(ML DL) May 4

Project

On

Amazon's Cell Phone Review

Using

Natural Lanaguage Preprocessing(NLP)

By Team 10

1) Sparsh Kumar Singh
2) A.Pooja
3) K.Anusha
4) P.Laxmi Poojitha

Table of Contents

1.INTRODUCTION

1.1 Overview :

Amazon is one of the best mobile applications which is considered as a treasure trove of all different categories of mobiles along with extra information like rating and comments on that particular mobile product which are submitted by the consumers. This rating and reviews provided by e-Commerce create a transparent system for other consumers which helps them to make an informed decision and purchase that product. Based on these ratings and reviews an application called "Amazon's cell phone reviews prediction system" has been developed.

1.2 Purpose:

The purpose of this project is to predict whether the reviews of different cell phones are either positive or negative whenever a person visits an Amazon's e-commerce website to purchase cell phones, he/she reads different reviews submitted by a different person and thus they conclude to purchase the product or not, sometimes consumers may get into a dilemma after reading a lot of reviews and reading all the reviews of a particular product may be time-consuming so in order to save the time and avoid getting into confusion, consumers can make use of an application which takes different reviews and predict whether the reviews are positive or negative.

2. LITEARATURE ANALYSIS

2.1 Existing Problem :

Generally, consumers while purchasing any mobile from amazon e-commerce website they go through many reviews on that product, and at last after spending alot of time they conclude themselves whether the reviews are positive or negative,sometimes they may fell to take decisions or may get in a dilemma.For this problem, different solutions have been provided like

---> Naive Bayes Model

All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The naiveBayes algorithm looks at particular keywords of a review to describe whether it is positive or negative, depending on the output set.

2.2 Proposed Solution :

The main objective of the project is to perform sentiment analysis accurately.

In the solution, Natural Language Processing tool kit has been used to develop a model, in order to build these models following steps were taken.

Data collection

In this step, different websites were visited like GitHub, Kaggle, etc and finally, review.csv file is selected for model building, from the collected csv file, only the required columns like rating and review are selected for each mobile model and rest all are deleted. Based up on the rating, review column has been classified as either positive or negative in separate column which act as output column.

Text Preprocessing

In this process entire dataset is imported though pandas and converted to dataframe

dataset = pd.read_csv('review.csv')

We start off by pre-processing the data, removing unnecessary words that don't help our prediction. Then, we take important words in their stemmed forms (e.g lov is the stem for loved, loving, or lovely). We then train the machine to learn which reviews are positive based on their word stems. After that we test the data using similar information, to see how accurately our machine can predict whether a review is positive or negative (1 or 0).

NLP tools are

NLTK which stands for Natural Language Toolkit. From NLTK, 2 classes are imported : the stopwords class, and the PorterStemmer class.

Vectorization

Vectorization allows us to extract features from our textual data, we take out all the words from each of the observations and pool them together in a sort of “bag” to reduce redundancy by not counting duplicates. This is done by importing the CountVectorizer class from sklearn.

Each word forms its own column in a way, and since there are so many words, there could be a huge number of columns. However, max_features have been specified with the maximum number of columns using the max_features parameter of the CountVectorizer class.

Then fit the column of words to the X (input) variable and specified y (output) variable as our second column in the dataset, which gives a 1 or 0, depending on whether the review was positive or negative respectively.

Split the data to train and test

I then had to split the dataset into a training set and a test set, and used a test size of 0.2, so that we have 800 values to train our dataset and 200 values to test it with.

Model Building (Artificial Neural Network)

In this process necessary libraries are imported like model is initialized with the help of Sequential,and then input layer, hidden layer and output layer is added to the model.For output layer activation function is sigmoid, since the datasets are in classification format.

