VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM 590014

Artificial Intelligence self-study Report on

"Sentiment Analyser"

By

Siddhant Sud (1BM17CS102)

Under the Guidance of

Asha G R
Assistant Professor, Department of CSE
BMS College of Engineering

Artificial Intelligence self-study carried out at

Department of Computer Science and Engineering
BMS College of Engineering
(Autonomous college under VTU)
P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019
2019-2020

BMS COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

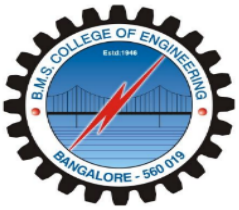
CERTIFICATE

This is to certify that the Artificial Intelligence self-study titled "Sentiment Analyser" has been carried out by Siddhant Sud (1BM17CS102), during the academic year 2019-2020.

Signature of the guide
Asha G R
Assistant Professor
Department of Computer Science and Engineering
BMS College of Engineering, Bangalore

BMS COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



DECALARATION

We, Siddhant Sud (1BM17CS102) students of 6th Semester, B.E, Department of Computer Science and Engineering, BMS College of Engineering, Bangalore, hereby declare that, this Artificial Intelligence self-study work entitled "Sentiment Analyser" has been carried out by us under the guidance of Asha G R, Assistant Professor, Department of CSE, BMS College of Engineering, Bangalore during the academic semester Jan-May 2020

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Siddhant Sud (1BM17CS102)

INTRODUCTION

Sentiment analysis (SA) is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. SA is used to analysis public opinion about certain topics, ranging from elections to brand pricing plans and brand identities..SA can be used to review public social media accounts, this would help in the personality classification of social media accounts.Sorting Data in huge scales. Concealed information which comes as of the social media in the sort of un-structured data can be structured.[12]. As imperative resources of real-time opinion, Twitter, texts and the other social networks have fascinated substantial interests of the research industry and Consistent criteria. It's estimated that people only agree around 60-65% of the time when determining the sentiment of a particular text. Tagging text by sentiment is highly subjective, influenced by personal experiences, thoughts, and beliefs. By using a centralized sentiment analysis system, companies can apply the same criteria to all of their data, helping them improve accuracy and gain better insights.

Description:

An Artificial Intelligence model is implemented in three steps:

1. Pre processing
2. Training
3. Evaluation

The first step is to collect, clean and organise the data. The second step is related to use the cleaned data sets to train the model through examples. By doing that, we generate a classifier model. This model contains a predict function which can receive any new commentary and classify it as a positive or negative. The last step is to validate if the model is good enough to solve the purposed problem. There's no step by step validation, because it depends entirely on the purposed problem.

Objective:

- To identify whether a given input sentence is positive or negative in sentiment
- To understand and implement Naive Bayes Algorithm in AI
- To visualize confusion matrix

Outcomes of the project:

- To create an accurate model to analyze the sentiment for English sentences
- We have gone through the research paper: Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues
- To create a tool to instantly identify a given sentence as positive or negative

Tools used:

Programming language:- python

Framework**:**

- **NumPy** is a package in Python used for Scientific Computing. NumPy package is used to perform different operations. The ndarray (NumPy Array) is a multidimensional array used to store values of same datatype. These arrays are indexed just like Sequences, starts with zero.
- **Matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter
- **Scikit-learn** is a free machine learning library for **Python**. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports **Python** numerical and scientific libraries like NumPy and SciPy.
- **pandas** is a **Python package** providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive.

Detailed Design:

As mentioned above the program can be divided into three parts.We will now go in-depth of the parts.

Data Gathering

- **Access Data :** There are three source files, amazon dataset, IMDB data set, yelp dataset. we extract the dataset and put it into a directory called “data”
- **Preprocessing:** We covert the directory into a nested list, which each outer list item containing an inner list with the text data and the text rating.

```
[['A very, very, very slow-moving, aimless movie about a distressed, drifting young man.  ',
  '0'],
 ['Not sure who was more lost - the flat characters or the audience, nearly half of whom walked out
  '0'],
 ['Attempting artiness with black & white and clever camera angles, the movie disappointed - became
```

- **Splitting the data:** We split the data set into a training set and an evaluation set.

