FIT3036 FINAL REPORT

THE PREDICTION OF SUICIDE IN FACEBOOK DATA USING MACHINE LEARNING ALGORITHMS

CHUN HIN CHAN
24727164

Contents

1. Abstract

The rise of suicide rate in Australia over the past 10 plus year has led me to study the change of their emotion in order to identify the hidden message from the suicide high risk people. Since suicide is a serious mental health problem and the rise of suicide rate is not a good social phenomenon, prevent it is a most important task. In this paper, I propose a suicide related words detection system, for predicting the suicidal acts by using Facebook data. I use a tool of data mining to extract the Facebook post comments for text classification based on machine learning classification algorithms. The test results show that the method for detecting the suicidal acts using Facebook data and the machine learning algorithms verify the effectiveness of performance in term of accuracy and error rate.

2. Keywords

Facebook; machine learning; suicide; effectiveness

3. Introduction

The project objective is to make a platform that involves two basic functions: the first basic function is provide testing of several machine learning algorithms with the Facebook dataset in order to find out the most appropriate machine learning algorithm for Facebook to determine whether the suicide keywords exist within user messages or not, and the second basic function is provide a checking function to check whether the message entered by user consist any suicide related keywords.

The project requirements can be separated into two types: functional requirements and non-functional requirements. Functional requirements include: the system can extract data set from Facebook, the system can check all the message in the data set by using different machine learning algorithm, the system can return diagrams of the machine learning algorithm performance testing, the system can accept the user to enter a message within the message box for message checking function, and the system must return a message dialog (either warning message or safe message) after the user press the check button. Non-functional requirements include: the response time of the system should within 5 minutes, and the accuracy of the machine learning algorithm must above 70%.

As part of the project, there are several constraints that faced during the project. First of all, it is hard to extract enough Facebook posts with suicide related keywords because the number of people that tends to suicide is still less than normal people. Secondly, the extracted Facebook dataset do not have enough data attributes, which we only need the message attributes, so that the Decision Tree accuracy is not accurate at all.

4. Background

In the recent years, suicide problems become normalized in Australia. According to the statistics from Lifeline, the overall suicide rate in Australia in 2015 is the highest suicide rate in 10-plus years, which was 12.6 people per 100,000 people. This is not a good social phenomenon. In order to prevent this situation, people's emotions are the most important signal for us to concern about. Therefore, social communication media can help us to understand their current emotion as these are the most popular platforms that they used to express their feelings, and then to make a prediction on their probability to suicide. Since there already had a research on Twitter about suicide prediction researched by Birjali, Beni-Hssane and Erritali in 2016, therefore another famous social communication media – Facebook has been selected as the target social communication media in the project.

4.1. Big Data

Big data refers to large amount of both structured and unstructured data that traditional data processing application software is unable to deal with them. Big data can be characterized by the three Vs, which are volume, velocity and variety. Volume refers to the amount of data, velocity refers to the speed of data processing and variety refers to the number of types of data. Besides the definition of big data, Facebook also make use of its big data. According to Monnappa, A. (2015), Facebook's usage of big data can be divided into four parts: tracking cookies, facial recognition, tag suggestions and analyzing the Likes. Therefore, as the purpose of the project is focus on prediction, therefore, the most suitable big data technologies to be used in the project to predict the suicide problems in Facebook should be predictive analytics.

4.2. Machine Learning Algorithm

Machine learning is an automated learning of a concept given some examples of data. And in the project, the target machine learning algorithms are Decision Tree and Naive Bayes, therefore, the target classifiers are Decision Tree classifier and Naïve Bayes classifier.

Decision trees are one of the most widely used and practical methods in machine learning, which use existing data attributes and values to classify new instances or profile existing data. According to the online lecture *Decision Tree Classifier* (2010), Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time times it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

Naive Bayes also are one of the most widely used and practical methods in machine learning. According to Ray S. (2017), Naive Bayes algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors and Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

4.3. Dataset Collection

As part of the project, the dataset used currently is one of the main components of the project. In order to extract the dataset from Facebook, I get the Facebook API key from the Facebook App Developer at first, and then use it into the FacePager to extract the target Facebook page comments as a csv file.

