**SCTR's Pune Institute of Computer Technology**
**Dhankawadi, Pune**

# A PROJECT REPORT ON

- Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data.


**SUBMITTED BY**
41112
Omkar Bhosale
**Under the guidance of**
Prof. D. D. Bhaiya




DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2023-24


**DEPARTMENT OF COMPUTER ENGINEERING**

**SCTR's Pune Institute of Computer Technology**
**Dhankawadi, Pune**
**Maharashtra 411043**

## CERTIFICATE

This is to certify that the SPPU Curriculum-based Mini Project
- Build a machine learning model that predicts the sentiment of the tweets as positive or negative.

**Submitted by**

Omkar Bhosale 41112

has satisfactorily completed the curriculum-based Mini Project under the guidance of Prof. D. D. Bhaiya towards the partial fulfillment of the final year Computer Engineering Semester VII, Academic Year 2023-24 of Savitribai Phule Pune University.

**Date:**

**Place:** PUNE **Name & Sign of Project Guide:**

## Acknowledgment

It gives me great pleasure to present the mini project on - Build a machine learning model that predicts the sentiment of the tweets as positive or negative.

First of all, I would like to take this opportunity to thank my guide Prof. D. D. Bhaiya for giving me all the help and guidance needed. I am grateful for his kind support and valuable suggestions that proved to be beneficial in the overall completion of this project.

I am thankful to our Head of the Computer Engineering Department, Dr. G. V. Kale, for her indispensable support and suggestions throughout the internship work. I would also genuinely like to express my gratitude to the CC, Prof. Samadhan Jadhav, for his constant guidance.

Finally, I would like to thank my mentor, Prof. DD Bhaiya for his constant help and support during the overall process.

**Title:**

- Implement ML model for sentiment analysis.

**Problem Statement:**

Build a machine learning model that predicts the sentiment of the tweets as positive or negative.

**Objective:**

To build a model for classification:

- To analyze its performance on "training.1600000.processed.noemoticon" Dataset.
- To use different ML and Feature Selection concepts to optimize the model's performance.

**Theory:**

Binary Classification-

Binary classification is a fundamental task in machine learning and statistics. It involves categorizing data into one of two possible classes or categories, often referred to as the positive class and the negative class. In binary classification, the goal is to build a predictive model that can determine which class a given data point belongs to based on its features or attributes.

Bidirectional LSTM-

Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) that is well-suited for sequence data, including text. It is designed to overcome the vanishing gradient problem of traditional RNNs, making it capable of capturing long-range dependencies in sequential data. LSTM cells have three gates (input, output, and forget) that control the flow of information within the network, allowing it to store and retrieve information over extended time steps.

Bidirectional LSTM (Bi-LSTM)

Bidirectional LSTMs take the concept of LSTM a step further. They consist of two LSTM layers, one processing the input sequence in a forward direction and the other processing it in reverse. The outputs of both directions are concatenated to capture information from both past and future context. This bidirectional approach enables the network to understand the context of a word or phrase in relation to both preceding and succeeding words.

Key Characteristics bidirectional LSTM:

1. Contextual Understanding: Bi-LSTMs excel at capturing contextual information. They

are able to model dependencies between words that traditional LSTMs might miss, making them particularly effective for tasks like sentiment analysis where context is crucial.
2. Parallel Processing: The bidirectional nature of the network allows for parallel processing of both past and future contexts, which can lead to faster convergence during training.
3. Flexibility: Bi-LSTMs can be used for various NLP tasks, including part-of-speech tagging, named entity recognition, and machine translation, in addition to sentiment analysis.

Advantages of Bidirectional LSTM in Sentiment Analysis:
1. Contextual Understanding: Bi-LSTMs can capture nuanced sentiments by considering both the preceding and succeeding words, providing a deeper understanding of the text.
2. Effective with Long Sequences: They are capable of handling longer text sequences and are more resilient to the vanishing gradient problem compared to traditional LSTMs.
3. State-of-the-Art Performance: Bidirectional LSTMs have achieved state-of-the-art results in various NLP benchmarks and competitions, making them a preferred choice for sentiment analysis.
4. Generalizability: Bi-LSTMs can generalize well across different languages and domains, making them a versatile choice for sentiment analysis tasks.

Confusion Matrix:
A confusion matrix is a fundamental tool in classification tasks that provides a comprehensive summary of the performance of a machine learning model. It is a table that allows us to understand the quality of predictions made by the model by comparing the predicted labels to the true labels. Here's a brief overview of the confusion matrix:
Key Components:
1. True Positives (TP): The number of instances correctly predicted as the positive class. These are the cases where the model correctly identifies positive outcomes.
2. True Negatives (TN): The number of instances correctly predicted as the negative class. These are the cases where the model correctly identifies negative outcomes.
3. False Positives (FP): The number of instances incorrectly predicted as the positive class. These are the cases where the model incorrectly classifies negative instances as positive.
4. False Negatives (FN): The number of instances incorrectly predicted as the negative class. These are the cases where the model incorrectly classifies positive instances as negative.

OUTCOME:
1. Improved Sentiment Classification Accuracy: The use of Bidirectional LSTM in your sentiment analysis model likely resulted in higher accuracy in classifying sentiments

compared to traditional LSTM or other methods. This improvement could be quantified by evaluating the model's performance on a test dataset.

2. Better Handling of Contextual Information: Bidirectional LSTM's ability to capture context from both past and future words in a text could have led to more accurate sentiment predictions. This is especially important for tasks where context plays a significant role, such as determining sentiment in reviews or social media posts.
3. Efficient Training and Faster Convergence: The parallel processing nature of Bidirectional LSTMs might have led to faster convergence during model training. This means that your model likely required fewer epochs to achieve the desired level of accuracy, which can save computational resources and time.
4. Versatility and Adaptability: Bidirectional LSTMs can be applied to a wide range of NLP tasks beyond sentiment analysis. Your project's outcomes may include the versatility of the model, showcasing its potential to excel in various applications, such as part-of-speech tagging, named entity recognition, and more.

**Result:**

The model performed decently with an accuracy of 74 percent on test data.

```
In [32]: from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
         confusion_matrix(test_data.sentiment.to_list(), y_pred_1d)

Out[32]: array([[112749,  47793],
                [ 36327, 123131]])

In [33]: print(classification_report(list(test_data.sentiment), y_pred_1d))

                   precision    recall  f1-score   support

         Negative       0.76      0.70      0.73    160542
         Positive       0.72      0.77      0.75    159458

         accuracy                           0.74    320000
        macro avg       0.74      0.74      0.74    320000
     weighted avg       0.74      0.74      0.74    320000

In [ ]:
```

**Conclusion:**

Bidirectional LSTM is a powerful and effective choice for sentiment analysis due to its ability to capture contextual information from both directions in a sequence. By

considering the advantages and disadvantages, it is essential to assess the specific requirements of a project and the available resources before choosing this model. When used judiciously, Bi-LSTM can yield highly accurate sentiment analysis results, making it a valuable tool in the field of natural language processing.