Sarcasm Detection

Maheswara Reddy

51518487

Problem Statement

- To predict whether the given statement is sarcastic or not.
- The following example shows this.
  - I love being ignored. (Sarcastic)
  - I love being pampered. (Non Sarcastic)
- Inputs/Dataset to be used:
  - Input is taken from "https://nlp.cs.princeton.edu/SARC/" version 2.0
- Output:
  - Presentations and Report.
  - Codes, Environment Setup
  - Final metrics values.

Proposed Solution

Abstract

- To predict whether the given statement is sarcastic or not
- Sarcasm, as linguist Robert Gibbs noted, includes “words used to express something other than and especially the opposite of the literal meaning of a sentence.”
- However, it’s not always easy to figure out if a writer is being sarcastic.
- Sarcasm thrives in ambiguous situations – and that’s the main issue.
- Sarcasm transforms the polarity of an apparently positive or negative utterance into its opposite.
- Automatic detection of sarcasm is still in its infancy. One reason for the lack of computational models has been the absence of accurately-labeled naturally occurring utterances that can be used to train machine learning systems.
- While speaking, people often use heavy tonal stress and certain gestural clues like rolling of the eyes, hand movement, etc. to reveal sarcastic. In the textual data, these tonal and gestural clues are missing, making sarcasm detection very difficult for an average human.

Data

- Data set is taken from "https://nlp.cs.princeton.edu/SARC/" version 2.0
- The dataset is divided into four categories.
  - Main Balanced: This is the primary dataset which contains a balanced distribution of both sarcastic and non-sarcastic comments.
  - Main imbalanced: To emulate real-world scenarios where the sarcastic comments are typically fewer than non-sarcastic ones, this can be used an imbalanced version of the Main dataset.
  - Pol Balanced: This is the subset of man which contains politics related comments with balanced distribution of sarcastic and non-sarcastic comments.
  - Pol imbalanced: This is the subset of man which contains politics related comments with unbalanced distribution of sarcastic and non-sarcastic comments.
  - For more information read readme file in the above link.

Data Pre-Processing

Train Data

	7vq9q	c07jfvv	c07jy05	1	0
0	7xdys	c07o37s	c07o350	1	0
1	bln1z	c0ndefe	c0ndajx	1	0
2	bm9yo	c0nh0jw	c0nhdes	1	0
3	bpkof	c0nyigy	c0ny03s	0	1
4	bpuo1	c0nzcjq	c0nz11j	0	1

Test Data

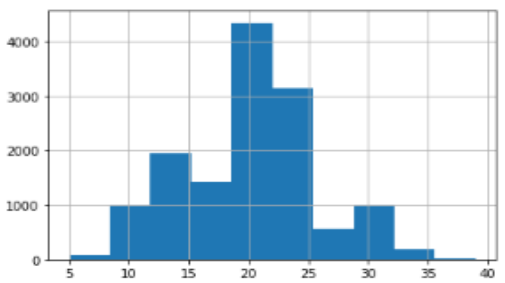
	hqa1x	c1xiujs	c1xj4e2	1	0
0	i0v01	c205da7	c201mb5	1	0
1	i6i1y	c21btxl	c21bxjw	1	0
2	i77mp	c21hz0p	c21jnd1	1	0
3	xie15	c5mw3ss	c5muofa	1	0
4	xmaqj	c5nlo2y	c5nz71q	0	1

