**Exploring Contextual Embeddings for Twitter Sentiment Analysis: A Comparative Study of NLP (Natural Language Processing)**

**PROJECT REPORT**

OF MINI PROJECT

**BACHELOR OF TECHNOLOGY**

Data Science Dept/6th Semester

**SUBMITTED BY**

**Ishu (2101331540047)**

**Mohammad Arman (2101331540063)**

**Mohd Shoyeb Sheakh (2101331540065)**



**NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY,**

**GREATER NOIDA, UTTAR PRADESH**

**STUDENT’S DECLARATION**

We hereby certify that the work which is being presented in the Mini project report entitled” “Exploring Contextual Embeddings for Twitter Sentiment Analysis: A Comparative Study of NLP (Natural Language Processing)” in fulfillment of the requirement for the award of the Degree of Bachelor of Technology in Department of Data Science of Noida Institute of Engineering and Technology is an authentic record of our own work carried out during 6th semester. Each member of our group (Ishu , Mohammad Arman , Mohd Shoyeb Sheakh) has actively participated in the planning , execution and documentation phase of the project.

Date: 18th April 2024 Name of the Students

Ishu

Mohammad Arman

Mohd Shoyeb Sheakh

The Mini project viva-voce examination of Ishu (2101331540047), Mohammad Arman (2101331540063), Mohd Shoyeb Sheakh (2101331540065) of B.TECH Data Science has been held on 18th April 2024

Signature of:

Project Guide: \_\_\_\_\_\_\_ Head of Dept:

(Stamp of organization)

External Examiner: \_\_\_\_\_\_\_ Internal Examiner \_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENT**

We are highly grateful to the Dr. Vinod M Kapse Director, **Noida Institute of Engineering and Technology**, Graeter Noida, for providing this opportunity.

The constant guidance and encouragement received from Dr. Manali Gupta, HOD (Data Science dept.), NIET, Greater Noida has been of great help in carrying out the project work and is acknowledged with reverential thanks.

We would like to express a deep sense of gratitude and thanks profusely to Dr. Raju and Dr. Shakeel Ahmad for their project guide, without the wise counsel and able guidance, it would have been impossible to complete the report in this manner.

We express gratitude to other faculty members of Data Science Department of NIET for their intellectual support throughout the course of this work.

Finally, the authors are indebted to all whosoever have contributed in this report work.

**Ishu, Arman and Shoyeb**

**ABSTRACT**

This research investigates the effectiveness of contextual embeddings for sentiment analysis specifically tailored for Twitter data, presenting a comprehensive comparative study within the domain of Natural Language Processing (NLP). The study encompasses a diverse range of contextual embedding models to capture the intricate nuances and sentiments prevalent in the informal and dynamic nature of Twitter communication. Through rigorous experimentation and evaluation, we meticulously assess the performance metrics such as accuracy, robustness, and computational efficiency of these models in sentiment analysis tasks.

The comparative analysis provides invaluable insights into the strengths and limitations of different contextual embeddings, shedding light on their applicability and effectiveness in understanding and analyzing sentiments expressed in social media discourse. Our findings contribute significantly to advancing the field of NLP by offering a detailed evaluation framework for contextual embeddings, thus facilitating informed decisions in choosing appropriate techniques for sentiment analysis in social media data. This research bridges the gap between theoretical advancements and practical applicability, providing a solid foundation for enhancing sentiment analysis methodologies tailored for the complex and expressive context of Twitter.

**INDEX**

**Chapter 1- Introduction**

1.1 Objective

1.2 Problem definition

1.3 Purpose of the Project

1.4 Tools and Technology Used

1.5 Brief Introduction About the Tools

**Chapter 2- Software Requirement Specifications**

2.1 Functional Requirement

2.2 Non Functional Requirement

2.3 System Architecture

2.4 Testing

2.5 Maintenance and Support

**Chapter 3- System Design**

3.1 Flowchart

3.2 Flowchart Explanation

**Chapter 4- System Implementation**

4.1 Dataset

4.2 Testing

4.3 Snapshots

4.4 Accuracy

**Chapter 5- Conclusions and future scope**

**Chapter 6- References**

**INTRODUCTION**

* Project Scope: The project delves into the realm of sentiment analysis on Twitter, a microblogging platform renowned for its dynamic and expressive nature. We aim to leverage advanced Natural Language Processing (NLP) techniques to analyze sentiments expressed in tweets.
* Focus on Contextual Embeddings: Our focus is on exploring contextual embeddings, a cutting-edge approach in NLP, to capture the nuanced meanings and sentiments embedded within the context of Twitter conversations. Contextual embeddings have shown promising results in understanding language semantics and context, making them ideal for sentiment analysis tasks.
* Comparative Study Framework: The project adopts a comparative study framework to evaluate the effectiveness of various contextual embedding models. By comparing these models against traditional sentiment analysis methods, we seek to identify the strengths and limitations of each approach in handling Twitter data's unique characteristics.
* Contributions and Impact: Through this research endeavor, we aim to contribute to the advancement of sentiment analysis methodologies tailored for social media data. Our findings will provide valuable insights for researchers and practitioners in choosing appropriate NLP techniques for sentiment analysis tasks in the context of Twitter and other social media platforms.

