# Sarcasm Detection in News Headlines

Information Retrieval

Under the guidance of
Prof. Houwei Cao

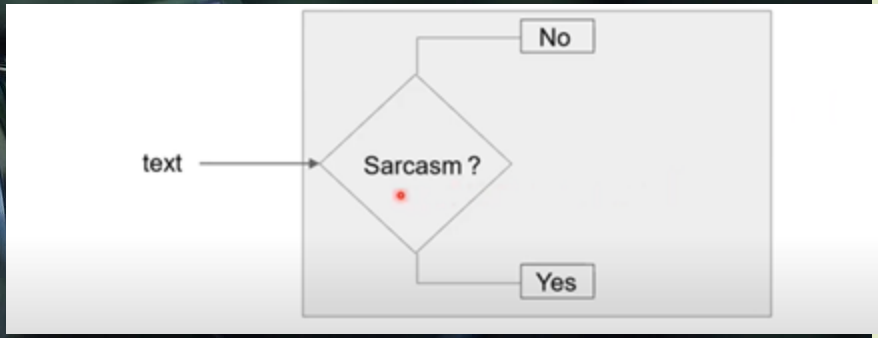
Group Members:
Surbhi Hirawat - 1301650
Jibbin Matthew- 1301096
Paridhi Nigam-1301148
Urja Shah- 1303753

# Contents

- Problem Statement
- Project Background
- Data Description
- Approach
- Models
- Experiments and results
- Project Demo
- Comparison
- Conclusion
- Future Work
- References

# PROBLEM STATEMENT

- This Project aims at detecting sarcastic comments within the news headline
- We aim to detect news headline that may have been written with sarcastic intent and hence misrepresented actual facts.
- **Sarcastic comments** generally tend to report false information with an intent to invert the sentiment of the expression that they seek to report.
- It is technically challenging sentiment analysis problem
- The fact that we cannot use a set of reference words without prior context ,to detect sarcasm,makes our task complicated.

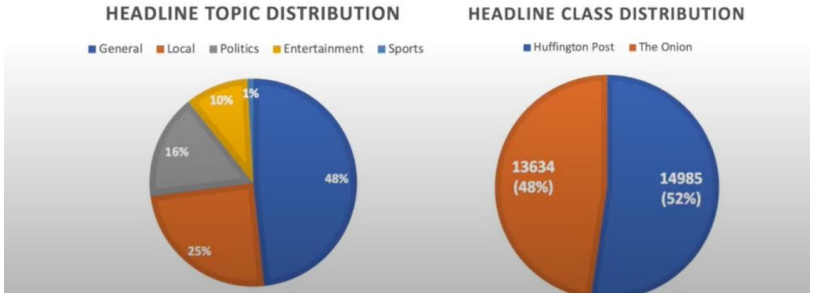


# Project background

- **Aim**- to build a sarcasm model that detect headline news using machine and deep learning.
- Attempt to understand how a computer learns the patterns of sarcasm.
- **Motivation**: Sarcasm is purely context-based, a common phenomena in social media and is inherently difficult to detect, which makes it sometimes difficult for humans to interpret.
- Studies- focused on social media content or review analysis which are usually noisy in terms of labels and language.
- To overcome- sarcasm detection in News Headlines.
- The **media** regularly seem to engage sarcasm in their news headline to get the attention of people. However, people find it tough to detect the sarcasm in the headline news, hence receiving a mistaken idea about that specific news and additionally spreading it to their friends, colleagues, etc.
- Consequently, an intelligent system that is able to distinguish between can sarcasm none sarcasm automatically is very important.

# DATA SOURCES

- The **data** for news headlines was sourced from Kaggle.
- The dataset contains news headlines collected from two news sources – The Onion and HuffPost.
- The Onion is known for sarcastic headlines, thus becomes a good source of data for this project.
- The headlines collected from HuffPost are non-sarcastic headlines.
- The dataset contains a total of 26709 observations
- Using a dataset with a proper mix of sarcastic and non-sarcastic headlines will be helpful to build non overfitting models.




| | article_link | headline | is_sarcastic |
|---|---|---|---|
| 0 | https://www.huffingtonpost.com/entry/versace-b... | former versace store clerk sues over secret 'b... | 0 |
| 1 | https://www.huffingtonpost.com/entry/roseanne-... | the 'roseanne' revival catches up to our thorn... | 0 |
| 2 | https://local.theonion.com/mom-starting-to-fea... | mom starting to fear son's web series closes ... | 1 |
| 3 | https://politics.theonion.com/boehner-just-wan... | boehner just wants wife to listen, not come up... | 1 |
| 4 | https://www.huffingtonpost.com/entry/jk-rowlin... | j.k. rowling wishes snape happy birthday in th... | 0 |