Json Data

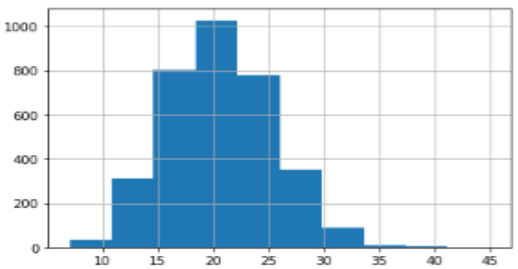
	7uxqr	7vewt	7vq9q	c07jfvv	7w0as	c07kq5w	c07myx2	7xdys	c07o37s	7xvzm	c07pjk9	7z0nk	c07sjyk	7yyaz	c07sfaw	c07sm7t	
text	Nancyt Pelosi messes up.. 500 Million Jobs los...	Netflix CEO: "Please raise my taxes"	The Six Million Dead Jews of World War ONE!	Oh right, *both* wars were just jewish conspir...	GOP says it is necessary to spend my tax dolla...	DO NOT QUESTION THE HIVE MIND!	Yup, all Republicans think exactly the same way.	WSJ begins the Jeb Bush campaign for 2016	Good luck with that.	Breaking a crucial campaign promise: Obama Def...	Right, lets wait 4 more years until he can pro...	Cop Who Shoved Cyclist Fired by NYPD, Faces 4 ...	This is why folks are getting arrested for tak...	OK, I understand why food prices went up last ...	Nobody forces you to either eat at a restauran...	But if there is a demand for cheaper soda capi...	D H
author	Fishbum	jdl2003	[deleted]	Erobern	fangolo	_pi	jk1150	[deleted]	Mastrmind	[deleted]	ysaberi	Orangutan	flannelback	Wordie	uriel	son-of-chadwardenn	
score	0	1733	0	6	891	1	4	14	2	129	6	1115	288	207	2	0	
ups	2	1985	20	6	1058	1	4	24	2	241	6	1353	288	312	2	0	
downs	4	252	23	0	167	0	0	10	0	112	0	238	0	105	0	0	

Data Pre-Processing

- Read the Train and Test data from csv files
- ID's data is separated with space and Result data is also separated with space. Then concatenated these columns.
- Read the JSON file and Transposed the date to get the ID's in a column.
- Merged the JSON data with CSV files and removed unnecessary columns.
- Converted the data to lower
- Removed punctuation from the text
- Removed Stop Words from the data
- Lemmatization have been done on Train and Test data
- Counted individual words and created a dictionary word with count
- Vocabulary is converted to int for Train and Test Data
- Still some of the punctuation are left out in data, Assigned 0 to those punctuations.
- Train and Test data after processing, data distribution is specified in pictures next to the text.
- Transformed the data to Tensors before exposing to model
- Split-ted train data with 80:20. 20 percent of data is used for validation to tune the hyper parameters.

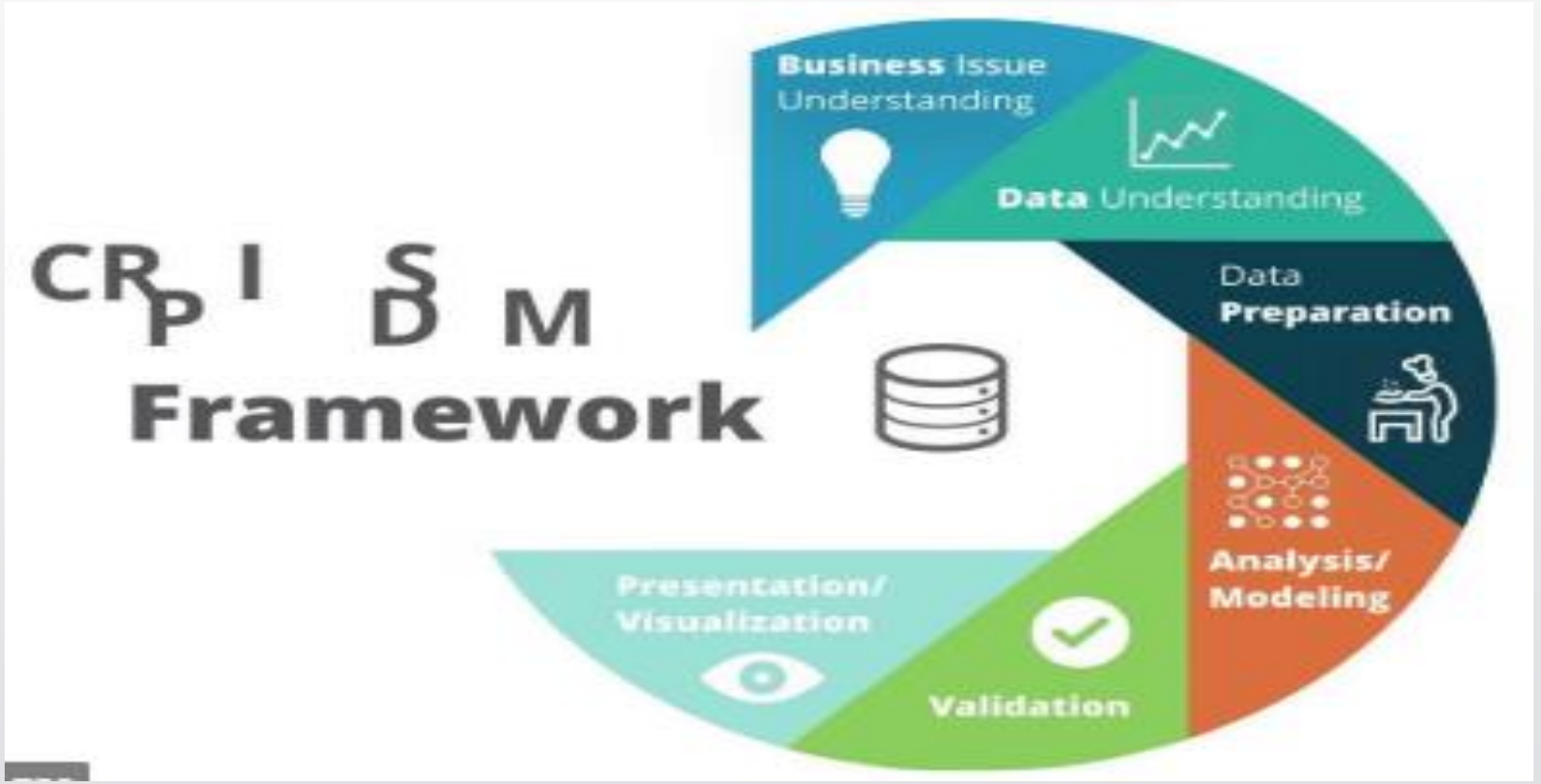



```
count    13666.000000
mean        20.415264
std          5.267803
min          5.000000
25%         17.000000
50%         21.000000
75%         25.000000
max         39.000000
dtype: float64
```