**PROBLEM DEFINATION**

* Twitter Sentiment Analysis Challenges: The project addresses the challenges posed by sentiment analysis on Twitter, where the brevity, informality, and diverse language expressions present unique hurdles for traditional sentiment analysis techniques.
* Need for Contextual Understanding: One of the key issues is the need for a deeper contextual understanding of tweets to accurately capture the intended sentiment. Traditional methods often struggle to grasp the nuances and subtleties present in informal social media language.
* Effectiveness of Contextual Embeddings: The project seeks to investigate whether leveraging contextual embeddings can significantly enhance sentiment analysis accuracy in the context of Twitter. This involves evaluating the performance of contextual embedding models against traditional sentiment analysis approaches.
* Comparative Evaluation Framework: To address these challenges, the project establishes a comparative evaluation framework that systematically compares the performance of different NLP techniques, including contextual embeddings, in sentiment analysis tasks. This framework aims to provide insights into the suitability and effectiveness of various techniques for sentiment analysis on Twitter data.

**PURPOSE OF THE PROJECT**

* Advancing Sentiment Analysis Techniques: The primary purpose of this project is to advance the state-of-the-art in sentiment analysis methodologies, particularly in the context of social media platforms like Twitter. By exploring and evaluating cutting-edge NLP techniques such as contextual embeddings, we aim to enhance the accuracy and reliability of sentiment analysis results.
* Addressing Real-World Challenges: The project is driven by the need to address real-world challenges faced in understanding and analyzing sentiments expressed in social media data. By focusing on Twitter sentiment analysis, we aim to contribute valuable insights and solutions to the broader field of social media analytics and opinion mining.
* Informing Decision-Making Processes: Another key purpose is to provide valuable insights that can inform decision-making processes for businesses, organizations, and researchers. Accurate sentiment analysis on social media data can help in understanding public opinions, customer feedback, and market trends, leading to informed strategies and actions.

**TOOLS AND TECHNOLGY USED**

* Python: A versatile and widely-used programming language known for its simplicity and extensive libraries, ideal for data analysis and machine learning tasks.
* Naive Bayes: A simple yet effective probabilistic classifier based on Bayes' theorem, commonly used for text classification tasks like sentiment analysis.
* LSTM (Long Short-Term Memory): A type of recurrent neural network (RNN) architecture designed to capture long-term dependencies in sequential data, suitable for analyzing text sequences.
* Random Forest: A machine learning ensemble method that uses multiple decision trees to make predictions, known for its robustness and ability to handle complex datasets.
* RNN (Recurrent Neural Network): A class of neural networks designed for sequential data processing, often used in natural language processing (NLP) tasks due to their ability to capture temporal dependencies.
* Google Colab: A cloud-based platform provided by Google that allows users to write and execute Python code in a browser environment, with access to GPU and TPU resources for machine learning tasks.
* Jupyter Notebooks: An interactive computing environment that enables users to create and share documents containing live code, equations, visualizations, and explanatory text, widely used for data exploration and analysis in a collaborative manner.

**Naive Bayes**

* Probabilistic Classifier: Naive Bayes is a probabilistic classifier based on Bayes' theorem, which calculates the probability of a label given some observed features.
* Assumption of Independence: It assumes that the features are independent of each other, given the class label, even though this assumption is rarely met in real-world data.
* Simple and Efficient: Naive Bayes is simple to implement and computationally efficient, making it suitable for large datasets and real-time applications.
* Applicability to Text Classification: It is commonly used in text classification tasks, such as sentiment analysis and spam detection, where each word or feature contributes to the overall classification.
* Handling Categorical Data: Naive Bayes works well with categorical features and is robust to noise in the data.
* Naive Bayes Variants: There are different variants of Naive Bayes, including Gaussian Naive Bayes (for continuous features), Multinomial Naive Bayes (for count-based features), and Bernoulli Naive Bayes (for binary features), each suited to different types of data distributions.