4.4. Project Information

For the project information, the main project risks are: 1) cannot extract dataset from Facebook by python; 2) taking too long on tasks of coding the machine learning algorithms; 3) computer occurs technical problems. According to the priority of the main project risks, their probabilities are medium, high and rare. And the corresponding risk reduction strategies are: 1) find another tool to extract dataset from Facebook, like R studio, FacePager; 2) reduce the time of task on testing and coding the other parts; 3) continue the project in another computer. However, the risks that actually encountered only included the first one of them, and the other risks that actually encountered are the lack of Facebook comments data with suicide related words and the time usage is over on learning the machine learning algorithms. Therefore, during the project, the corresponding risk reduction strategies are: 1) using FacePager to extract Facebook comments data; 2) enlarger the scale of required data by resampling the data; 3) reduce the number of machine learning algorithms for the project.

4.5. Resource Requirements

The hardware requirements of the project is that a desktop computer or a notebook computer is required, and the software requirements of the project are JetBrains PyCharm Community, Python, Python Graphic User Interface, Python Facebook API, Python tkinter package, Python pandas package, Python numpy package, Python matplotlib.pyplot package, Python scikit-learn package, Python textblob package, Facebook App Developer, FacePager.

4.6. Tasks and Timeline

Task	duration	Start Date	End Date
Write the project proposal	9 days	24-Aug-2017	1-Sep-2017
Write the function to get the list of suicide related keywords	1 day	5-Sep-2017	5-Sep-2017
Write the function to get user input message	1 day	5-Sep-2017	5-Sep-2017
Write the function to return a message dialog	1 day	5-Sep-2017	5-Sep-2017
Write the function to get messages from the dataset extracted from Facebook	4 days	7-Sep-2017	11-Sep-2017
Write the function for target machine algorithm – Naïve Bayes	5 days	11-Sep-2017	16-Sep-2017
Write the function for target machine algorithm – Decision Tree	5 days	16-Sep-2017	21-Sep-2017
Write the function to plot graph	1 day	21-Sep-2017	21-Sep-2017
Unit Testing	5 days	25-Sep-2017	30-Sep-2017
System Testing	5 days	30-Sep-2017	5-Oct-2017
Deal with the problems / justify the coding	3 days	5-Oct-2017	8-Oct-2017
Write the test report	11 days	9-Oct-2017	20-Oct-2017
Write the final report	11 days	9-Oct-2017	20-Oct-2017

5. Method

5.1. Methodology

The project can be separated into several component parts: the construction of the suicide related keywords; extracting the posts and comments from Facebook; the automatic classification by using different machine learning algorithms; and the analysis of the result.

With these component parts, the process of method ology can be defined as the following steps: train the machine learning algorithms by using the Facebook posts training data at first. Then, put the Facebook posts testing data and the trained machine learning algorithms into the classifier together. Finally, the prediction comes out.

Before starting the project system testing, the suicide related keywords should be selected out and grouped as a group. And the finalized suicide related keywords are "suicidal; suicide; kill myself; my suicide note; my suicide letter; end my life; never wake up; can't go on; not worth living; ready to jump; sleep forever; want to die; be dead; better off without me; better off dead; suicide plan; suicide pact; tired of living; don't want to be here; die alone; go to sleep forever".

5.2. Internal Design

The overall internal design is that user run the software and it will give out an interface with three options, which are option 1 for determining which machine learning algorithm is the best for Facebook to classify the message with suicide related keywords; option 2 for determining whether the message entered by the user contains suicide related keywords or not; option 3 for exiting the program. If the user selected option 1, the program will request for the Facebook data set and then build up the machine learning algorithm models. After the models built up, the accuracy and the error rate of each of the models will be calculated in order to plot the graph to show the result. If the user selected option 2, the program will request the user to enter a message and then determine whether the message included any suicide related keywords. The program will return a warning message if suicide related keywords exist or return a safe message if suicide related keywords are not exist. Once the user selected option 3, the program will stop running.

