Ternary Classification technique based on Factorization

D.Dineshwar
dineshsupertechno@gmail.com

Abstract: Now-a-days it's quite common that from start-ups to MNC companies, individuals to large organizations are turning towards automation of tasks. Here some of the automation task is done using "If-Else" but in some tasks we require computer understand the input and give the output . In these tasks the system directly interacts with user with natural language. In such task we require an algorithm that captures the intension of user and presents the output. This paper focuses on the ternary classification of tasks using factorization method. This paper concludes in presenting a new approach to classify the text into it appropriate class.

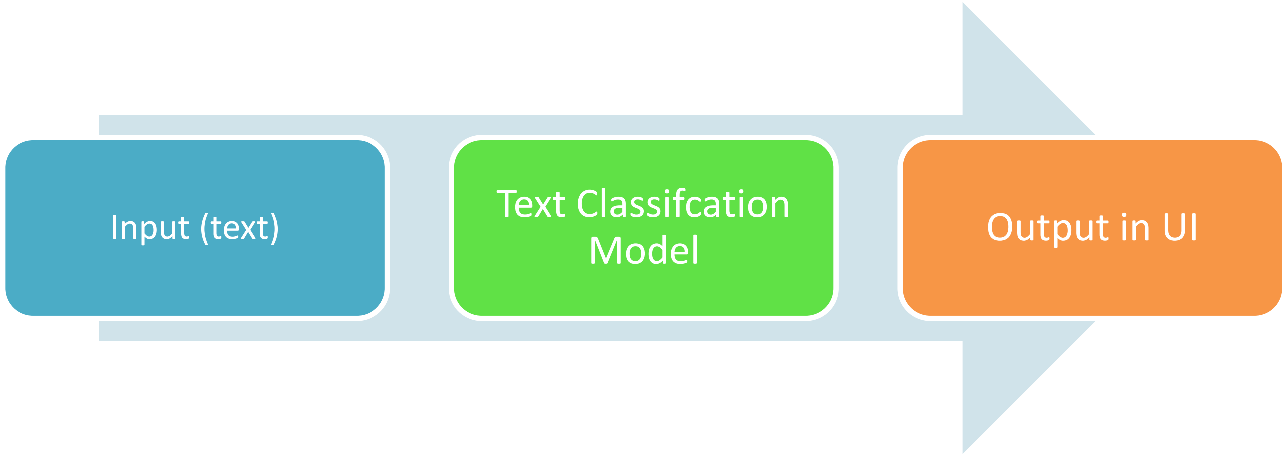
Keywords: Automation, Classification, Factorization.

INTRODUCTION

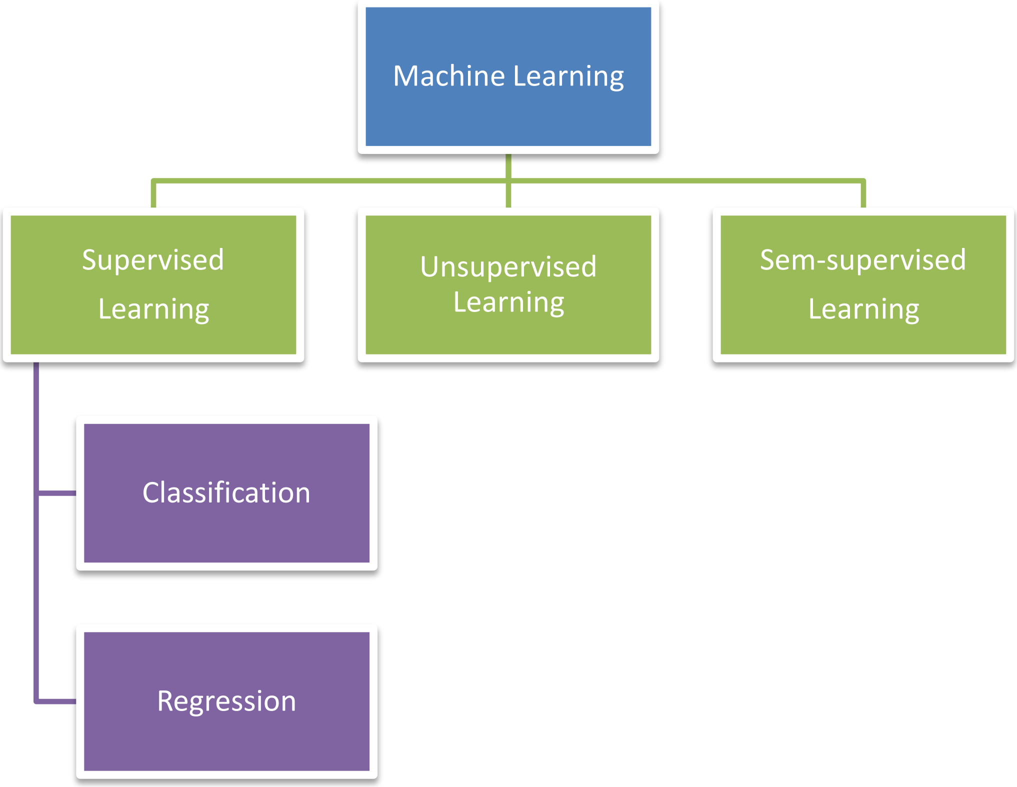
Text classification is machine learning technique that assigns the text input to set of pre-defined categories. This is one of NLP (Natural Language Processing) technique used to find the category of text input. Text classification works on textual data.

