### AL (ML DL) APRIL 15

**Project**

**On**

**Analysis of Amazon Cell Phone Reviews**

Using

Python,Python Web Frame Works,NLP

By Team 20

1) MERUGU VIJAYASREE

2) PARRE NIKHITHA

3) PUPPALA NAIMESHA

4) SNEHALATHA SAMBAGANI

5) MAVLURI.SWATHI

|  |  |  |  |
| --- | --- | --- | --- |
| **1** |  | **INTRODUCTION** | |
|  |  | 1.1 Overview | |
|  |  | 1.2 Purpose | |
| **2** |  | **LITERATURE SURVEY** | |
|  |  | 2.1 Existing problem | | |
|  |  | 2.2 Proposed solution | | |
| **3** |  | **THEORITICAL ANALYSIS** | |
|  |  | 3.1 Block diagram | |
|  |  | 3.2 Hardware / Software designing | | |
| **4** |  | **EXPERIMENTAL INVESTIGATIONS** | | |
| **5** |  | **FLOWCHART** | |
| **6** |  | **RESULT** |  |
| **7** |  | **ADVANTAGES & DISADVANTAGES** | | |
| **8** |  | **APPLICATIONS** | |
| **9** |  | **CONCLUSION** | |
| **10** |  | **FUTURE SCOPE** | |
| **11** |  | **BIBILOGRAPHY** | |
|  |  | **APPENDIX** | |

1. **Introduction** 
   1. **Overview:**

Language is a method of communication with the help of which we can speak, read and write. For example, we think, we make decisions, plans and more in natural language; precisely, in words. However, the big question that confronts us in this AI era is that can we communicate in a similar manner with computers. In other words, can human beings communicate with computers in their natural language? It is a challenge for us to develop NLP applications because computers need structured data, but human speech is unstructured and often ambiguous in nature.

in this sense, we can say that Natural Language Processing (NLP) is the sub-field of Computer Science especially Artificial Intelligence (AI) that is concerned about enabling computers to understand and process human language. Technically, the main task of NLP would be to program computers for analyzing and processing huge amount of natural language data.

**1.2 Purpose:**

In recent years, the ML methods have become popular as they allow researchers to improve the prediction accuracy of concrete properties and are used for various engineering applications. The ML methods have been used to increase the prediction accuracy of concrete properties, and the data derived from the literature sources were used

Regression models tend to be used for the prediction of the compressive strength of high-strength concrete. These models also demonstrate how the concrete compressive strength depends on the mixing ratios.

Previous studies evaluated the amount of the concrete component materials and compared their results to the published data. In this study, the ML regression methods were compared to predict the compressive strength and slump values of the cube samples. The samples were prepared by accounting for seven simultaneously controllable effect variables in the laboratory. The study aimed to determine the most successful regression method by comparing the decision tree (DT), random forest (RF), support vector machine (SVM)