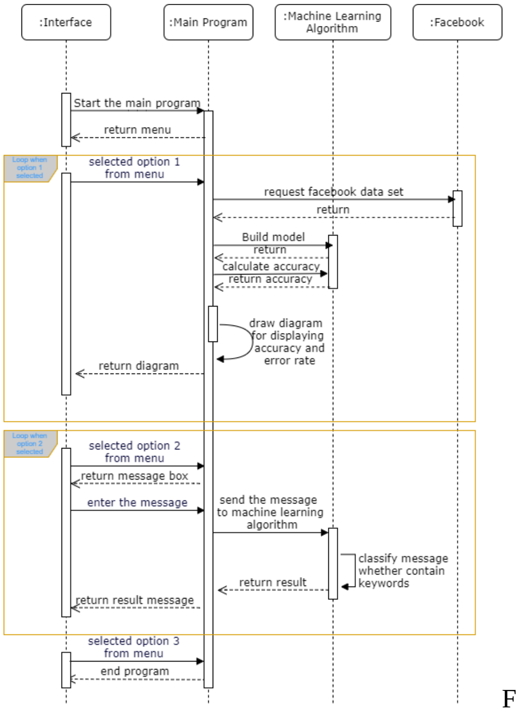

:Interface
:Main Program
:Machine Learning Algorithm
:Facebook
Start the main program
return menu
Loop when option 1 selected
selected option 1 from menu
request facebook data set
return
Build model
return
calculate accuracy
return accuracy
draw diagram for displaying accuracy and error rate
return diagram
Loop when option 2 selected
selected option 2 from menu
return message box
enter the message
send the message to machine learning algorithm
classify message whether contain keywords
return result
return result message
selected option 3 from menu
end program

Figure 1: Sequence Diagram

5.3. Software Architecture

The overall structure of the code included:

- Five class: UI class, Menu class, PageOne class, PageTwo class, PageThree class
- One public function: accuracy_calculate() for calculate the accuracy of different machining learning algorithm models' text classifier
- Two private functions in Menu class:
    - start(): to start the program
    - exit(): to end the GUI
- Three private functions in PageOne class:
    - accuracy_performance() : go to PageTwo class
    - message_check(): go to PageThree class
    - back(): return to the pervious page
- Two private functions in PageTwo class:
    - plotgraph(): to plot the graph
    - back(): return to the pervious page
- Three private functions in PageThree class:
    - check_message(): to check the message input
    - msg_window(): to pop out a dialog window
    - back(): return to the pervious page

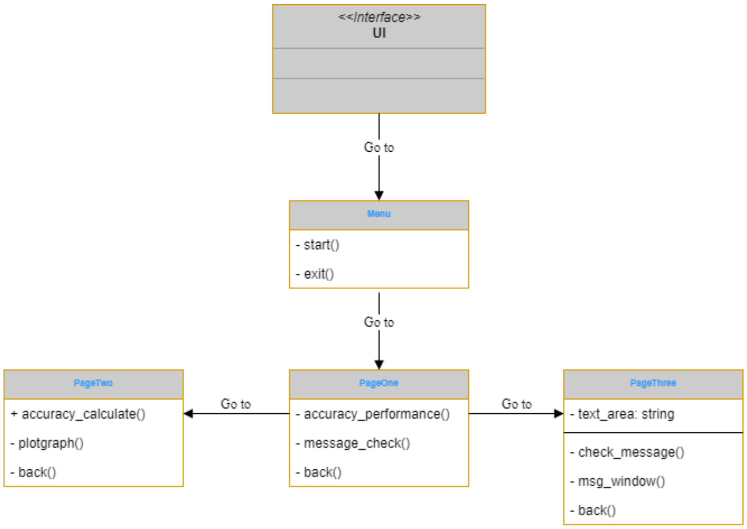


Figure 2: Class Diagram

5.4. Key Algorithms

There are two main key algorithms used in the project: accuracy performance and message check.