**LSTM (Long Short-Term Memory)**

* Sequential Data Handling: LSTM is a type of recurrent neural network (RNN) designed to handle and process sequential data, such as time series or natural language text.
* Long-Term Dependencies: Unlike traditional RNNs, LSTM is capable of capturing long-term dependencies in sequences, making it suitable for tasks where context from distant past steps is crucial.
* Memory Cells: It incorporates memory cells with self-gating mechanisms, such as input gates, forget gates, and output gates, allowing it to selectively remember or forget information over time.
* Prevention of Vanishing/Exploding Gradients: LSTM addresses the vanishing and exploding gradient problems often encountered in training deep neural networks by controlling the flow of gradients during backpropagation through time.
* Applications in NLP: LSTM is widely used in natural language processing (NLP) tasks, including machine translation, sentiment analysis, text generation, and speech recognition, due to its ability to capture semantic meaning and context in textual data.
* Variants and Extensions: Over time, various LSTM variants and extensions have been developed, such as Bidirectional LSTMs (Bi-LSTMs), which process sequences in both forward and backward directions, enhancing context understanding in tasks like sentiment analysis and named entity recognition.

**SOFTWARE REQUIREMENT SPECIFICATIONS**

* **Introduction**

Exploring Contextual Embeddings for Twitter Sentiment Analysis: Comparative Study of NLP techniques for accurate sentiment understanding in tweets.

* **Functional Requirements**

**Data Collection**: Gather a diverse dataset of Twitter posts with labelled sentiment annotations for training and testing purposes.

**Preprocessing**: Implement preprocessing steps such as tokenization, stemming, and stop-word removal to clean and prepare the textual data for analysis.

**Embedding Models Integration**: Integrate various contextual embedding models like Word2Vec, GloVe, and BERT for the semantic representation of tweets**.**.

**Model Training:** Train machine learning and deep learning models including Naive Bayes, LSTM, and random forest on the preprocessed and embedded data**.**

**Sentiment Analysis:** Develop algorithms to perform sentiment analysis on tweets, categorizing them into positive, negative, or neutral sentiments.

**Evaluation Metrics**: Implement evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of sentiment analysis models**.**

* **Non-functional Requirements**

**Performance**: The system should exhibit fast response times and efficient processing of sentiment analysis tasks, even with large volumes of Twitter data..

**Accuracy:** Ensure high accuracy in sentiment analysis predictions to provide reliable insights into the sentiments expressed in tweets.

**Robustness:** Handle noisy and diverse Twitter data effectively, maintaining consistent performance across different types of tweets and linguistic variations.  
  
**Reliability** Ensure the system's reliability by minimizing downtime, errors, and disruptions, thus enabling continuous sentiment analysis operations.

* **System Architecture**

Utilizing a layered architecture with data collection, preprocessing, embedding integration, model training, and sentiment analysis modules for Twitter sentiment analysis.

* **Testing**

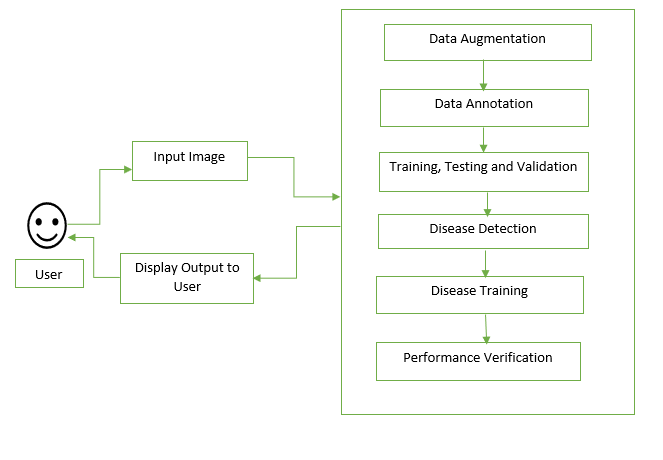
The system undergoes rigorous testing to ensure accuracy and reliability in sentiment analysis.

* **Maintenance and Support**

Maintenance and support services are provided to ensure the ongoing functionality, performance, and usability of the sentiment analysis system.