1. **Literature Survey**

**2.1 Existing Problem:**

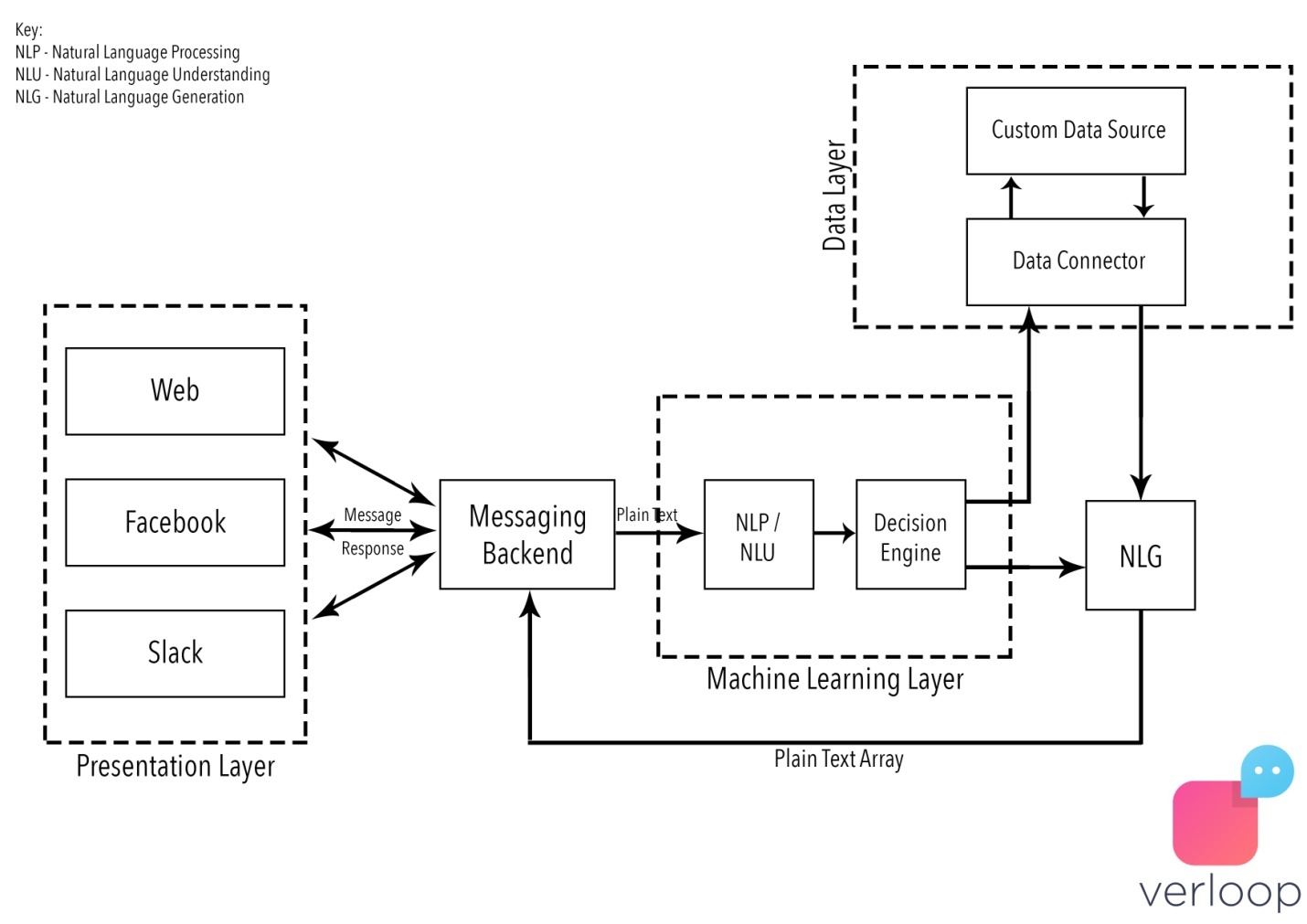
90 percent of the consumers read online reviews before they decide to purchase any Mobile phone from any e-commerce website .Online Mobile applications has revolutionised the way consumers purchase mobile phones online as these apps have all the information regarding any mobile phone at users finger tips. Amazon is one of the best mobile applications which is considered as a treasure trove of all mobile reviews, and their review system is accessible across all channels presenting reviews in an easy-to-use format. So,There should be a system which analyses thousands of reviews of unlocked mobile phones sold on Amazon.com to find insights with respect to reviews, ratings.

**2.2 Proposed Solution :**

This project aims at building a model to predict the helpfulness of the review and the rating based on the review text. Corpus-based and knowledge-based methods can be used to determine the semantic similarity of review text. We will be using Natural language processing to analyse the sentiment ( positive or a negative) of the given review . A sample web application is integrated to the model built.

1. **Theoretical Analysis :**

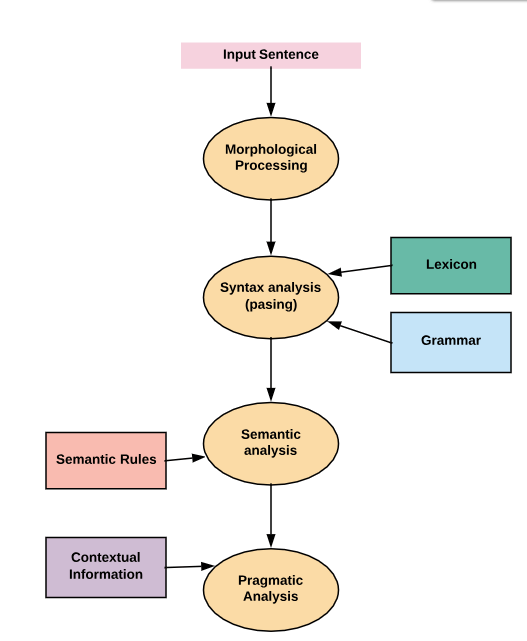
**3.1 Block Diagram:**



**NOTE :**

Comparative machine learning experiments have become an important methodology in empirical approaches to natural language processing (i) to investigate which machine learning algorithms have the ‘right bias’ to solve specific natural language processing tasks, and (ii) to investigate which sources of information add to accuracy in a learning approach. Using automatic word sense disambiguation as an example task, we show that with the methodology currently used in comparative machine learning experiments, the results may often not be reliable because of the role of and interaction between feature selection and algorithm parameter optimization. We propose genetic algorithms as a practical approach to achieve both higher accuracy within a single approach, and more reliable comparisons.

**3.2Block Diagram:**



**3.4Hardware / Software designing:**

Python, Python Web Frame Works, Python for Data Analysis, Python for Data Visualization, Data Pre-processing Techniques, Machine Learning, Regression Algorithms

1. **Experimental Investigation :**

**Data format:**

product/productId: B00006HAXW

product/title: Rock Rhythm & Doo Wop: Greatest Early Rock

product/price: unknown

review/userId: A1RSDE90N6RSZF

review/profileName: Joseph M. Kotow

review/helpfulness: 9/9

review/score: 5.0

review/time: 1042502400

review/summary: Pittsburgh - Home of the OLDIES

review/text: I have all of the doo wop DVD's and this one is as good or better than the

1st ones. Remember once these performers are gone, we'll never get to see them again.

Rhino did an excellent job and if you like or love doo wop and Rock n Roll you'll LOVE

this DVD !!