The function of the first key algorithm – accuracy performance – is to read in the extracted Facebook data csv file row by row, classify the existence of the suicide related keywords and store the result in an array in form of (message, existence). Then, split the array into training dataset (80%) and testing dataset (20%) and use then to build up the models of the machine learning algorithms. After that, calculate the accuracy and error rate of every machining learning algorithm models and plot a performance graph as the result. Below is the pseudocode of the accuracy performance:

```
# Using the python package textblob, numpy, pandas, matplotlib.pyplot
Read in the Facebook data
For each row of the Facebook data:
        classify the existence of the suicide related keywords
        store in an array in form of (message, existence)
```

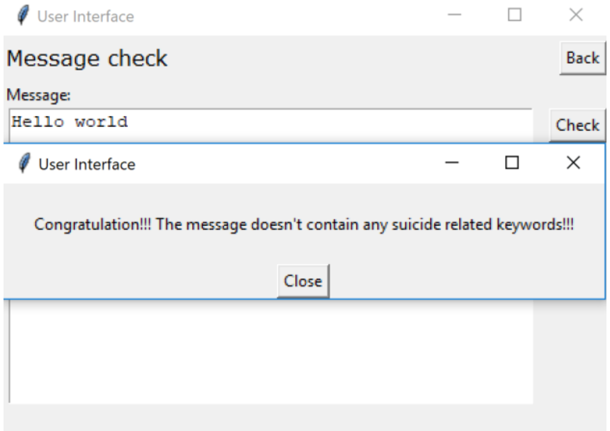
```
Split the array into training dataset (80%) and testing dataset (20%)
Build up the model of the machine learning algorithms
Calculate the accuracy of each models
Calculate the error rate of each models
Set the name of the models as the x-axis data
Set the accuracy and error rate as the y-axis data
Set the graph title, x-axis label and y-axis label
Plot the performance graph
```

The function of the second key algorithm – message check – is to read in the user input message, followed by tokenizing the message, and then comparing to the suicide related keywords group one by one to check whether the user input message include suicide related keywords or not. At last, return a dialog window with a message to notify the result. An error message means the user input message is empty. A congratulation message means that the suicide related keywords do not exists in the user input message. A warning message means that the suicide related keywords exists in the user input message. Below is the pseudocode of the message check:

```
# Using the python package sklearn.feature_extraction.text
Get the user input message from the text area
Tokenize the user input message and store as an array
Set counter i = 0, variable found = False
While the counter i is smaller than length of the array
        Check the array[i] is inside the suicide related keywords group or not
        If found, the variable found change to True
Return the result as the message of the dialog window
        Error message if the user input message is empty
        Congratulation message if variable found is False
        Warning message if variable found is True
```

6. Results

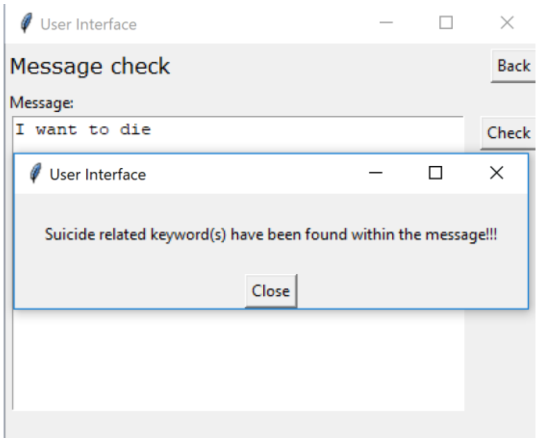
The expected outcomes of the project are: 1) extracting the keyword related to suicide from the Facebook message and classify the keyword by machine learning algorithms; 2) return message dialog window; 3) generate a graph on accuracy and error rate comparison of different machine learning classifiers. After running the project system, there are two main results. The first one is that a bar chart about the accuacies and error rates of all the machine learning classification models, and the other one is that a doalog window with different messages (congratulation message or warning message or error message).



