TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

A Minor Project Proposal

On

MACHINE LEARNING IMPLIED PERCEPTIVE THOUGHT READER

Submitted by:

ABHAY NEPAL 075BEI 003

DIPESH TRIPATHI 075BEI 013

GOKARNA ADHIKARI 075BEI 014

KSHITIZ DHAKAL 075BEI 015

Submitted to

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

LALITPUR, NEPAL

(December 12, 2021)

Abstract

AI, Artificial Intelligence is a rapidly evolving process in context of today's world. The fact that the machine can approach to the behavior of the human is the arguably the most interesting topic in present context.

The project also revolves around the idea of "machine approaching behavior of a human". We plan to create a *fully software based desktop application*, which uses *a suitable machine learning* technique to master psychology related task. In this project, primarily we aim to master our application *sentimental analysis*, besides we would like to incorporate the *personality prediction* and *handwriting font generation* features in our application as well.

The reasons for selecting this topic for our project is its versatility and the wide learning opportunity it will provide that includes multiple yet correlated machine-learning topics.

Table of contents

Introduction

Theoretical Background

Theoretical Background

Tools and technology

Expected Project

Introduction

The intended project is a fully software based desktop application named as "THREADER", which implements different machine learning techniques to perform human like behavior.

Background

Sentimental analysis analyses the textual data and processes the emotions hidden into them. Sentimental analysis is one of the rising and popular topics in the field of artificial intelligence. For a machine to approach the behavior of a human, it is mandatory that it can interpret the emotions of the statements expressed by a human. The topic is popular because it has a versatile application in various fields of industry.

Conclusively, analysis of sentiment deals with the people opinion about a particular topic or product. It is competent in understanding the people's opinion and provide a basis of evidence and reasoning on particular product.

Motivation

In present context, communication through texting via internet is becoming more and more popular. Vast majority of organizations use online feedback system to enhance the quality of service they provide. For this, the sentimental analysis is one of the vital aspect for knowing how they are performing from the perspective of their customers. A human could do the feedback analysis but the task seems very tedious and time consuming.

Therefore, a computer software that can analyze the sentiment of large number of feedbacks received by an organization seems a very good idea. A software that analyses the abundant textual information presented in the internet or other medium to provide future insight for the organization and aid the public to take decision on their purchase seems a very practical project. On the other hand, multinational big tech giant companies like Meta (Facebook), Alphabet (Google), Twitter, Amazon etc. are also investing huge resources on the research and development of a good sentimental analysis model.

Objectives

We plan to create a fully software based desktop application, which uses a suitable machine learning algorithm to master psychology related task. In this project, primarily we aim to master our application sentimental analysis, besides we would like to incorporate the personality prediction and handwriting font generation features in our application as well.

Problem Statement

For the best result, we will need a lot of data and great computational power. Since we are confined with limited resources in terms of computational power and data, the obtained model may not be perfect. However, the model will definitely approach the perfection relative to the data we will be able to collect. In case of personality prediction task, since a human behavior cannot be generalized as a whole, however the prediction would be based on the generalization of the human behavior that can lead to the wrong predictions on some cases. The handwriting font generation would be the task that would involve the image processing technique that needs a high computational power. The computational power of the available resources may not be sufficient to produce a good model.

Scope of project

Threader has versatile scope and serve numerous applications. The project can be used as a various useful tools.

i.   Support and feedback: It helps for the analysis of the product based on the sentimental analysis of the feedbacks done by the customers. It can assure a good and respectable place for the voice of both customers and employees.

ii.   Personality detector: It can help an organization review the employees, recognize a good candidate on an interview in case of new hiring etc.

iii. Recommender system

iv.   Social media monitoring

v.   Chat bot

Literature Review

Sentimental analysis analyses the textual data and processes the emotions hidden into them. For a machine to effectively analyze the emotions of a text written or stated by a human is onerous task. Hence, the accuracy parameters comes into play while developing the application.

Different papers had been published about the sentimental analysis in the past; three most relevant of them are discussed here.

i.   **Topic**: A Sentimental Education: Sentimental Analysis using subjectivity summarization n based on Minimum Cut.
     Author: BoPang and Lillian Lie [Cornell University]
     Description: This paper proposed that SVM and NB are better technique for improving the performance of a model up to 86.4 %.

ii.   **Topic**: Automatic Sentiment Analysis in Online Text
     Author: Erik.Boiy, Pieter Hens, Koen Dschacht and Marie Francine Moens

Description: This paper shows the varying level of accuracy when symbolic and machine-learning methods were applied to different social network dataset.

iii.   **Topic**: Sentiment Analysis in data of Twitter using Machine learning algorithm
         Author: Dr. Sefer Kurnaz, and Mustafa Ahmed Mahmood.
         Description: This paper proposes a new technique that offers an accuracy of 98 % when compared with Deep learning method, SVM and Maximum Entropy method.