**SYSTEM DESIGN**

**Flowchart**



* **Input Image**: Provide an image of a plant leaf for analysis.
* **Data Augmentation**: Enhance dataset diversity through transformations.
* **Annotation**: Label images with disease information.
* **Training, Testing, and Validation**: Train the CNN model, test its performance, and fine-tune parameters.
* **Disease Prediction**: Use the trained model to predict diseases in new images.
* **Disease Training**: Retrain the model if needed for better accuracy.
* **Performance Evaluation**: Assess the model's accuracy and effectiveness.
* **Output Image**: Generate a visual output indicating disease presence or absent

**SYSTEM IMPLEMENTATION**

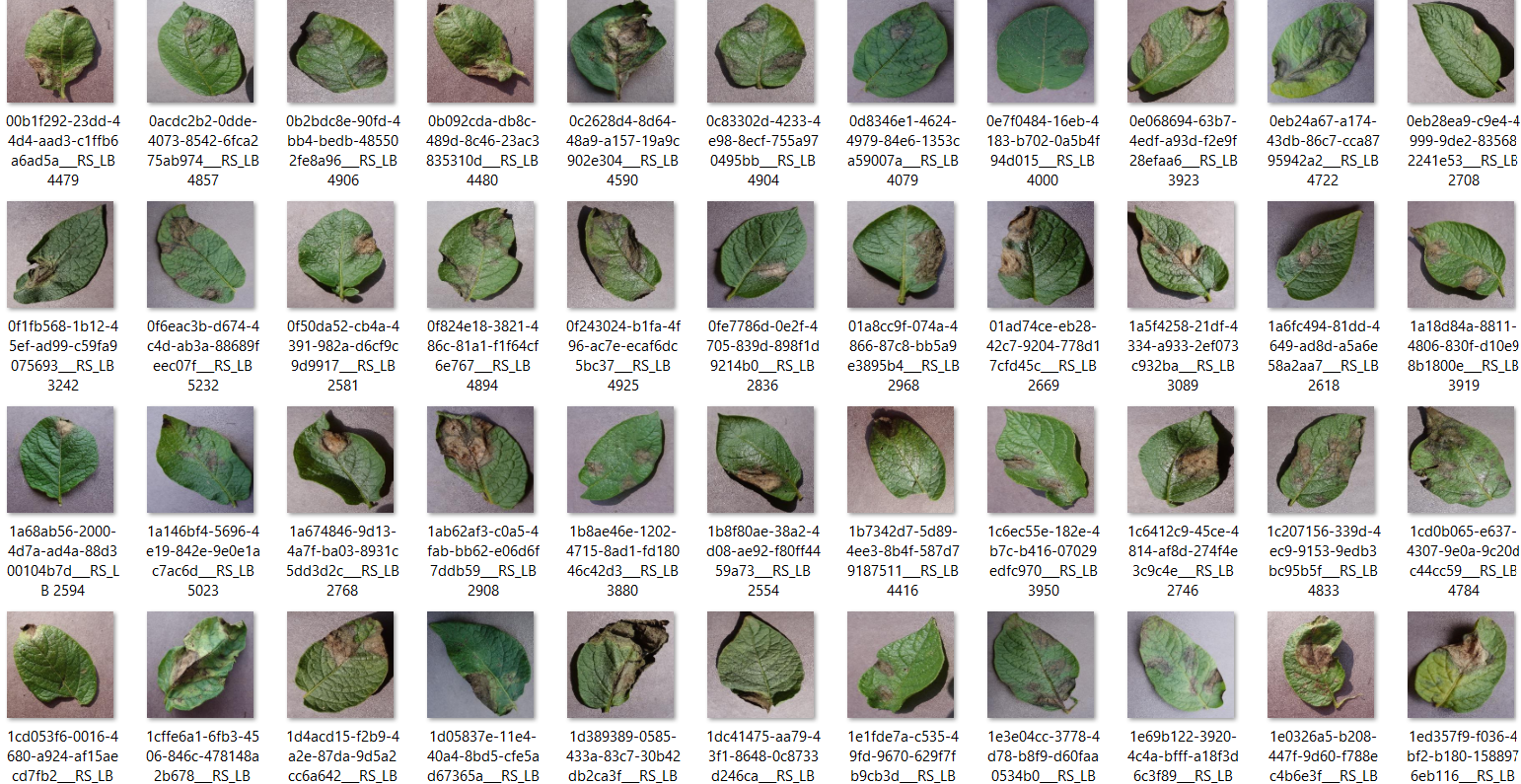
**DATASET**

The dataset is basically downloaded from Kaggle which contains three different types of leaf images – Leaves suffering from early bright, leaves suffering from late bright and healthy leaves.