Hello world

Congratulation!!! The message doesn't contain any suicide related keywords!!!

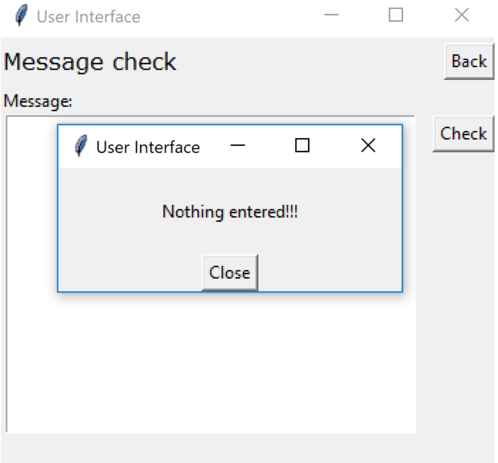
Figure 3: Congratulation message dialog window



I want to die

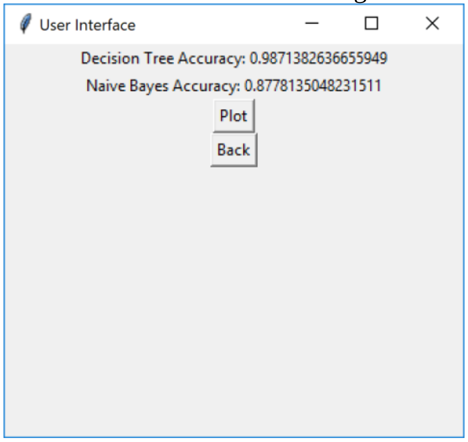
Suicide related keyword(s) have been found within the message!!!

Figure 4: Warning message dialog window



Nothing entered!!!

Figure 5: Error message dialog window



Decision Tree Accuracy: 0.9871382636655949

Naive Bayes Accuracy: 0.8778135048231511

Figure 6: Accuracy of every models

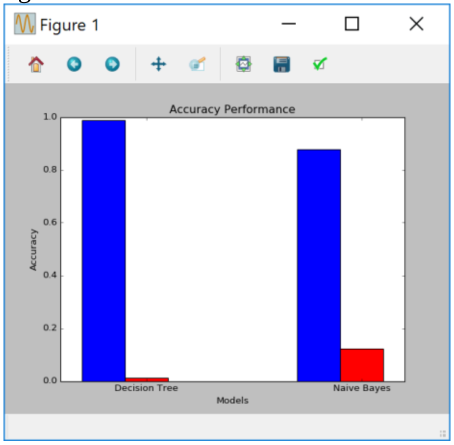


Figure 7: Performance graph of every models

7. Analysis & Discussion

This figure presents the statistics of the classification of the Facebook posts data, which are 15% of Facebook posts with suicide related keywords and 85% of Facebook posts without suicide related keywords. By figure 8, the limitation of the project also show out, which is hard to extract enough Facebook posts with suicide related keywords due to the overall number of people that tends to suicide is still less than normal people.