After analyzing the works done on these fields, we found out that following algorithms are used in existing research works:

i.   Rule bases
ii.   Naïve Bayes
iii.   Support Vector Machine
iv.   Multilayer perception
v.   Maximum entropy
vi.   Decision Tree
vii.   Convolution neural network
viii.   Bayesian Network

Theoretical background

The project will incorporate a suitable machine learning algorithm for achieving the entire desired task.

Machine learning

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning.

Approaches

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system.

Supervised learning

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. Through iterative optimization of an objective function, supervised learning

algorithms learn a function that can be used to predict the output associated with new inputs. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task. Types of supervised learning algorithms include active learning, classification and regression.

Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function. However, unsupervised learning encompasses other domains involving summarizing and explaining data features.

Reinforcement learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov decision process (MDP). Many reinforcement-learning algorithms use dynamic programming techniques. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP, and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

Algorithms

Any machine learning algorithms can be analyzed based on their three major components.

Representation

Different methods can be incorporated to represent machine learning algorithms into various types:
i.   Support vector machines
ii.   Decision Trees
iii. Set of rules / Logic programs
iv.   Instances
v.   Graphical models (Bayes/Markov nets)

vi.   Neural networks

vii.   Model ensembles

Evaluation

A machine learning algorithm can be evaluated on basis of following parameters:

i.   Accuracy

ii.   Position and recall

iii.   Squared error

iv.   Likelihood

v.   Posterior probability

vi.   Cost / Utility

vii.   Margin

viii.   Entropy

ix.   K-L divergence

Optimization

There are mainly three optimization technique for machine learning algorithm:

i.   Combinational Optimization

ii.   Convex Optimization

iii.   Constrained optimization

Based on these three components we would select the best algorithm fit for our project.

Methodology

The software development approach, the system block diagram and overall system design process is discussed on this section.

Software development approach

We intend to develop the project as a desktop application GUI (Graphical User Interface) based on JAVA that incorporates different machine learning techniques, written in PYTHON to achieve entire desired tasks.

The overall software development process can be stated as:

i.   Data collection
  - Data collection will be done through survey and some data could be extracted from internet

ii.   Training and testing
  - The best suitable machine learning algorithm is chosen
  - The obtained data is trained and tested for model development

iii.   Model development
  - Developing machine learning model based in train/test result

iv.   GUI development
  - Graphical user interface development as a desktop application

