SARCASM DETECTION IN NEWS HEADLINES

BY

SURBHI HIRAWAT, PARIDHI NIGAM,

URJA SHAH, JIBBIN MATTHEW.

GRADUATE STUDENT,

DEPARTMANT OF DATA SCIENCE

UNDER GUIDANCE OF

PROF. HOUWEI CAO

AT

NEW YORK INSTITUTE OF TECHNOLOGY, MANHATTAN

INDEX

SR.NO TOPIC

1. ABSTRACT
2. INTRODUCTION
3. PROBLEM STATEMENT
4. PROCEDURE AND EXPLAINATION
5. OUTPUT
6. OBSERVATION AND RESULT
7. CONCLUSION
8. REFERENCE
9. **Abstract.**

Even for people, let alone computers, it can be challenging to discern sarcasm, a crucial component of communication. Newspapers frequently appear to use sarcasm in their headlines to engage readers. Sarcasm is a witty, frequently ironic comment used to mock or convey disgust. Sarcasm is a cutting, incisive, or harsh phrase; a stinging remark. Sentence detection task classification as "sarcastic" or "non-sarcastic". Lack of intonation and lack of facial emotions make it difficult.

1. **Introduction**

The atmosphere of the present day can be characterised as data-driven. This project seeks to identify sarcastic remarks in news headlines. Our goal is to identify news headlines that may have been intentionally sarcastic and hence misled the truth. Sarcastic remarks typically provide inaccurate facts with the intention of distorting the meaning of the utterance that they are intended to report.

Technically speaking, the sentiment analysis issue is difficult. Our work is compounded by the fact that we are unable to recognise sarcasm using a collection of reference phrases without any prior context.[2]

Purely contextual, sarcasm is challenging for people to recognise and understand. In terms of labels and vocabulary, noisy. to avoid news headlines that are sarcastic. The media frequently uses sarcasm in its news headlines to grab readers' attention. Therefore, it is crucial to have an intelligent system that can recognise sarcasm from none automatically.

The information for the news headlines came from Kaggle. The dataset includes news headlines compiled from The Onion and HuffPost, two news sources. Since The Onion is recognised for its ironic headlines, it serves as a reliable source of information for this project. The headlines that were taken from HuffPost are not ironic. [5] 26709 observations in total are included in the dataset. To create non-overfitting models, it will be beneficial to use a dataset that contains an appropriate balance of sardonic and non-sarcastic headlines.

Pre-Process: Lemmatizer, Tokenizer, ASCII Conversion, Stop-Word Removal Models for classification: CNN, BERT, and LSTM Precision, Recall, Accuracy, F1 Score, and Confusion Matrix are used to evaluate the model. [3] Therefore, after gathering the data for the data source, we pre-process the data by tokenizing it, removing words from it, and then performing stemming and lemmatization. Following that, characteristics are extracted. The use of a rule-based technique and classification to identify sarcasm follows. and a choice is taken after receiving the results of the detection.

1. **Problem statement**

Even for people, let alone computers, it can be challenging to discern sarcasm, a crucial component of communication. Newspaper headlines frequently seem to use sarcasm to draw the reader in. Sarcasm is a witty, frequently ironic comment used to mock or convey disgust. Sarcasm is a cutting, incisive, or harsh phrase; a stinging remark. Sentence detection task classification as "sarcastic" or "non-sarcastic". Lack of intonation and lack of facial emotions make it difficult.

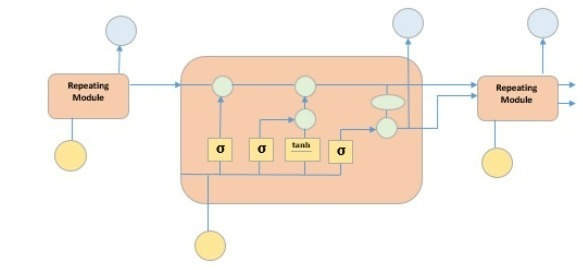
1. **Procedure And Explanation**

Recurrent neural networks of a particular variety can be trained to recognise long-term dependencies in data. This is made possible by the model's recurring module, which consists of four levels that interact with one another.

Above, input is depicted as yellow circles, pointwise operators are depicted as green circles, four neural network layers are depicted as yellow boxes, and cell state is depicted as blue circles. They have the capacity to selectively learn, unlearn, or retain knowledge from each of the units thanks to a cell state, three gates, and an LSTM module.