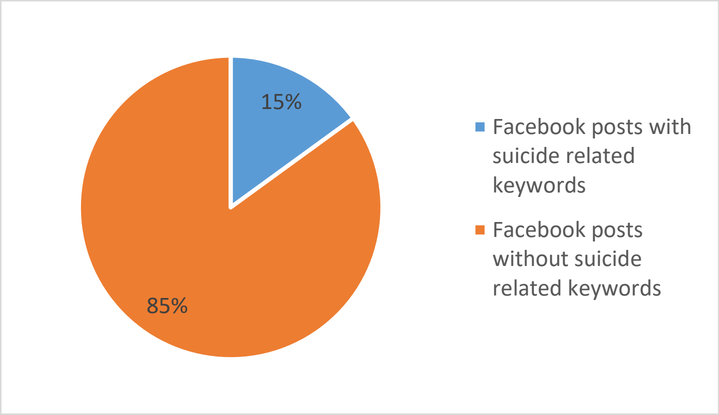


Figure 8: classification of the Facebook posts

As the project objectives mentioned above, the comparison of different machining learning algorithms is the main target of the project. After reading in the extracted Facebook data into the project system, different machining learning algorithm models work with the data and return the performance result as a bar chart. As the result shown in the figure 7, the accuracy of the Decision Tree model classifier is 98.7%, the error rate of the Decision Tree model classifier is 1.3%, the accuracy of the Naïve Bayes model classifier is 87.8% and the error rate of the Naïve Bayes model classifier is 12.2%. Although the accuracy of the Decision Tree model classifier is much higher than the accuracy of the Naïve Bayes model classifier, the best option for the classifier still be Naïve Bayes. The reason is that the number of data attributes is not enough to provide information to make an accurate Decision Tree. Therefore, the extremely high accuracy Decision Tree model relatively not reliable.

8. Future work

Consider as the future works, I plan to compare more different machine learning algorithms, which I have to understand the machine learning algorithm at first. Also, improvement of the current techniques is considering in order to enhance the accuracy of the current methods.

9. Conclusion

In conclude, the project system show that using different machine learning on the social communication media Facebook can be a preventive force in the fight against suicide.

(Word counts: 3009)

10.Bibliography

- Birjali, M., Beni-Hssane, A., and Erritali, M. (2016). Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms. International Arab Conference on Information Technology (ACIT'2016).
- Brownlee, J. (2016). *Spot-Check Classification Machine Learning Algorithms in Python with scikit-learn.* Retrieved from https://machinelearningmastery.com/spot-check-classification-machine-learning-algorithms-python-scikit-learn/
- *Decision Tree Classifier*. (2010). Retrieved from http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html
- Loria, S. (2017). *Tutorial: Building a Text Classification System.* Retrieved from http://textblob.readthedocs.io/en/dev/classifiers.html
- Monnappa, A. (2015). *How Facebook is Using Big Data - The Good, the Bad, and the Ugly. Retrieved* from https://www.simplilearn.com/unlocking-data-science-webinar
- Press G. (2016). *Top 10 Hot Big Data Technologies.* Retrieved from https://www.forbes.com/sites/gilpress/2016/03/14/top-10-hot-big-data-technologies/#748b0a5865d7
- Rouse, M. (2013). *What is 3Vs (volume, variety and velocity)? - Definition from WhatIs.com.* Retrieved from http://whatis.techtarget.com/definition/3Vs
- Ray, S. (2017). *6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)*. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
- Sherlock, D. (2014). *Extract Facebook Data and save as CSV.* Retrieved from https://www.youtube.com/watch?v=S9kYApoR8U4

11.Appendices

All the materials (coding + presentation + workbook + test report + final report) can be accessed and downloaded from my GitHub https://github.com/chcha49/Computer-Science-Project-Assignment-2

The presentation materials can also be accessed and downloaded from http://moodle.vle.monash.edu/mod/assign/view.php?id=4151156

The workbook can also be accessed and downloaded from http://moodle.vle.monash.edu/mod/turnitintooltwo/view.php?id=4258079

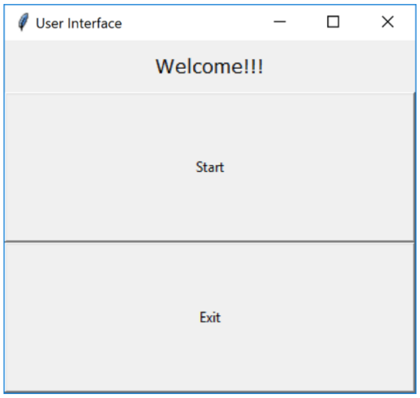
The test report can also be accessed and downloaded from http://moodle.vle.monash.edu/mod/turnitintooltwo/view.php?id=4258063