The text classifiers can be used to categorize data extracted from documents, textual inputs or data from web. For example we can categorize the articles in newspaper. This is one of the basic task in NLP with wide range of applications from sentiment analysis to detecting the urgency based on document.

The text classifier takes the input textual data and performs some operation and outputs the category using an GUI.



Overview of Machine learning:



1. **Supervised Learning:**

In this method we train or build the model by providing it with both features/inputs as well as labels. Here in this method the model trains based on inputs and labels that are provided during training, so that model learns how to map inputs to output.

In supervised learning we have two categories REGRESSION, CLASSIFICATION.

a) **Regression:** It is a supervised machine learning technique which aims to predict the quantitative values such as individual's salary, stock market price…etc
Techniques: 1) Linear Regression
2) Logistic Regression
3) Ridge Regression

b) **Classification:** It is a supervised machine learning technique whose aim is to predict the qualitative value such whether patient has heart disease or not, whether it rains or not..etc.
Techniques: 1) Support vector machines

2) K-Nearest Neighborhood
3) Decision Trees

2. **Un-Supervised Learning:**

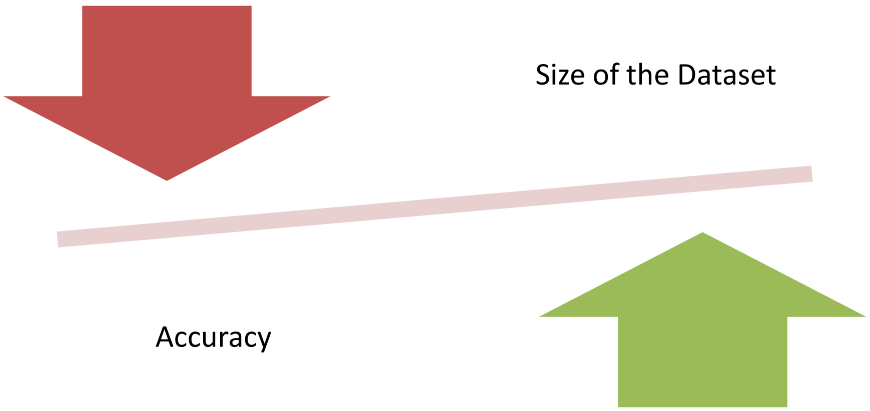
In this method we provide the model only with features, which means the model should learn itself as we don't have any labels given during training. Here the machine itself find the pattern among the data and labels itself.

Techniques: 1) Clustering

3. **Semi-Supervised Learning:**

In this method we provide some labeled data and some un-labeled data during training process. Here the main task of this method is label the unlabeled data using the labeled data provided.