v.   Integrating models and the GUI
   •   Integrating the developed model with the GUI
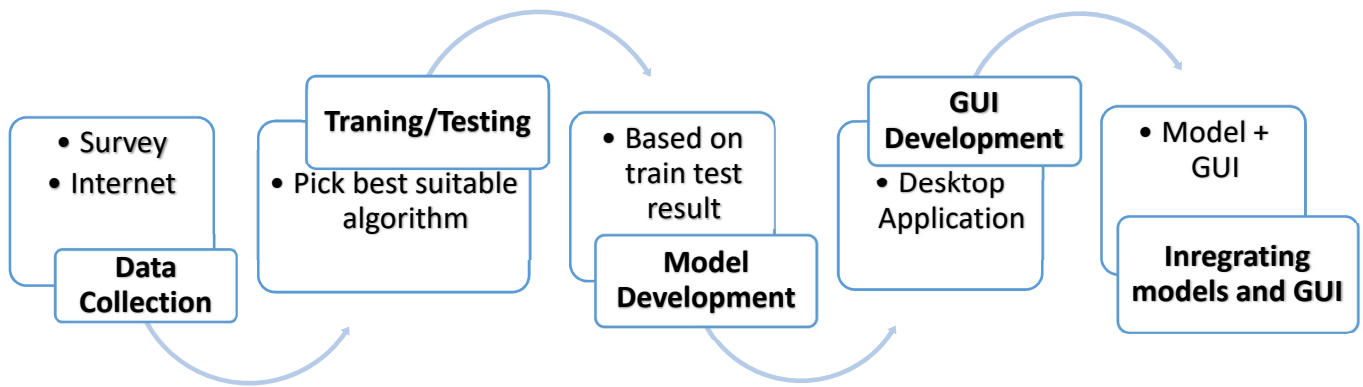
Chart 1: Software development approach
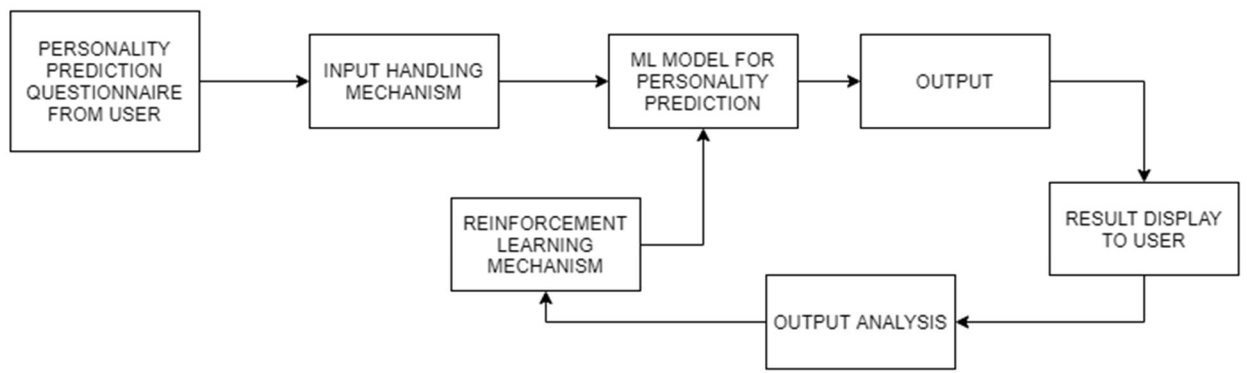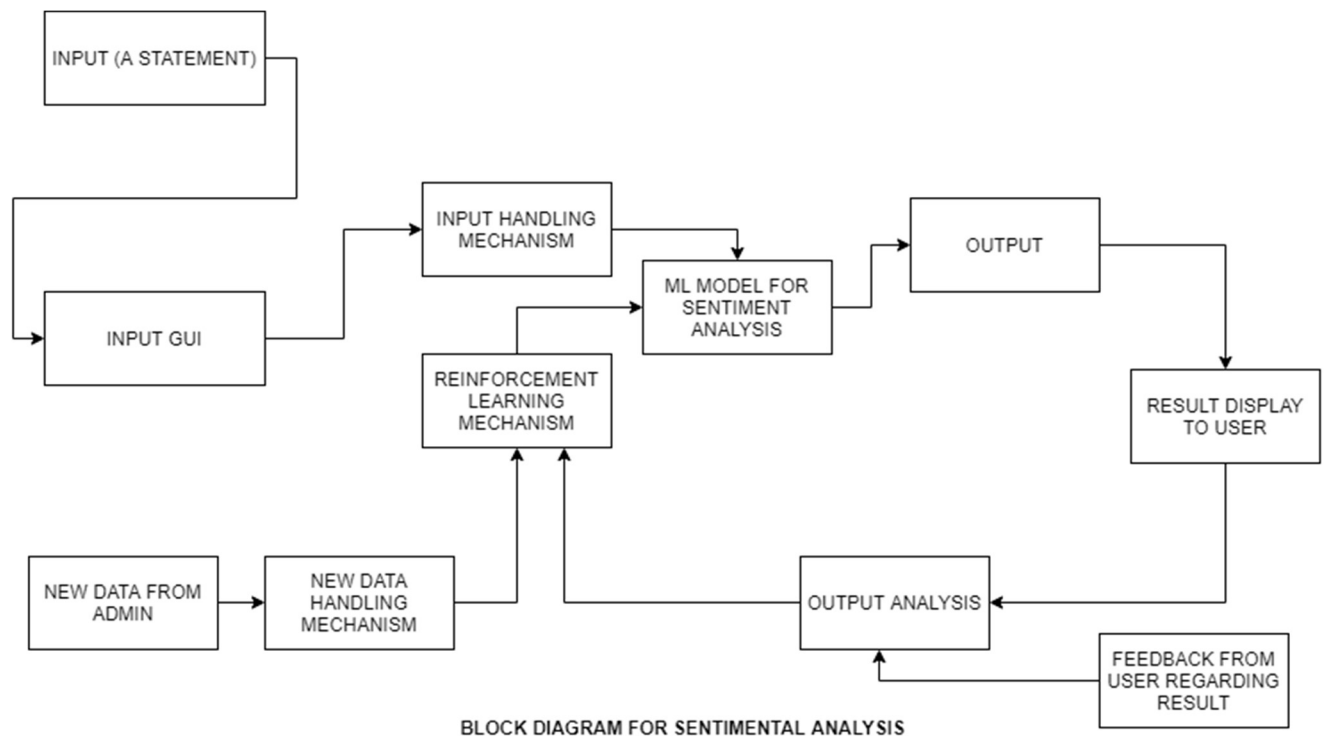
System block diagram
The project mainly focuses on the sentimental analysis and it incorporates other features like personality prediction and handwriting font generation.

The comprehensive system block diagram of all the features are shown below. For the development of the respective ML model (Machine Learning model), the best suitable algorithm will be chosen.