Training Data

- **Splitting the text and value :** We split the data into two lists.one contains all the text and the

data	[[\'A very, very, very slow...\'t be going back.", \'0\']]
training_result	['0', '0', '0', '0', '1', '...0', '0', '1', '1', '1', '0']
training_text	[\'A very, very, very slow-...ss I won\'t be going back."]
vectorizer	CountVectorizer(analyzer='w...nizer=None, vocabulary=None)

second contains the rating.

- **Count Vectorisation :**Creating vectors that have a dimensionality equal to the size of our vocabulary, and if the text data features that vocab word, we put a '1' in that dimension. Every time we encounter that word again, we increase the count, leaving 0s everywhere we did not find the word even once**.**

$$\textbf{Score} = \frac{\textbf{How many times the word appears in a positive sentence}}{\textbf{How many time it appears at all}}$$

This is done by using the .**tranform ()** function

- **Naive Baues Algorithm :** Bayes' Theorem provides a way that we can calculate the probability of a hypothesis given our prior knowledge.Bayes' Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d)$$

Where

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis).

This is done by implementing the **BernoulliNB.fit()** function.

Classification

The BernoulliNB (NB for Naive Bayes) will generate our classification model. Now we can use the predict function of this model to try to predict the sentiment of any text.

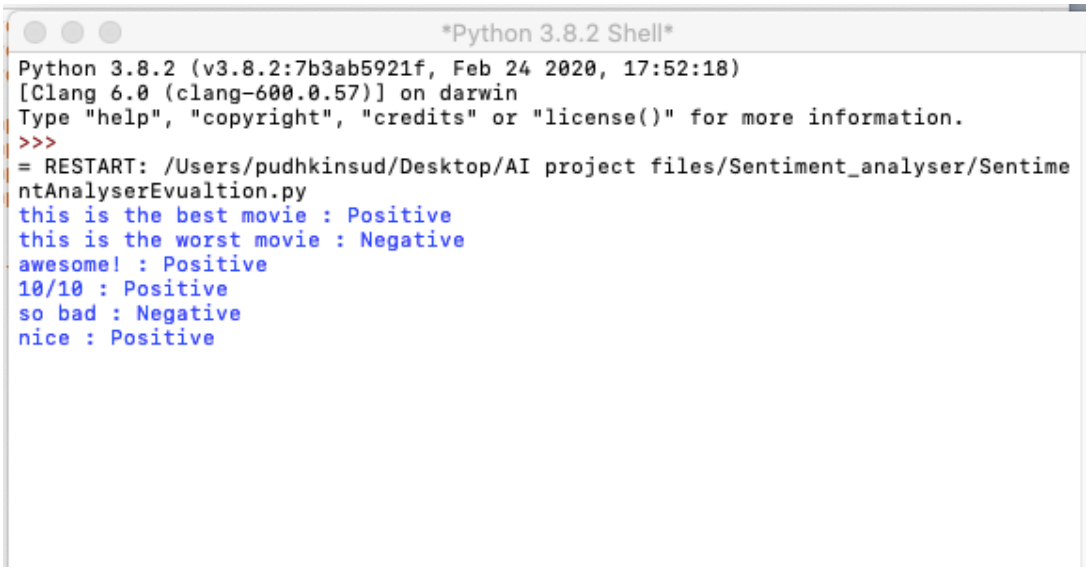
Evaluation

We check our predicted answer against the actual answer counting the right answers from the evaluation dataset.A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualisation of the performance of an algorithm.It allows easy identification of confusion between classes.