# Approach:

PRE PROCESSING

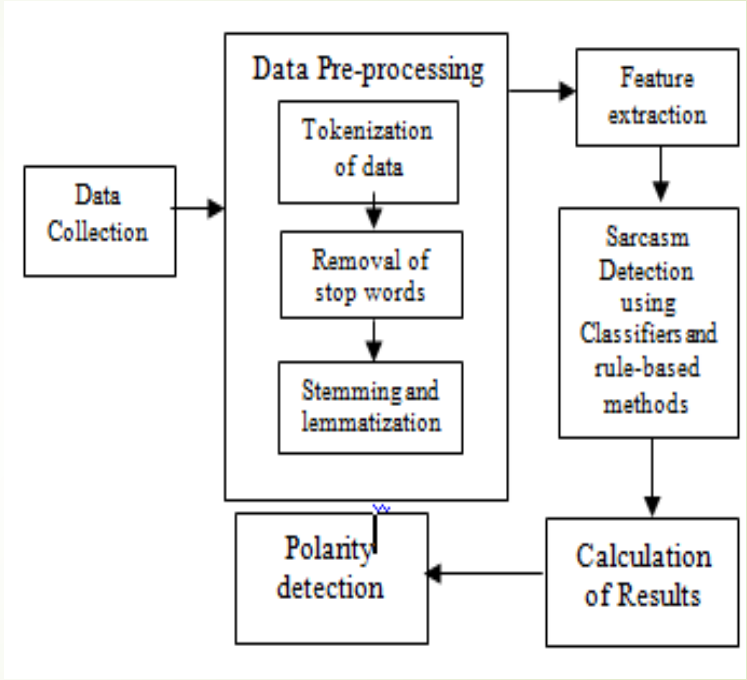
-Tokenization
-ASCII Conversion
-Stop-word Removal
-Lemmatizer

CLASSIFICATION MODELS

-CNN
-LSTM
-BERT

MODEL EVALUATION

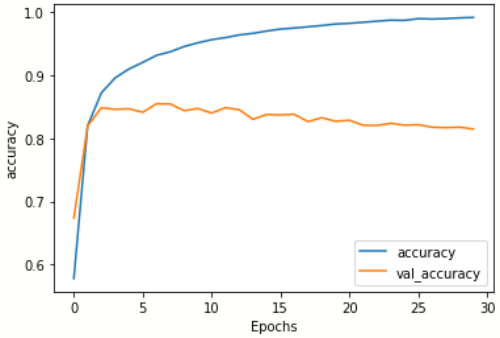
-Precision
-Recall
-Accuracy
-F1 Score
- Confusion Matrix

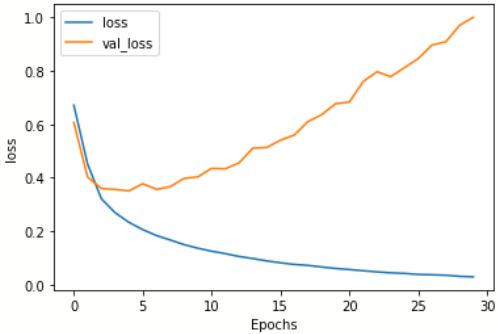


# CNN

- Embedding(100),Average Pooling,Dense(24)
- Binary Cross Entropy+Adam
- Epoch=30

- ACCURACY=85%

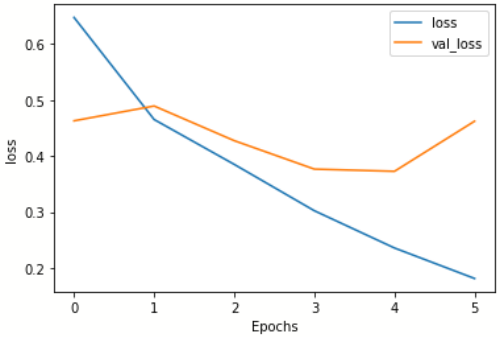


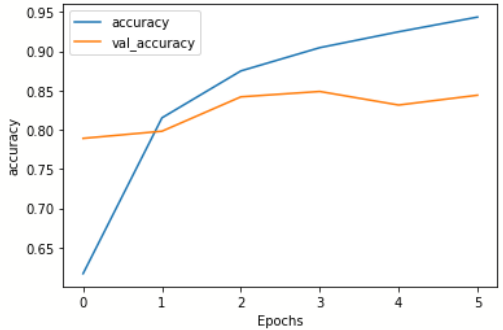



# LSTM

- Embedding Layer(100)+LSTM(100)+Average Pooling+Dropout(.5)+Dense(128)+Dropout(.5)+Dense(64)+Dropout(.5)
- Binary Cross Entropy+Adam
- Eopch=6,Batch Size=100