BLOCK DIAGRAM FOR SENTIMENTAL ANALYSIS

```
INPUT (A STATEMENT)
        |
        v
INPUT GUI ──> INPUT HANDLING MECHANISM ──> ML MODEL FOR SENTIMENT ANALYSIS ──> OUTPUT ──> RESULT DISPLAY TO USER
                                              ^
                        REINFORCEMENT LEARNING MECHANISM
                                              ^
NEW DATA FROM ADMIN ──> NEW DATA HANDLING MECHANISM
                                              
                                    OUTPUT ANALYSIS <── 
                                              ^
                                    FEEDBACK FROM USER REGARDING RESULT
```

BLOCK DIAGRAM FOR SENTIMENTAL ANALYSIS

BLOCK DIAGRAM FOR PERSONALITY PREDICTION

```
PERSONALITY PREDICTION QUESTIONNAIRE FROM USER ──> INPUT HANDLING MECHANISM ──> ML MODEL FOR PERSONALITY PREDICTION ──> OUTPUT ──> RESULT DISPLAY TO USER
                                                                                          ^
                                                REINFORCEMENT LEARNING MECHANISM
                                                                      <── OUTPUT ANALYSIS <──
```

BLOCK DIAGRAM FOR PERSONALITY PREDICTION

BLOCK DIAGRAM FOR HANDWRITING FONT GENERATION

```
VALID HANDWRITING INPUT FROM USER ──> INPUT HANDLING MECHANISM ──> ML MODEL FOR FONT GENERATION [image processing] ──> OUTPUT
```

BLOCK DIAGRAM FOR HANDWRITING FONT GENERATION

Predicted timeline

We aim to complete our project in the span of 10 weeks with the predicted timeline as follows.

2 weeks	• Data collection
4 weeks	• ML Model development
2 weeks	• GUI development
1 week	• GUI and ML Model integration
1 week	• Testing and review

Table 1: Project timeline prediction

Tools and technology

Since the project is completely software based, the requirement is also fully software based. The programming language and their frameworks we will be using in our projects are described below.

Java

Java will be used for the development of the desktop application to provide an effective graphical user interface, and integrating with python. Following libraries or frameworks would be necessary.

Swing

Swing is a GUI widget toolkit for Java that will be used to develop the desktop application.

Jython

Jython is an implementation of the Python programming language designed to run on the Java platform.

Python

Python will be used to implement all the machine learning algorithms and machine learning model development. Following libraries would be necessary.

Scikit learn

Scikit-learn is a free software machine-learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines.

NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Tensor Flow

Tensor Flow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Anticipated project

We believe that we can complete our project on time and become successful to develop an effective and bug free desktop application with attractive GUI using JAVA and successfully integrate it with the machine learning models developed using PYTHON in the span of about 10 weeks. It will be able to analyze the sentiments of the textual statement, predict the personality of the user based on some questionnaire and generate a handwriting font based on the input done by the user.

Budget estimation

Since we will be using all open source software and libraries, the budget of our project will not be much; it is expected to be less than Rs. 10,000 excluding the computers/laptop we are going to use unless we plan to incorporate it in cloud based web application in future.

Particulars	Description	Cost
Computer/Laptop	Hardware setup for programming	Rs. 150,000
Visualization tools, APIs		Rs. 5,000
Reports and documentation		Rs. 3,000
	Grand Total	Rs. 158,000

Table 2: Project Budget estimate

�✱✱✱✱✱✱

REFERENCES

i.   *https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5107*
ii.   *https://monkeylearn.com/blog/sentiment-analysis-applications/*
iii.   *http://www.ijstr.org/*
iv.   *https://en.wikipedia.org/*
v.   *https://www.mrdbourke.com/2020-machine-learning-roadmap/*
vi.   *https://www.youtube.com/watch?v=KcVozMo--M4*
vii.   *Research papers:*
   a.   *https://www.cs.cornell.edu/home/llee/papers/cutsent.pdf*
   b.   *http://www.ijstr.org/final-print/apr2020/Literature-Review-On-Sentiment-Analysis.pdf*
   c.   *https://ieeexplore.ieee.org/document/6897213*
   d.   *E. Kouloumpis, T. Wilson, J. Moore, Twitter sentiment analysis: The good the bad and the omg! , Proc. 5th Int. AAAI Conf. Weblogs Social Media, pp. 538-541, 2011.*
   e.   *H. Saif, Y. He, H. Alani, Alleviating data sparsity for Twitter sentiment analysis, Proc. CEUR Workshop, pp. 2-9, Sep. 2012*
   f.   *Nupur Kalra, Deepak Yadav, Gourav Bathla. (2019), SynRec: A prediction technique using collaborative filtering and Synergy Score*