Prediction

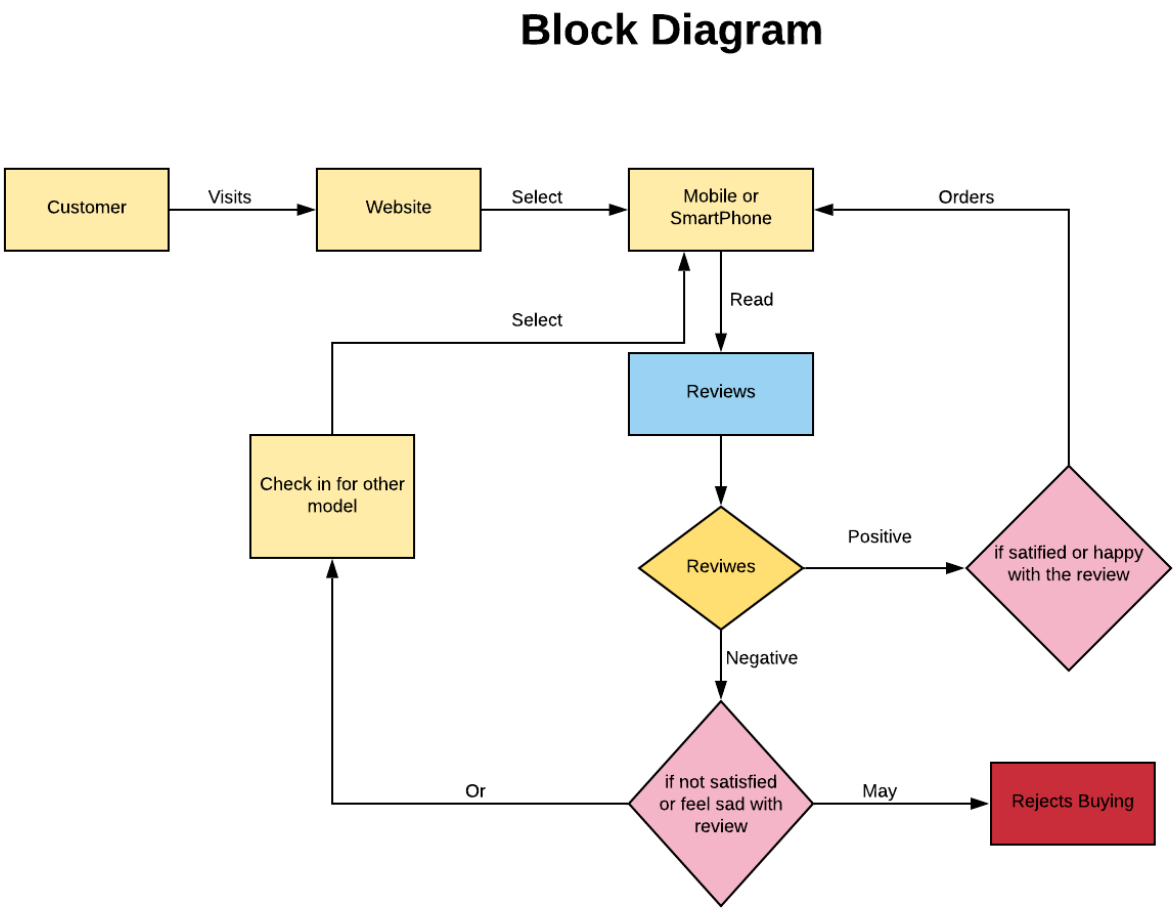
Once the model is get trained we test with the x_test data and then compare with the y_test .

Save the Model

save the model with the .h5 extension
model.save("phone_model.h5")