By allowing only a small number of linear interactions, the cell state in LSTM aids in the uninterrupted flow of information across the units. With the ability to add or delete data from the cell state, each component has an input, an output, and a forget gate. A sigmoid function is used by the forget gate to choose which data from the previous cell state should be disregarded.

Using point-wise multiplication of "sigmoid" and "tanh," the input gate regulates the information flow to the active cell state. The output gate ultimately determines which information should be transferred to the following concealed state.



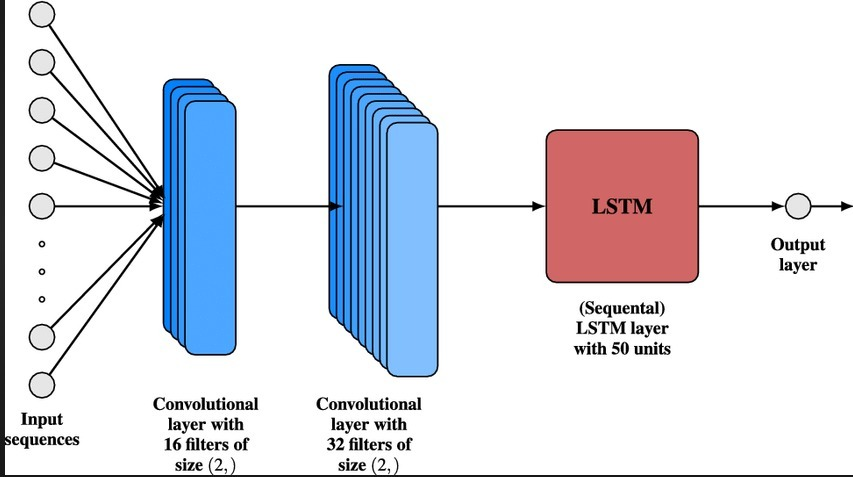
With the use of CNN, a particular kind of neural network model, we can extract more precise representations of picture material. CNN uses the raw pixel data from the image, trains the model, and then automatically extracts the features for improved classification. Traditional image recognition needs you to define the image characteristics.

A deep learning neural network designed for analysing organised arrays of data, such as images, is known as a convolutional neural network, or CNN. Design components like lines, gradients, circles, or even eyes and faces are very successfully recognised by CNN in the input image. Convolutional neural networks are extremely effective for computer vision because of this feature. CNN does not require any pre-processing and may run straight on an underdone picture.

A feed-forward neural network with up to 20 layers is a convolutional neural network.

The convolutional layer, a specific sort of layer, is what gives convolutional neural networks their power. Each of the numerous convolutional layers that make up CNN is capable of identifying increasingly complex forms. These layers are stacked on top of one another. Handwritten digits can be recognised with three or four convolutional layers, while human faces may be distinguished with 25 layers.

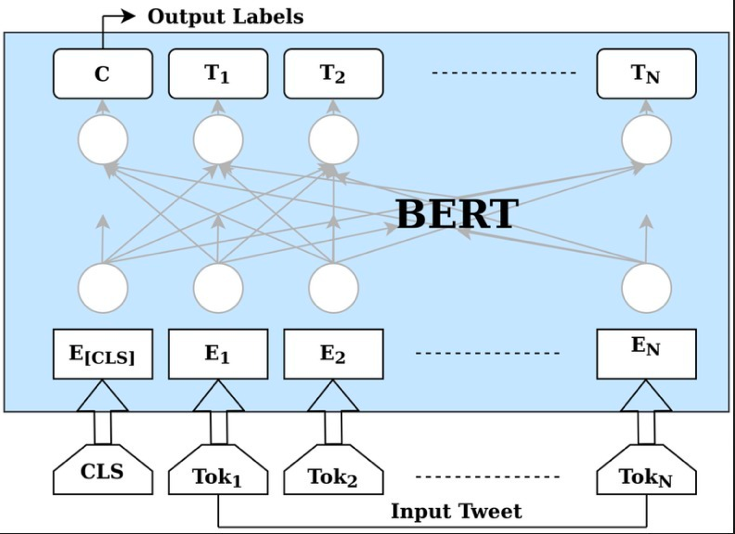
The goal of this field is to enable machines to perceive the environment similarly way humans do and to apply that understanding for a variety of tasks, including picture and video recognition, image inspection and categorization, media reconstruction, recommendation systems, natural language processing, etc.



Open source machine learning software for natural language processing is called BERT (NLP). BERT uses the text around it to construct context, which helps computers grasp the meaning of ambiguous words in text. With the use of question and answer datasets, the BERT framework may be adjusted after being pre-trained on Wikipedia content.

