ABSTRACT

The current way of checking subjective papers is adverse. Evaluating the Subjective Answers is a critical task to perform. When a human being evaluates anything, the quality of evaluation may vary along with the emotions of Person. In Machine Learning, all results are only based on the input data provided by the user.

Our proposed system uses machine learning and NLP to solve this problem. Our Algorithm performs a task like Tokenizing words and sentences, Part of Speech tagging, chunking, lemmatizing words and Wordnetting to evaluate the subjective answer. Along with it, our proposed algorithm provides the semantic meaning of the context. Our System is divided into two modules. The first one is extracting data from the uploaded answer documents and organizing it in the proper manner and the second is applying ML and NLP to the text retrieved from the above step and giving marks to them. The main purpose of this project is to reduce man power.

The software will take a scanned document of the answer as an input and then after the preprocessing step it will extract the text of the answer. This text will again go through processing and build a model of keywords and feature sets. The evaluator also provides the model answer sets, keywords and question specific things. The Model answer sets and keywords are categorized as mentioned will be the input as well. Classifier will then, based on the training, give marks to the answers. Marks to the answer will be the final output.

CONTENTS

LIST OF FIGURES

1.INTRODUCTION

The manual system for evaluation of Subjective Answers for technical subjects involves a lot of time and effort of the evaluator. Evaluating subjective answers is a critical task to Perform. When a human being evaluates anything, the quality of evaluation may vary along with the emotions of the person. Performing evaluation through computers using intelligent techniques ensures uniformity in marking as the same inference mechanism is used for all the students. In Machine Learning, all results are only based on the input data provided by the user. Our Proposed System uses machine learning and NLP to solve this problem.

Our Algorithm performs a task like tokenizing words and sentences, Part of speech tagging, Chunking, chinking, lemmatizing words and Wordnetting to evaluate the subjective answer. Our system will evaluate answers based on some keywords and also manpower will be saved. Our System is divided into two modules, Extracting the data from the printed documents of answer and organizing it in the proper manner and Applying ML and NLP to the text retrieved from the above step and giving marks to them. The software will take a printed copy of the answer as an input and then after, it will extract the text of the answer. This text will again go through processing to build a model of keywords and feature sets. Model answer sets and keywords categorized as mentioned will be the input .Based on the keywords written in the answer and the keywords in the dataset the application will provide marks in the certain range.Marks to the answer will be the final output. The need for online examination is mainly to overcome the drawbacks of the existing system.The main aim of the project is to ensure user friendly and more interactive software to the user.

The online evaluation is a much faster and clear method to define all the relevant marking schemes.It brings much transparency to the present method of answer checking the answers to all the questions after the extraction would be stored in a database.The database is designed as such that it is very easily accessible.The work of checking hundreds of answer sheets which more or less contains the same answer can be a burden task.This system can be used instead in order to reduce their burden. It will save a lot of effort and time on teachers' part. The human efforts applied in this repetitive task can be saved and spent more in other academic endeavors.

The obvious human mistakes can be reduced to obtain an unbiased result. The system calculates the score and provides results fairly quickly. This system can be widely used in academic institutions such as schools, colleges, coaching and institutes for checking answer sheets . It can also be implemented in different organizations which conduct competitive examinations.

1.1 Prerequisites :

1.1.1 Software Requirements :

1. Python version 3.0
2. Python IDE - Jupyter
3. Data science libraries - Pandas, Numpy, Sklearn, Matplotlib, PIL

1.1.2 Hardware requirements:

1. CPU - 8 to 16 with each octa core processor in a distributed network.
2. RAM - 128 to 256 GB
3. Storage – 30 to 50 GB

1.2 Python:

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

It's often used as a “scripting language” for web applications. This means that it can automate a specific series of tasks, making it more efficient. Consequently, Python (and languages like it) is often used in software applications, pages within a web browser, the shells of operating systems and some games.

It is used for:

- Web development (server-side),
- Software development,
- Mathematics,
- System scripting.

Python does the things as follow:

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

Advantages of python are mentioned below:

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way or a functional way.