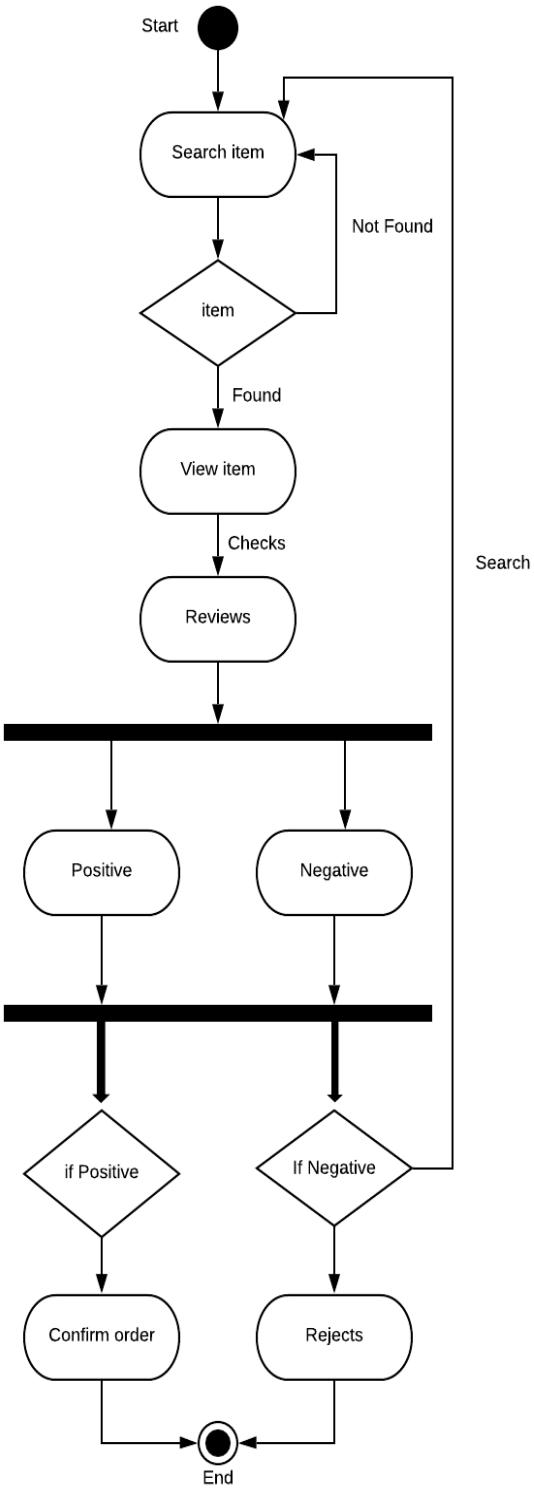
3.THEORITICIAL ANALYSIS

3.1 Block Diagram:

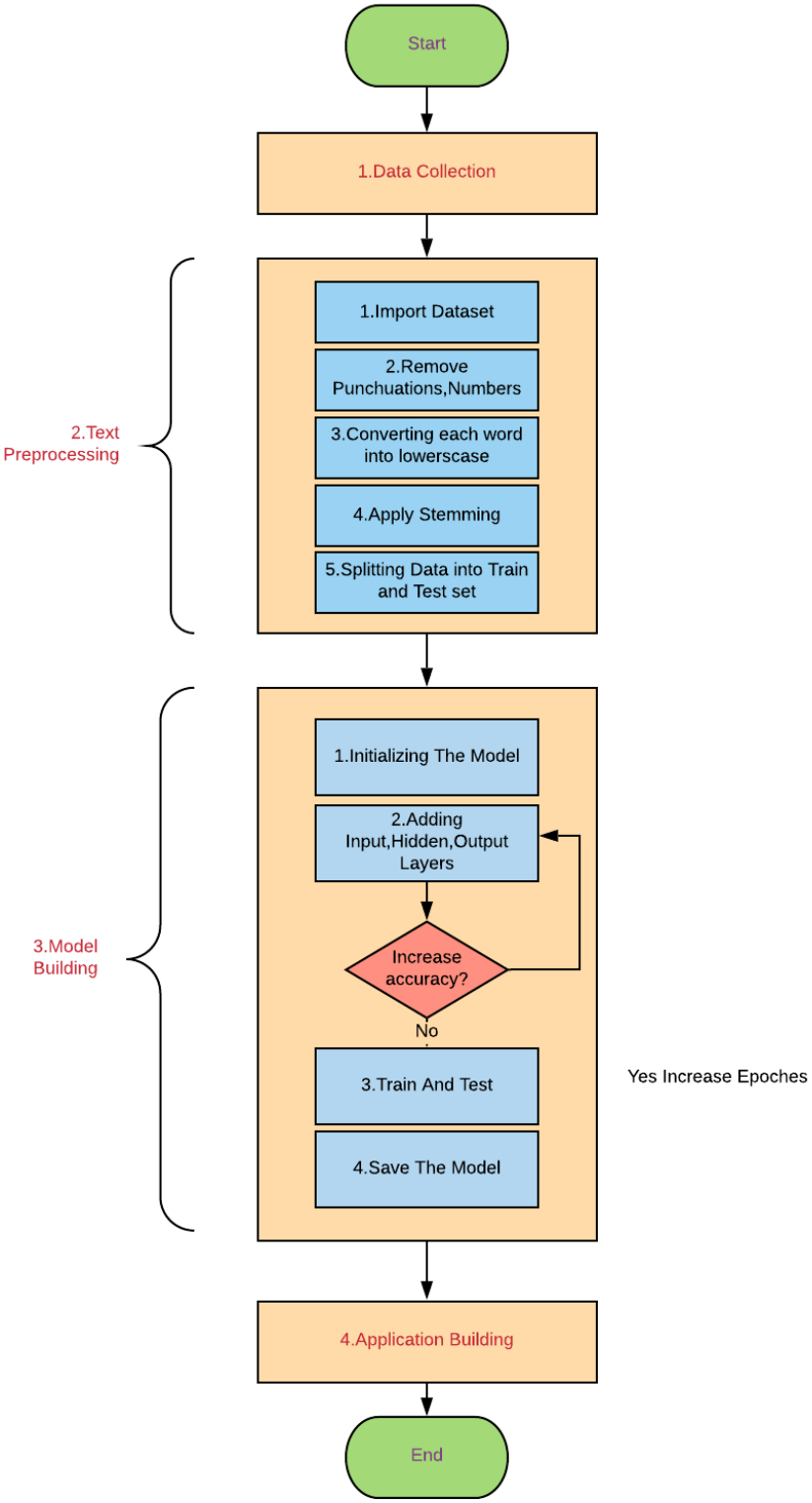


3.2 Hardware/Software Desigining :

Acitivity Diagram



4. FLOW CHART


Start
1.Data Collection
2.Text Preprocessing
1.Import Dataset
2.Remove Punchuations,Numbers
3.Converting each word into lowerscase
4.Apply Stemming
5.Splitting Data into Train and Test set
3.Model Building
1.Initializing The Model
2.Adding Input,Hidden,Output Layers
Increase accuracy?
No
Yes Increase Epoches
3.Train And Test
4.Save The Model
4.Application Building
End

5. RESULT

After applying NLP and then fit the ANN model the accuracy was 99.34%.
the model has been tested with different reviews like

Review	***Result***
1. Awesome Phone	Positive
2. I hate this phone	Negative
3. Great phone	Positive

6. ADVANTAGES AND DISADVANTAGES

ADVANTAGES :

1. This is used to analyze the reviews of a particular product before purchasing it.

2. Natural Language Processing is used to predict the good quality mobile phones based on the ratings and reviews given by the consumers.

3. It is also helpful for the customers to identify whether a purchased mobile is good or not based on the prediction given by the model.

4. Based on this prediction the customers can easily identify the good quality product without having much knowledge regarding the product they.

5. Even it can be used for analyzing the way the released brand is growing outside the world.

DISADVANTAGES :

1. There are some chances that the prediction will not be correct because of more number of characters in the reviews.

2. Even the predictions tend to fail if the customers fail to provide a needful information needed by us to predict about a particular product .

7. APPLICATIONS

1. Used to identify the good quality mobile phones.
2. It can be implemented on the website.

8. CONCLUSION

We have created a sentiment analysis web page which works for sentiment analysis.
The vast majority of data that exists is unstructured text data hence the Natural language processing seeks to extract information from text data through a variety of statistical techniques . To get a sense of the text across a large set of text data, it is sometimes useful to analyze the most commonly occurring words. Text data, in this case the Amazon reviews consists of descriptions of over a million of peoples reviews on unlocked cell phones. Here Natural Processing Language allows us to calculate the positive and negative sentiment present within the text. As we have divided

the sentence into words converted them into lowercase then again performed stemming on it and later on combined them into sentence. Later on after visualizing and identifying the direction of sentiment, we tried to build a model to predict whether a review is positive or negative for reviewers who did not clearly specify their rating. Then this estimation we used to compare against the actual sentiment to see how well the model predicts.