The matrix will contains four different values:

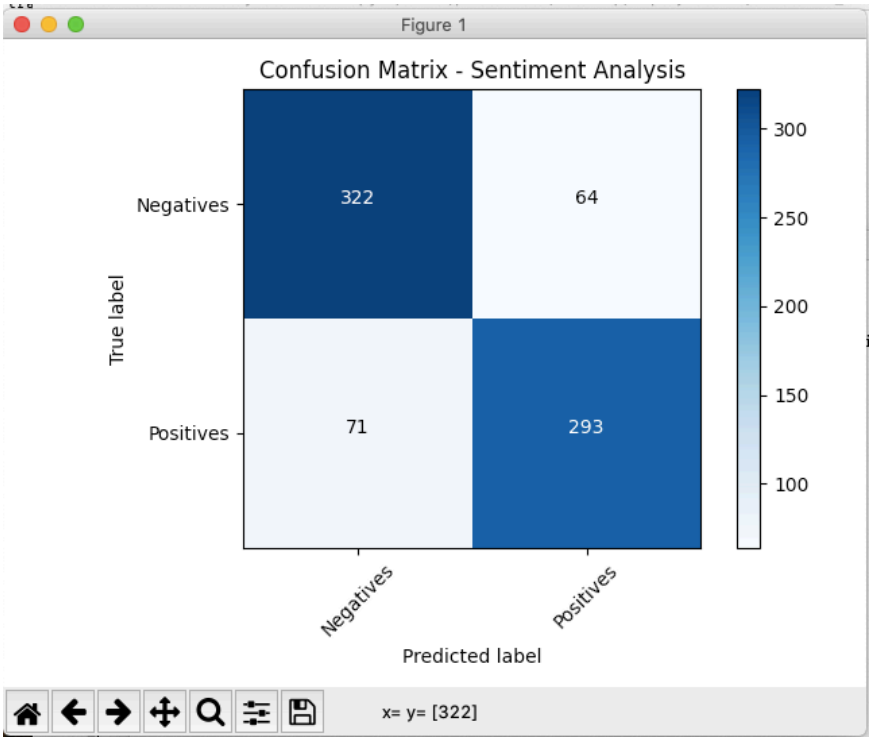
Negative predictions when the actual result is negative;
Negative predictions when the actual result is positive;
Positive predictions when the actual result is negative;
Positive predictions when the actual result is positive;

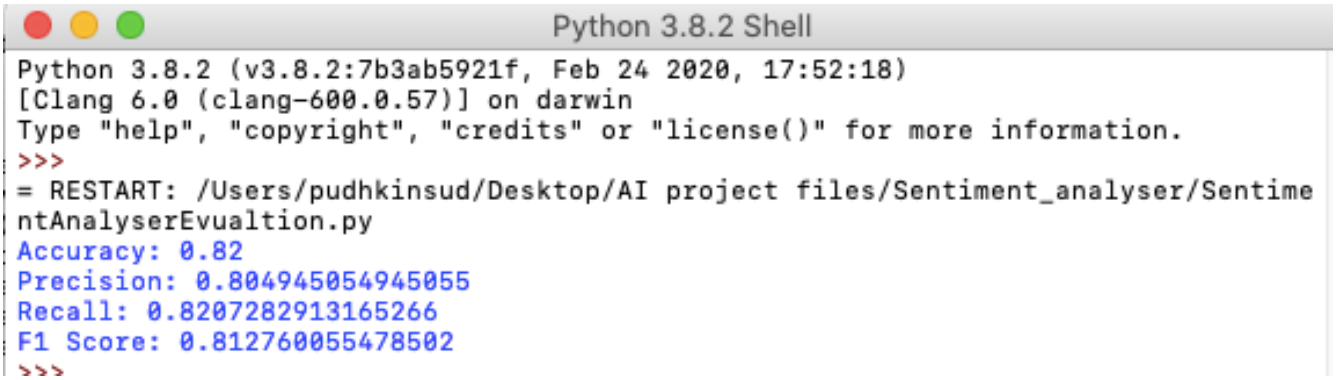
Screen Shots



Displaying whether entered data is positive or negative

Confusion Matrix





Displaying the Accuracy , Precision, Recall value and F1 Score

Formulas:

- accuracy = (true_positives + true_negatives) / (true_positives + true_negatives + false_positives + false_negatives)
- precision = true_positives / (true_positives + false_positives)
- recall = true_positives / (true_positives + false_negatives)
- f1_score = 2*(recall * precision) / (recall + precision)

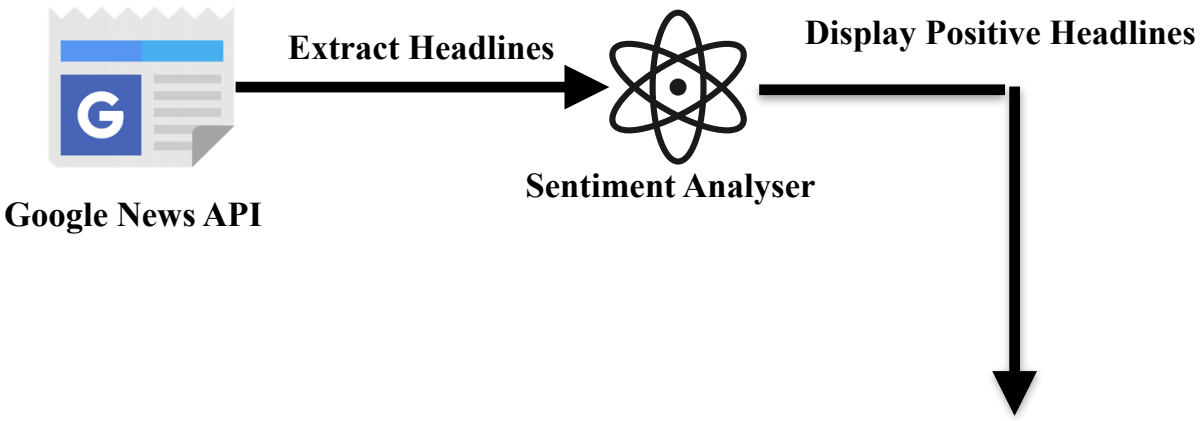
Conclusion and future enhancements:

The sentiment analyser provides consistency in critiquing.It's estimated that people only agree around 60-65% of the time when determining the sentiment of a particular text. The Analyser has no bias and is consistent with the analysing parameters.

The sentiment analyser can be used in many fields

- Businesses Companies can use the analyser on social media platforms to understand which trends are working and which aren't. They can also run any content that they put out through the analyser to check if it matched with their brand images (this is possible with the help of personality sentiment analysers that work on a similar principle but have a wider variety of classification).
- PR crises can be detected early and resolved.By analysing the sentiment of companies, their stock market trends can be predicted.
- Predicting of election resulting by analysing the sentiment of people for a certain candidate can help in prediction.

We used the analyser to extract headlines from google APIs and only display the headlines that are positive in nature.




```
= RESTART: /Users/pudhkinsud/Desktop/AI project files/Sentiment_analyser/GoodNewsInfo.py
38
5       UN chief: World should follow South Korea on COVID-19 fight - Times of India
 https://timesofindia.indiatimes.com/world/rest-of-world/un-chief-world-should-follow-south-korea-on-c
ovid-19-fight/articleshow/75482882.cms
10      Coronavirus: 14 more Nanded pilgrims test positive in Mohali, Ferozepur - The Tribune India
 https://www.tribuneindia.com/news/punjab/coronavirus-14-more-nanded-pilgrims-test-positive-in-mohali-
ferozepur-78957
11      Reliance Jio working on Zoom rival: 10 things to know - Gadgets Now
 https://www.gadgetsnow.com/slideshows/reliance-jio-working-on-zoom-rival-10-things-to-know/photolist/
75484420.cms
14      Experts give thumbs up to RIL's rights issue, term it as an opportunity for investors - Money
control.com
 https://www.moneycontrol.com/news/business/markets/experts-give-thumbs-up-to-rils-rights-issue-term-i
t-as-an-opportunity-for-investors-5210361.html
>>>
```