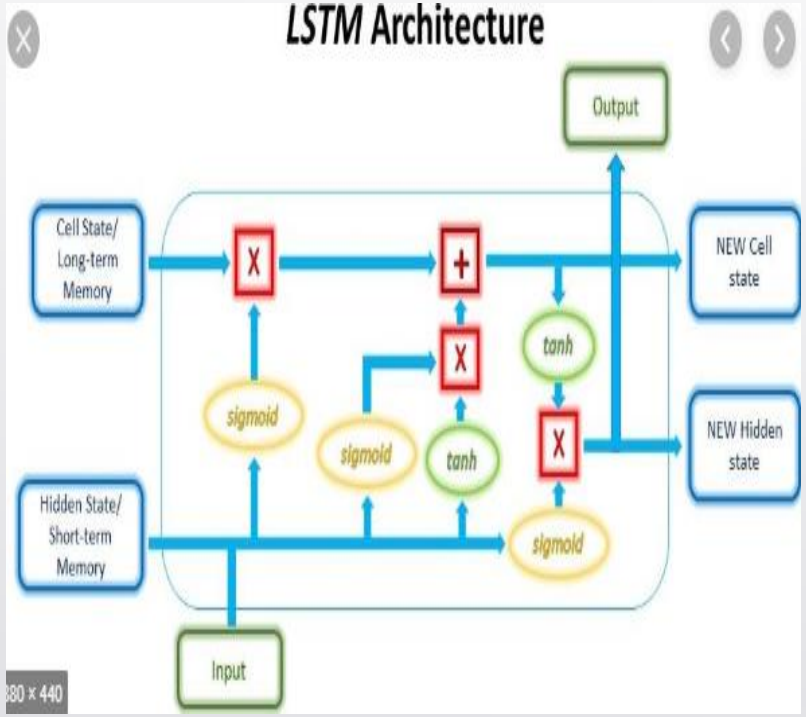



```
count    3404.000000
mean       20.420388
std         5.334077
min         7.000000
25%        17.000000
50%        21.000000
75%        25.000000
max        45.000000
dtype: float64
```

Modelling


CRISP DM
Framework
Business Issue Understanding
Data Understanding
Data Preparation
Analysis/ Modeling
Validation
Presentation/ Visualization