**How to parse (in Python):**

import gzip

import simplejson

def parse(filename):

f = gzip.open(filename, 'r')

entry = {}

for l in f:

l = l.strip()

colonPos = l.find(':')

if colonPos == -1:

yield entry

entry = {}

continue

eName = l[:colonPos]

rest = l[colonPos+2:]

entry[eName] = rest

yield entry

for e in parse("all.txt.gz"):

print simplejson.dumps

1. **Advantages and Disadvantages:**

**Advantages:**

The **benefits of natural language processing** are innumerable. **Natural language processing** can be leveraged by companies to improve the efficiency of documentation processes, improve the accuracy of documentation, and identify the most pertinent information from large databases.

**Disadvantages:**

* Complex Query Language- the system may not be able to provide the correct answer it the question that is poorly worded or ambiguous.
* The system is built for a single and specific task only; it is unable to adapt to new domains and problems because of limited functions.

**6)Applications:**

1. Machine Translation. Everyone knows what is a manual translation — we translate information from one language into another. ...
2. Speech Recognition. ...
3. Sentiment Analysis. ...
4. Question Answering. ...
5. Automatic Summarization. ...
6. Chatbots. ...
7. Market Intelligence. ...
8. Text Classification.

**7)Abstract:**

Deep learning methods employ multiple processing layers to learn hierarchical representations of data, and have produced state-of-the-art results in many domains. Recently, a variety of model designs and methods have blossomed in the context of natural language processing (NLP). In this paper, we review significant deep learning related models and methods that have been employed for numerous NLP tasks and provide a walk-through of their evolution. We also summarize, compare and contrast the various models and put forward a detailed understanding of the past, present and future of deep learning in NLP.

**8)Conclusion:**

This study contributes to both theory and practice. By building on the foundation of the economics of information, we provide a theoretical framework to understand the context of online reviews. Through the application of the paradigm of search and experience goods we offer a conceptualization of what contributes to the perceived helpfulness of an online review in the multistage consumer decision process. The type of product affects information search and evaluation by consumers. We show that the type of product moderates the effect of review extremity and depth on the helpfulness of a review. We ground the commonly used measure of helpfulness in theory by linking it to the concept of information diagnosticity As a result, our findings help extend the literature on information diagnosticity within the context of online reviews. We find that review extremity and review length have differing effects on the information diagnosticity of that review, depending on product type.

We also found that length increases the diagnosticity of a search good review more than that of an experience good review. This is consistent with classification of search and experience goods, in that it is easier to gather information on product quality for search goods prior to purchase. In the context of an online retailer, information comes in the form of a product review, and reviews of search goods lend themselves more easily to a textual description than do reviews of experience goods.

**9)Future Scope :**

* Data Retrieval with Web Scraping.
* Text wrangling and pre-**processing**.
* Parts of Speech Tagging.
* Shallow Parsing.
* Constituency and Dependency Parsing.
* Named Entity Recognition.
* Emotion and Sentiment Analysis.

**10)Bibliography :**

* + **Books:**
* Natural Language Processing Courses | Stanford Online
* Natural Language Processing (NLP): Deep Learning in Python | Udemy
* How to Build, Train, and Deploy Machine Learning Models with Amazon SageMaker
* **Algorithms**
* Thesmartbridgeteachable.com

**11)Appendix :**

### import numpy as np

### import matplotlib.pyplot as plt

### import pandas as pd

### data = pd.read\_csv('reviews.csv')

### df=data

### df=df[['Reviews','Rating']]

### df=df[df['Rating']!=3]

### df=df.reset\_index(drop=True)

### df['sentiment']=np.where(df['Rating'] > 3, 1, 0)

### import re

### import nltk

### nltk.download('stopwords')

### from nltk.corpus import stopwords

### from nltk.stem.porter import PorterStemmer

### ps = PorterStemmer()

### data=[]

### for i in range(0,1000):

### review=df["Reviews"][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)

### data.append(review)

### 

### from sklearn.feature\_extraction.text import CountVectorizer

### cv = CountVectorizer(max\_features=1500,min\_df=5,ngram\_range=(1,2))

### X = cv.fit\_transform(df["Reviews"].apply(lambda x: np.str\_(x)))

### y=df.iloc[:,2].values

### y.shape

### y=y.reshape(-1,1)

### import pickle

### pickle.dump(cv, open("cv.pkl", "wb"))

### from sklearn.model\_selection import train\_test\_split

### X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.20, random\_state = 0)

### import keras

### from keras.models import Sequential

### from keras.layers import Dense

### model = Sequential()

### model.add(Dense(output\_dim = 8, init = 'uniform', activation = 'relu', input\_dim = 1500))

### model.add(Dense(output\_dim = 8, init = 'uniform', activation = 'relu'))

### model.add(Dense(output\_dim = 1, init = 'uniform', activation = 'sigmoid'))

### model.compile(optimizer = 'adam', loss = 'binary\_crossentropy', metrics = ['accuracy'])

### model.fit(X\_train,y\_train,batch\_size=32,epochs=10)

### model.save('mymodel.h5')