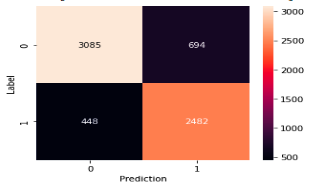
Every output element is related to every input element in the BERT (Bidirectional Encoder Representations from Transformers) deep learning model, and the weights between them are dynamically determined depending on their relationship.[6] In NLP, this is referred to as attention.

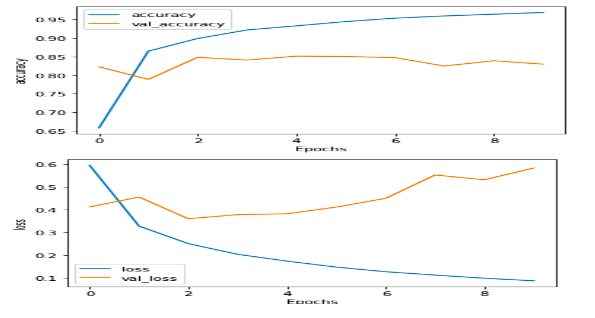
Historically, language models could only sequentially read text input from right to left or from left to right, but not both. BERT is distinctive.

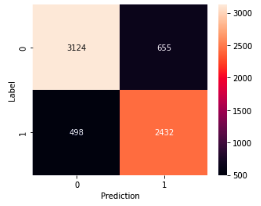


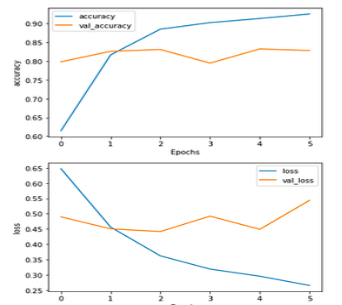
1. **OUTPUT**

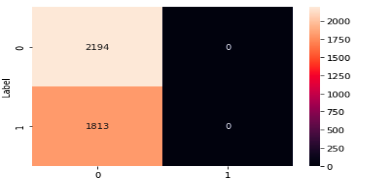
* CNN

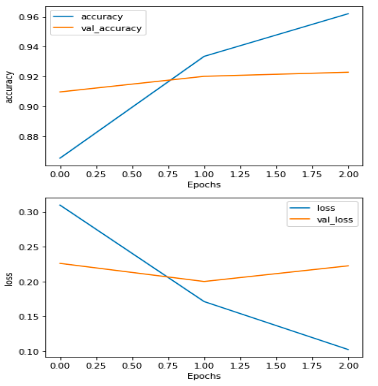
****



* LSTM



* BERT



1. **OBSERVATION AND RESULTS**

Due to the fact that news headlines are produced by professionals using formal language with no spelling problems and casual use, this new dataset offers the following benefits over the pre-existing Twitter datasets. This lowers the density and raises the likelihood of discovering pre-trained embeddings. The headlines acquired, unlike tweets that react to other tweets, stand alone. This would make it easier to identify the genuine sarcastic expressions.

1. **Conclusion**

We developed a variety of Deep Learning models throughout the course of our research and found a training model with 128 layers, dropout, and density once again. The focus mechanism is contrasted with the model, which was mostly built utilising the most modern transformers.

The Bert has a huge embedding dimension, allowing the model to learn more and function more effectively. In the future, we can test sarcasm detection across a wide range of languages.

Adding more intricate layers will help us develop our model even further. Given that it is a basic model, additional optimization is necessary to improve performance.

**8. References**

1. P. Mehndiratta, S. Sachdeva and D. Soni, "Detection of Sarcasm in Text Data using Deep Convolutional Neural Networks", Scientific International Journal for Parallel and Distributed Computing, vol. 18, September 2017.
2. W. Zhang, W. Xu, G. Chen and J. Guo, "A Feature Extraction Method Based on Word Embedding for Word Similarity Computing", Communications in Computer and Information Science, vol. 496, pp. 160-167, January 2014.
3. E. Lunando and A. Purwarianti, "Indonesian Social Media Sentiment Analysis With Sarcasm Detection", Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS, pp. 195-198, September (2013).
4. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Language in Social Media, pp. 30-38, June (2011).
5. R.N Waykole and A.D Thakare, "A Review Of Feature Extraction Methods For Text Classification", International Journal of Advance Engineering and Research Development, vol. 5, April 2018.
6. Paul Mandal and Rakeshkumar Mahto, Deep CNN- LSTM with Word Embeddings for News Headline Sarcasm Detection, 2019.