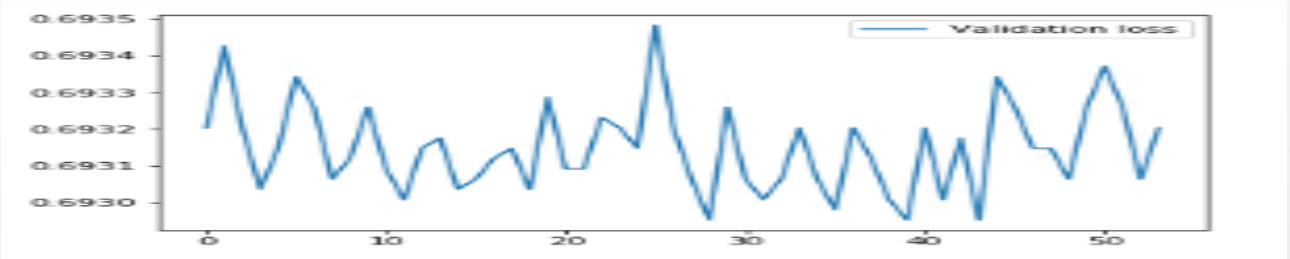
Modelling



- I used RNN LSTM deep learning technic to classify sarcastic comments.
- Total Vocab size is used. For im-balanced data size is increased to fit all the words.
- 400 Embedding dimensions, 512 hidden dimensions and 10 layers were considered to build LSTM network.
- Output size was considered as 1

Performance Measures

- Different learning rates were considered and almost for all the learning rates I got a accuracy of 50%.
- These I tried for 4 epochs.
- But even I tried for 100 epochs with learning rate of 0.01, I got the same accuracy and even validation loss was not changing it was around 0.69, even the training loss was around 0.69.
- The above slides procedure have been used for im-balanced data also.



Learning Rate	Accuracy	Test Loss
0.1	49.9	0.785
0.2	49.9	0.706
0.5	49.9	0.696
0.01	49.9	0.710
0.02	49.9	0.697
0.05	50.0	0.895
0.001	50.0	0.603
0.002	49.9	0.701
0.005	49.9	0.696
0.0001	49.9	0.711
0.0002	49.9	0.793
0.0003	49.9	0.695

Tools and Framework Used

- Python, Pandas, NumPy and Pytorch

Tools and Framework Used

- Python, Pandas, NumPy and Pytorch

Setup

- Project has been executed in Google Colab. So upload all the required input files in google drive. Then execute the below commands

  from google.colab import drive

  drive.mount('/content/gdrive')

- Imported below libraries
  - import pandas as pd
  - import numpy as np
  - import torch
  - import json
  - from string import punctuation
  - from spacy.lang.en import English
  - from spacy.lang.en.stop_words import STOP_WORDS
  - import en_core_web_sm
  - from collections import Counter
  - import pandas as pd
  - import matplotlib.pyplot as plt
  - import torch
  - from torch.utils.data import DataLoader, TensorDataset
  - import torch.nn as nn

Source Code

- Placed the code in below GitHub link:
- https://github.com/mahesh1982/HCLHackathon_SarcasticCalssification

Problems Faced and Future Implementations

- Reading a 2 GB Json File is so tricky.
- When I read it in Google Colab, its accommodating full RAM of colab. So I am unable to proceed with that and google colab session is restarted every time.

- I will implement this in Kaggle and in next submission I can implement it.
- I will use XLNet, BERT and RoBERTa for future implementations and I will submit it.

References

- https://nlp.cs.princeton.edu/SARC/
- https://arxiv.org/pdf/1704.05579.pdf
- https://theconversation.com/why-is-sarcasm-so-difficult-to-detect-in-texts-and-emails-91892
- https://www.sciencedirect.com/science/article/pii/S235286481630027X
- https://towardsdatascience.com/sentiment-analysis-using-lstm-step-by-step-50d074f09948
- https://github.com/mahesh1982/deep-learning-v2-pytorch/blob/master/sentiment-rnn/Sentiment_RNN_Solution.ipynb

About Me

- Currently Working as Technical Manager in UTC Aerospace ODC.
- Completed PG Diploma in Data Science from IIIT Bangalore in Jun 2019 with 3.3/4 CGPA
- Certified as Deep Learning Engineer from Udacity
- Certified as Data Scientist in R from Data Camp

Thank You