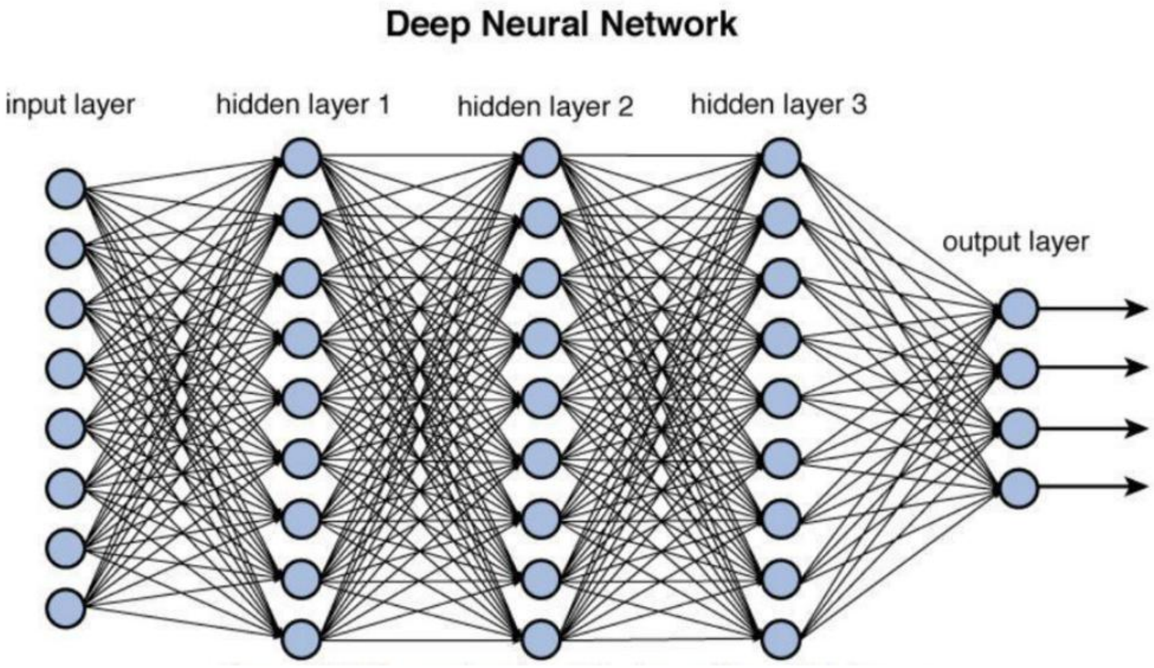
Baseline Approach



There is a trade-off between the size of dataset and the accuracy as they are inversely proportional in some cases. Here we have two different ways to solve this problem.

Method-I) Neural Networks:

We can classify a task using Deep Neural Networks, but in prior of feeding data to neural network we need to perform some operations on raw data like tokenizing, stemming. Apart from this neural networks should be feed with large datasets for better accuracy. So the down side of this technique is large dataset, high computational power and long duration to train the network.



Method-II) Classical Model:

There are many Classical models for text classification, but here we consider the model using frequency table. This model can be useful using relatively less data, but it doesn't have high accuracy. This model works with relatively small dataset, low computational power and less time to train the model but the down side is that it doesn't produce high accuracy as Neural Networks
.

Proposed Approach

This method could acquire better accuracy than Classical models while having medium sized datasets.

ALGORITHM (building predictive weights):

1. **Remove all stop words from the given dataset.**
2. **Now assign numbers 2, 3, 7 to classes in the given dataset.**
3. **Now multiply each and every example with even numbers starting from 2 to 'n' (i.e. n is number of example belongs to a single class in given dataset)**
4. **After multiplying with even numbers, now make a dictionary that contains all words which are keys and their values are the count of number associated to their example in given dataset and add them among all the classes in dataset.**
5. **Now check each and every key if it's values is a multiple of number's 2, 3 and 7. If it's a multiple of all three numbers then delete the key value pair from that dictionary**
6. **Running this process isolates the word which mainly constitutes the meaning of the task and removes the rest of**

ALGORITHM (Computing the output):

1) **Take the input and create 3 variables each one for each class.**
2) **Now add the values to variables which belong to particular class by looking at dictionary. Continue all this process for all words in given input.**
3) **Now find the maximum variable of all three variables.**
4) **Depending on maximum variable output the class belongs to respective class.**

Result and Conclusion

Successfully proposed a method that works relatively better on even medium size datasets. To be specific the above mentioned procedure works well for Task classification than on sentiment analysis.

CODE AVIALABLE AT :
https://github.com/dinesh9-ai/Ternary-Classification-technique-based-on-Factorization