9. FUTURE SCOPE

Sentiment analysis is an important concept and one of the most effective tools of improving the conversion rate. Reading the sentiment of consumers, not only enables businesses to reach out to their target audience, but also enables them understand their needs and feelings. It provides a bird-eye view to brands and let them observe and protect their prestige. Additionally, it automates a cumbersome process of going through millions of lines of text to better read and listen to the demands and concerns of consumers. That, in turn, helps manage unpredictable damaging scenarios and ease the cost of doing so. Daily, weekly, and monthly reports of sentiment analysis can help a brand improve its image, set its pricing appropriately, and improving its relationship with consumers. It can also be turned into a tool of tracking sector-wide trends and demands, including competitors contents and strategies, to contribute to a competitive advantage

10. BIBILOGRAPHY

1. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit

Authors: Steven Bird, Ewan Klein, and Edward Loper
This book is a helpful introduction to the NLP field with a focus on programming. If you want have a practical source on your shelf or desk, whether you're a NLP beginner, computational linguist or AI developer, it contains hundreds of fully-worked examples and graded exercises that bring NLP to life. It can be used for individual study, as a course textbook when studying NLP or computational linguistics, or in complement with artificial intelligence, text mining, or corpus linguistics courses.

2. Natural Language Understanding

Author: James Allen

This book is another introductory guide to NLP and considered a classic. While it was published in 1994, it's highly relevant to today's discussions and analytics activities and lauded by generations of NLP researchers and educators. It introduces major techniques and concepts required to build NLP systems, and goes into the background and theory of each without overwhelming readers in technical jargon.

11. APPENDIX

A. SOURCE CODE

```
import pandas as pd
import numpy as np
from google.colab import drive
drive.mount('/content/gdrive')
data = pd.read_csv('reviews.csv',error_bad_lines=False)
data.head()
data.shape
data['rating'].unique()
data.insert(8,"output","")

 for i in range(0,len(data['rating'])):
  num=data['rating'][i]
  if num>3:
    data['output'][i]=1
  else:
    data['output'][i]=0

data['review']=data['title']+data['body']

data=data.drop(["title","body"],axis=1)
y=data.iloc[:,6:7].values
y.shape
```

```
import re
import nltk          #natural language tool kit
nltk.download("stopwords")
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()

data['review'].isnull().any()
data['review'].fillna(data['review'].mode()[0],inplace=True)
data['review'].isnull().any()

list1=[]
for i in range(len(data['review'])):
    review=data["review"][i]
    review=re.sub("[^a-zA-Z]"," ",review)
    review=review.lower()
    review=review.split()
    review=[ps.stem(word) for word in review if not word in
set(stopwords.words('english'))]
    review=" ".join(review)
    list1.append(review)

#vectorization
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=3000)
x=cv.fit_transform(list1).toarray()
x.shape

!mkdir -p count1
import pickle
Filename="count1/countvectorizer.pkl"
with open(Filename,"wb") as file:
```

```
pickle.dump(cv,file)

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=
0)
x_train.shape

from keras.models import Sequential
from keras.layers import Dense
model = Sequential()
model.add(Dense(units =3000  ,init = "random_uniform",activation = "relu"))
model.add(Dense(units =3000  ,init = "random_uniform",activation = "relu"))

model.add(Dense(units = 1 ,init = "random_uniform",activation = "sigmoid"))
model.compile(optimizer = "adam",loss = "binary_crossentropy",metrics =
["accuracy"])
model.fit(x_train,y_train,epochs  = 8)

model.save("phone_model.h5")

y_pred=model.predict(x_test)
y_pred=(y_pred>0.5)
y_pred
y_test
y_p=model.predict(cv.transform(['bad phone']))
y_p
y_p>0.5
y_p=model.predict(cv.transform(['very bad']))
y_p=y_p>0.5

model.save("phone_review.h5")
```