11.1. Production and Deployment

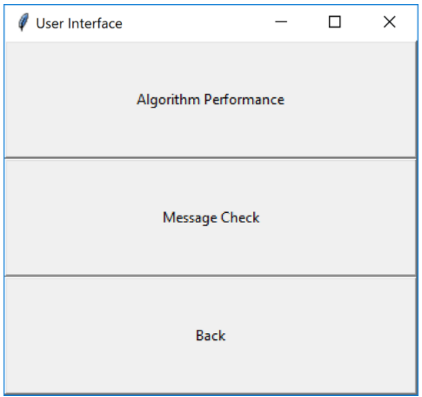
To run the project software, make sure you are using JetBrains PyCharm Community Edition 2017.2.3 and the Python environment is the same as my version. My version of Python is 3.5.2; the version of scipy is 0.19.1; the version of numpy is 1.13.1; the version of matplotlib is 1.5.1; the version of pandas is 0.18.1; the version of sklearn is 0.19.0; and the version of textblob is 0.13.0. You can check the version by using the version.py.

```
C:\Users\Derek\Anaconda3\python.exe C:/Users/Derek/PycharmProjects/fit3036/version.py
Python: 3.5.2 |Anaconda 4.1.1 (64-bit)| (default, Jul  5 2016, 11:41:13) [MSC v.1900 64 bit (AMD64)]
scipy: 0.19.1
numpy: 1.13.1
matplotlib: 1.5.1
pandas: 0.18.1
sklearn: 0.19.0
textblob: 0.13.0
```

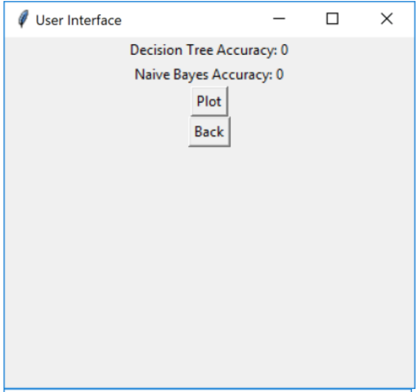
11.2. User Interface



The first user interface is the Welcome page with two buttons: start and exit. Button “start” uses to start the program and button “exit” uses to end the program.

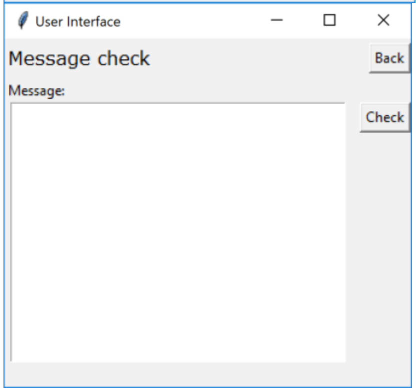


The second user interface is the Menu page with three buttons: algorithm performance, message check and back.
Button “Algorithm Performance” uses to go to the algorithm performance page. Button “Message Check” uses to go to the message check page. Button “Back” uses to return to the pervious page.



Decision Tree Accuracy: 0
Naive Bayes Accuracy: 0

The third user interface is the Algorithm Performance page with two buttons. Inside this page, the accuracies of all the models will be shown and the button “Plot” can be used to plot a bar chart about the accuracies (shown as blue bar) and the error rates (shown as red bar) of every models. (The bar chart is the one that shown in Figure 7.) Button “Back” uses to return to the pervious page.



The fourth user interface is the Message Check page with one text area and two buttons. The text area “Message” is for the user to input a message. Button “Check” uses to check whether the user input message include suicide related keywords with the better classifier in the option one by passing the user input message into the algorithm. Button “Back” uses to return to the pervious page.