores

Figure 7 Machine Learning Framework

V.CHALLENGES OF MACHINE LEARNING

Though machine learning algorithms are seemingly very popular these days with the huge amount of data, there are lot of challenges that are needed to be sorted out. The challenges are discussed below:

Insufficient Data used in Training the Machine Learning Model

For a human to identify a fruit as apple or orange is very simple, whereas, the same for a machine is quite difficult. This is because, the machine needs large amount of data to be trained first so that it can predict accurately in future. For complex applications such as image processing, speech processing, text analytics, etc., the amount of training data to be used may even be in millions.

Outliers and Missing Values Present in the Data

There are possibilities that the training data, that is used to train the machine learning model, can have outliers and missing values. These are to be treated before proceeding to build the machine learning model. The reason for treating these outliers and missing values are that they will lead to incorrect results in predict time are spent in data pre-processing. Some of the solutions for this type of challenge are:

- If the presence of outliers is detected, it can be simply discarded or the errors can be fixed manually.
- If missing values are present, the ratio of missing values is found. If the ratio is very less, then those data can be discarded. Else, the missing values can be filled using statistical measures of central tendency such as mean, median or mode depending on the type of missing data.

Bias and Variance

Bias is defined as the difference between the actual value and the value that is predicted by our model. Variance is the variability present in the predicted value for a given data point that says how the model is spread. The bias variance trade-off is a very important concept in machine

learning models. This can be easily understood using bulls-eye diagram.

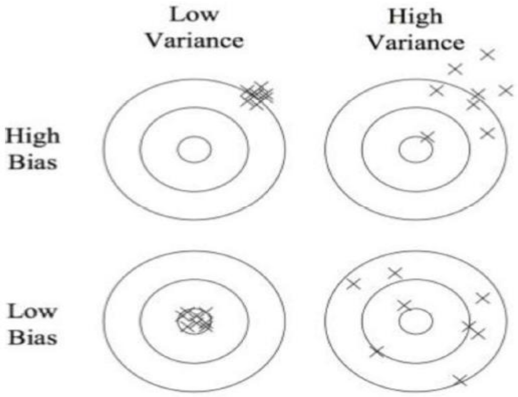


Figure 8 Bulls-eye diagram explaining bias and variance

The perfect scenario for a machine learning model is that the model should have low bias and low variance which is in the bottom left of the image. A model with high bias and high variance will not be considered as a good machine learning model.

VI.FEATURE SELECTION

Feature selection is the mechanism of selecting a subset of features from the given set of features in order to reduce the redundant and irrelevant features without much loss of information. A dataset may contain a wide variety of features, few of which may be redundant or irrelevant to that problem. In regression techniques, the feature selection can be done by deploying simple statistical techniques such as correlation and Variation Inflation Factor (VIF) among the features. Feature selection is very essential in regression techniques because without feature selection few problems such as multicollinearity can occur. Such multicollinearity can lead to poor performance metrics for the regression models. To avoid this problem, correlation and VIF can be very useful. Through these techniques, some features can be dropped.

Ensemble classifiers have shown that the combination of various classifiers along with the relevant feature selection methods provide greater accuracy in predictive analytics. Another striking advantage of using these feature selection methods is that the time used for training the model could be drastically reduced because of the limited number of features being used. In some scenarios, domain knowledge can also help in selecting the appropriate features. In scenarios where there is no prior knowledge on domains, there are some feature selection algorithms that can be used to select the necessary features. There are three main categories of feature selection techniques wrappers, filters and embedded methods. Wrapper methods use black box techniques to score the subsets of variables according to predictive power.

Filter methods perform the variable selection as a pre-processing irrespective of the chosen predictor. Embedded methods, however, select the variables as a training process and they depend on the learning machines. Two of the widely used feature selection algorithms are discussed below:

a. Chi-Square Feature Selection

Chi-square feature selection explained comprehensively in statistics is generally used to test the independence of two events. In feature selection, Chi-square method takes oneevent as the occurrence of an attribute and another event as the occurrence of the class.

b. Boruta Feature Selection Algorithm

Boruta is an all-relevant feature selection wrapper algorithm. The important features are identified by comparing the importance of original attributes with the importance that is achieved at random, done by estimation of using the permuted copies. This is taken from the literature of Kursa & Rudnicki (2010). When the Boruta is run on a dataset, an important method called as attStats is shown in an attribute-centred way. Boruta uses shuffling principle to iteratively compare the importance of attributes. The procedure adopted the implementation of Boruta algorithm is as follows:

- An extensive system is built where each descriptive variables are replicated. These values are randomly permuted across objects.
- Several random forest runs are performed on these variables maintaining randomness among the variables.
- For each run the importance values are calculated for all the Attributes.

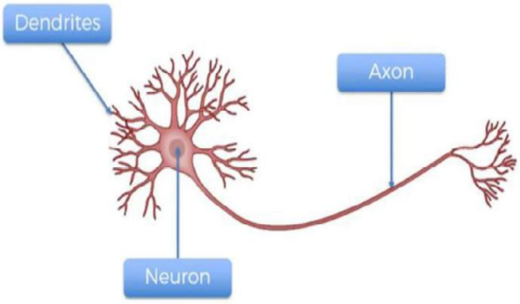
VII.TESTING AND VALIDATING

Once a machine learning model is developed, it is always a good practice to test and monitor how well the model works for new data. However, it is better to test the model before deploying rather than directly testing in the live environment. In order to test the machine learning model, the dataset is split into training dataset and test dataset. The model is trained using the training dataset and then tested with test dataset to understand how the model is performing. Usually, 80% of the data will be used as training dataset and 20% will be used for testing purposes. Since the model is built using training data, it will definitely produce some errors for the new samples (test data). This error is termed as generalization error. This error is also called as out-of-sample error. This value will give an estimate of how well our machine learning model behaves when it sees a new data.