- **ACCURACY=95%**

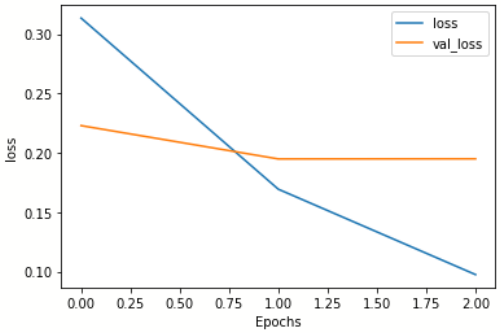


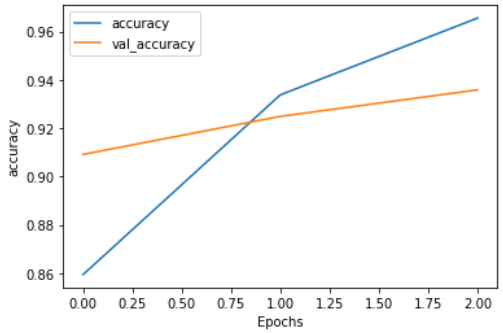



# BERT

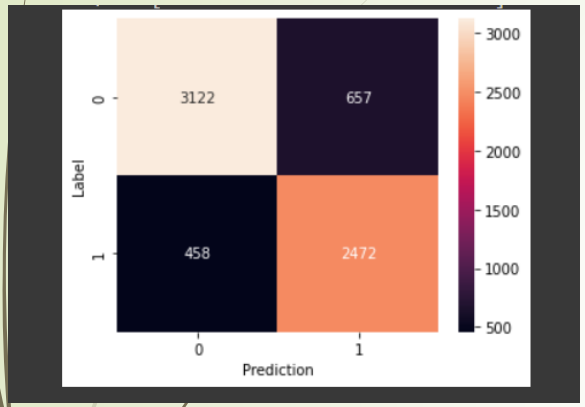
- Pre Trained Language Model(BERT BASE Uncased)
- Layers 128,Dropout(.2),Dense
- Epoch =3,Batch_Size=32

- ACCURACY=96%

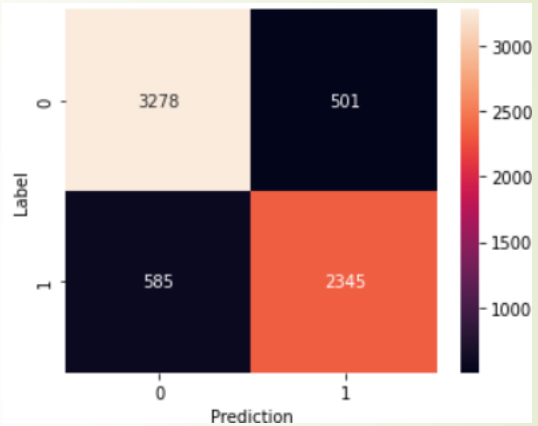


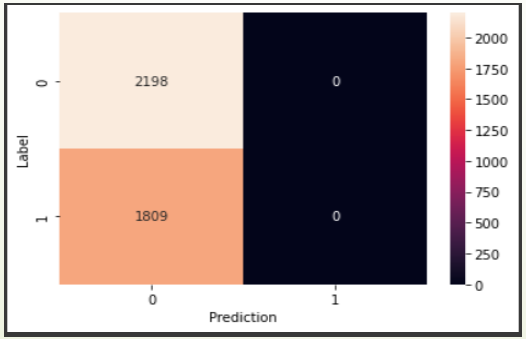
CNN




LSTM




BERT



# PERFORMANCE EVALUATION

| Model Name | Accur_Train | Accu_valid | Precision | Recall | F1_Score | Loss | Time |
|---|---|---|---|---|---|---|---|
| CNN | 96.88 | 82.98 | .83 | .83 | .83 | .0877 | 4ms/step |
| LSTM | 92.52 | 82.81 | .83 | .83 | .83 | .2653 | 12 ms/step |
| BERT | 96.22 | 92.29 | .55 | 1 | .71 | .1020 | 141ms/step |

# VISUALIZATION

Embedding projector - visualization of high-dimensional data (tensorflow.org)

mom starting to fear son's web series closest

## Sarcasm Detection Project

Analyzing headlines for checking if they are sarcastic

Enter a headline to classify

Classify Headline

Headline

*mom starting to fear son's web series closest*

## Result

It is a sarcastic headline!

# CONCLUSION & FUTURE WORK

- During our analysis, we constructed several Deep Learning Models and discovered that the model BERT is the most accurate way to solve the problem.
- The model is primarily built with the latest transformers and compared with the attention Mechanism
- The Bert has a large embedding dimension which allows the model to gain common knowledge and increase and model Performance.
- In future, we can ...