1.2.1 Machine learning in python :

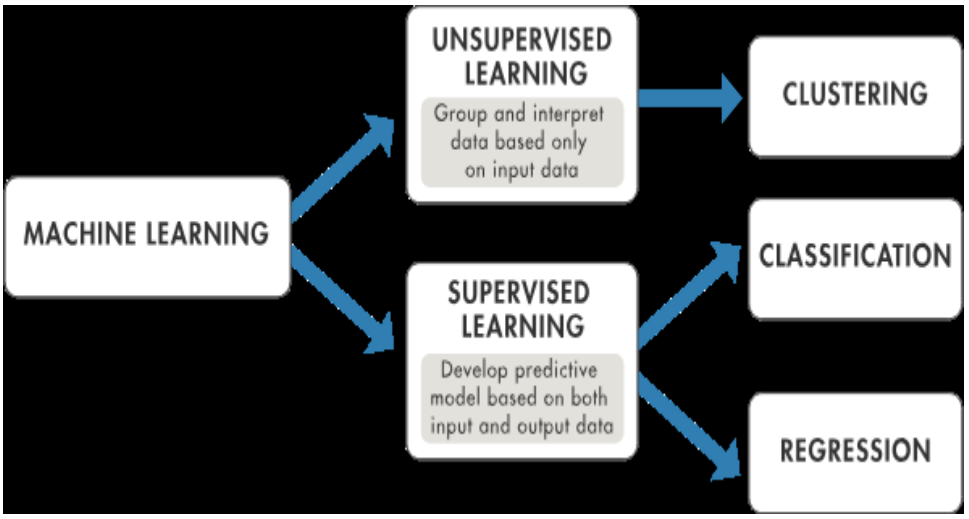

Fig 1.1 Machine learning overview

What is Machine Learning?

Machine learning is a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

In data science, an algorithm is a sequence of statistical processing steps. In machine learning, algorithms are 'trained' to find patterns and features in massive amounts of data in order to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions.

Machine learning is learning based on experience. As an example, it is like a person who learns to play chess through observation as others play. In this way, computers can be programmed through the provision of information in which they are trained, acquiring the ability to identify elements or their characteristics with high probability.

There are various stages of machine learning:

- Data collection
- Data storage
- Data analysis
- Algorithm development
- Checking algorithm generated
- The use of an algorithm to further conclusions

Machine learning algorithms are divided into two groups:

- Unsupervised learning
- Supervised learning

With Unsupervised learning, your machine receives only a set of input data. Thereafter, the machine is up to determine the relationship between the entered data and any other hypothetical data. Unlike supervised learning, where the machine is provided with some verification data for learning, independent Unsupervised learning implies that the computer itself will find patterns and relationships between different data sets. Unsupervised learning can be further divided into clustering and association. Supervised learning implies the computer ability to recognize elements based on the provided samples. The computer studies it and develops the ability to recognize new data based on this data. For example, you can train your computer to filter spam messages based on previously received information.

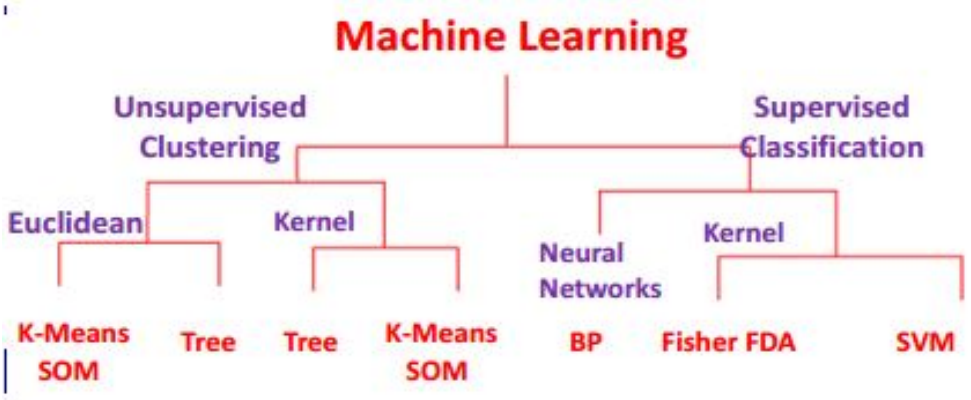

Fig 1.2.Types Of Machine Learning

Some Supervised learning algorithms include :

- Decision trees
- Support-vector machine
- Naive Bayes classifier
- k-nearest neighbours
- linear regression

Some Unsupervised learning algorithms include:

- K-means clustering
- Hierarchical clustering
- Anomaly detection
- Neural Networks
- Principal Component Analysis
- Independent Component Analysis
- Apriori algorithm

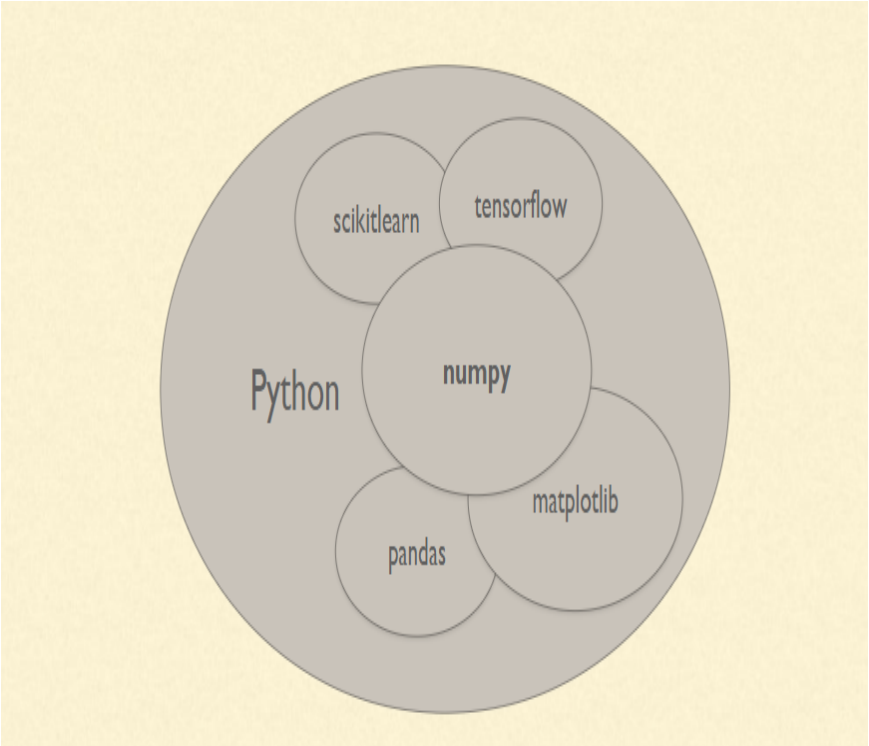

Fig 1.3 Numpy

1.2.1.1 Numpy :

NumPy is the fundamental package needed for scientific computing with Python.

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays.NumPy is an open-source numerical Python library.

Numpy array is a powerful N-dimensional array object which is in the form of rows and columns. We can initialize NumPy arrays from nested Python lists and access it elements.

Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called cleaning the data.

Pandas is a popular Python package for data science, and with good reason: it offers powerful, expressive and flexible data structures that make data manipulation and analysis easy, among many other things. The DataFrame is one of these structures. Those who are familiar with R know the data frame as a way to store data in rectangular grids that can easily be overviewed. Each row of these grids corresponds to measurements or values of an instance, while each column is a vector containing data for a specific variable. This means that a data frame's rows do not need to contain, but can contain, the same type of values: they can be numeric, character, logical, etc**.**

Now, DataFrames in Python are very similar: they come with the Pandas library, and they are defined as two-dimensional labeled data structures with columns of potentially different types.

In general, you could say that the Pandas Data Frame consists of three main components: the data, the index, and the columns.

Firstly, the DataFrame can contain data that is:

- A Pandas DataFrame
- A Pandas Series: a one-dimensional labeled array capable of holding any data type with axis labels or index. An example of a Series object is one column from a DataFrame.
- A NumPy ndarray, which can be a record or structured.
- A two-dimensional ndarray.
- Dictionaries of one-dimensional ndarray, lists, dictionaries or Series.

The difference between np.ndarray and np.array() .The former is an actual data type, while the latter is a function to make arrays from other data structures.

Structured arrays allow users to manipulate the data by named fields: in the example below, a structured array of three tuples is created. The first element of each tuple will be called foo and will be of type int, while the second element will be named bar and will be a float.

Record arrays, on the other hand, expand the properties of structured arrays. They allow users to access fields of structured arrays by attribute rather than by index. You see below that the foo values are accessed in the r2 record array.

1.2.1.3 SK-Learn :

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.