If the model has a very less training error but the generalization error is more, then the model is said to be over fitted. To avoid overfitting, cross validation can be deployed on the data. Cross validation is a technique where many folds of subset are taken and each model is trained with different training data in each fold and tested upon the remaining data. In this way, the overfitting problem can be solved.

VIII.DEEP LEARNING

Deep learning is a subset of machine learning which is inspired from the structure of human brain called as neural networks. The ultimate goal of deep learning algorithms is to mimic the activity of human brain. This is because, human brain is the most powerful tool which learns fast and adapt to the environment quickly. Neurons are the most important features of a neural network. A biological neuron consists of axons and dendrites as shown in the Figure.



Dendrites are the signal receivers of a neuron and axons are the signal transmitters of a neuron. The regions where signals are transmitted from one axon to another is called as synapse.

The importance of deep learning started very recently because it has lot of advantages. One of the main advantages is that feature selection is no longer a separate process. The neural network which uses weights for the features learns about the importance of the features during training phase. In traditional machine learning algorithms for image classification, feature selection has to be done separately and then the relevant features should be given as input to the algorithm. However, in deep neural networks, the image can directly be fed into the neural network architecture and the system will take care of the relevant features during the training phase.

Another striking feature about deep learning algorithms is that as the size of the data becomes huge, the performance also increases. For smaller amount of data there are chances that traditional machine learning algorithm performing better than deep learning algorithm. But when the data size explodes, there is a clear increase in performance by deep learning algorithms as depicted in the Figure.

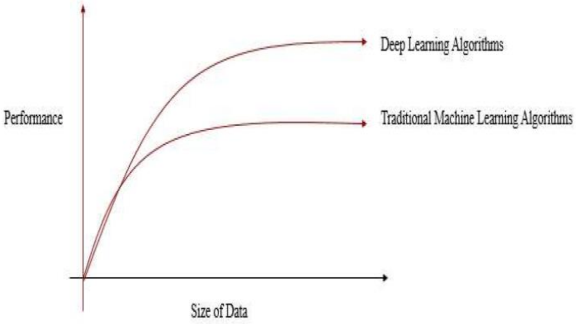


Figure 11 Performance of Traditional ML algorithms Vs Deep learning algorithms with the change in size of data

The deep neural networks have layered architectures with each layer consisting of a number of nodes. The input layer contains the features present in the input data. Each feature is taken as a node in the input layer. The next layer is called as hidden layer. This layer is mainly responsible for performing all the necessary operations including the feature selection, updating the weights, training the data, etc. There is no rule of thumb to choose the number of hidden layers in a neural network. This is a trade-off. The final layer in a neural network is the output layer. This layer provides the final output of the neural network. Neural network can be used in supervised learning as well as unsupervised learning. In order to train the system and learn the features, neural networks use a function called as activation function. Some of the widely used activation functions are:

- Threshold function
- Sigmoid Function
- Rectifier function
- Hyperbolic tangent function

IX. CONCLUSION

Work we have investigated certain machine learning algorithms for societal applications. Various performance measures were analyzed and few comparative studies on the other state-of-art algorithms with respect to the proposed algorithm were also done. A Pre-release analyser was built using the data collected from various users using EEG device which was aimed at helping the film makers in used in predicting the emotions of the viewers. The model was tested on the created dataset and it provided 85% accuracy in predicting the emotions of the viewers.

The proposed pre-release analyser using deep learning can provide an opportunity to the film makers to know in advance whether the expected emotions are evoked among the set of viewers. the future by taking other repeat and near repeat crimes and incorporating other advanced machine learning/deep learning techniques and hybrid architectures to study their performance measures.

REFERENCE

1) *Mingoia, J.; Hutchinson, A.D.; Wilson, C.; Gleaves, D.H. The relationship between social networking site use and the internalization of a thin ideal in females: A meta-analytic review. Front. Psychol. 2017, 8, 1351*
2) *Verduyn, P.; Ybarra, O.; Résibois, M.; Jonides, J.; Kross, E. Do social network sites enhance or undermine subjective well-being? A critical review. Soc. Issues Policy Rev. 2017, 11, 274–302.*
3) *Yang, C.-C. Instagram use, loneliness, and social comparison orientation: Interact and browse on social media, but don't compare. Cyberpsychology Behav. Soc. Netw. 2016, 19, 703–708.*
4) *Lup, K.; Trub, L.; Rosenthal, L. Instagram# instasad?: Exploring associations among instagram use, depressive symptoms, negative social comparison, and strangers followed. Cyberpsychology Behav. Soc. Netw. 2015, 18, 247–252.*
5) *Dionísio, N.; Alves, F.; Ferreira, P.M.; Bessani, A. Cyberthreat detection from twitter using deep neural networks. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.*
6) *Taherdoost, H.; Madanchian, M. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. Computers 2023, 12, 37.*
7) *Jayatilake, S.M.D.A.C.; Ganegoda, G.U. Involvement of machine learning tools in healthcare decision making. J. Healthc. Eng. 2021, 2021, 6679512.*