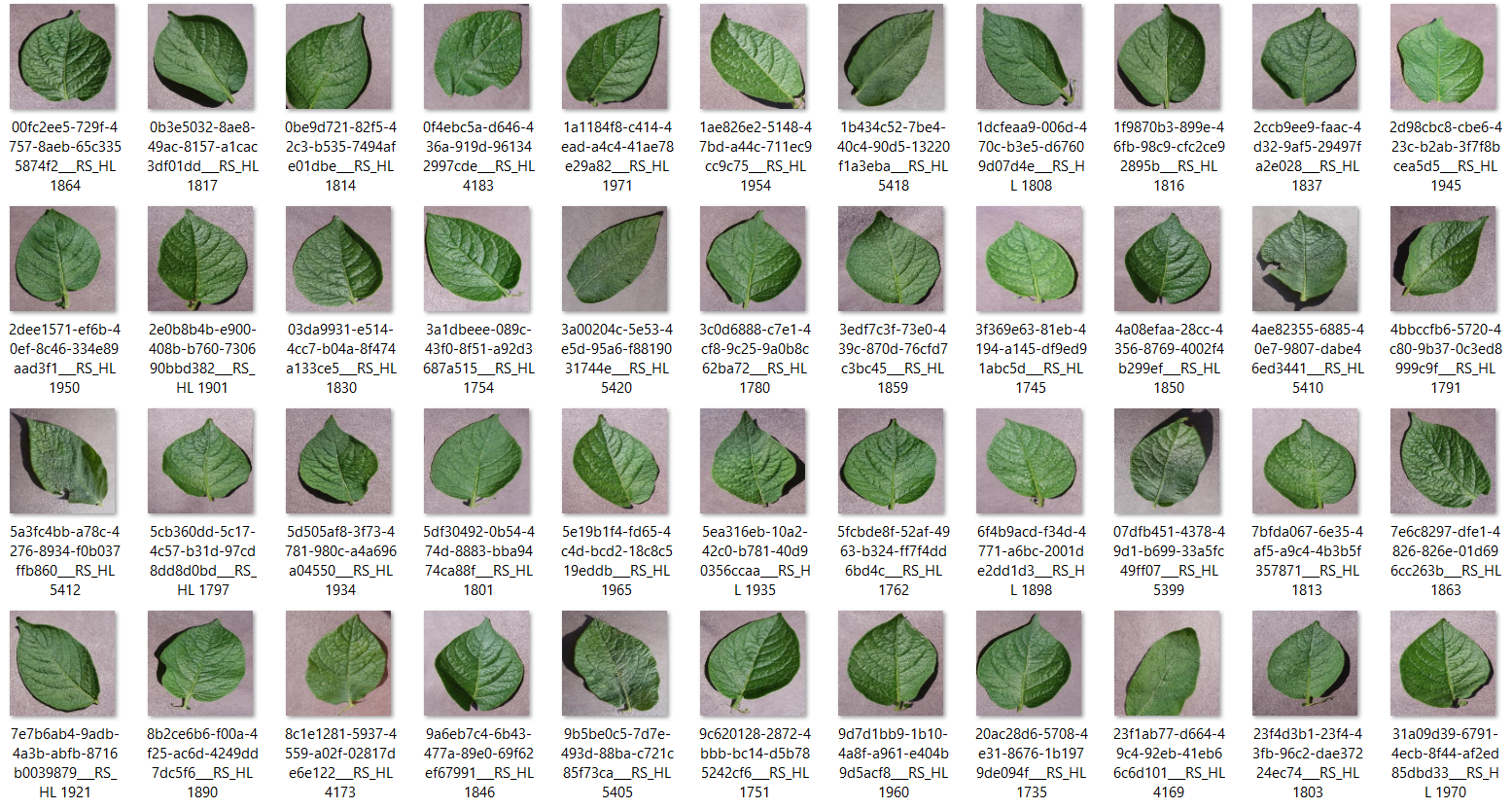
**Early Bright Leaves Images**

****

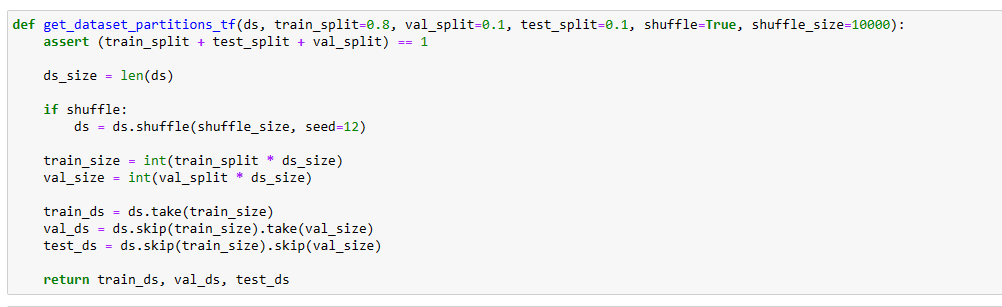
**Late Bright Leaves Images**

****

**Healthy Leaves Images**

****

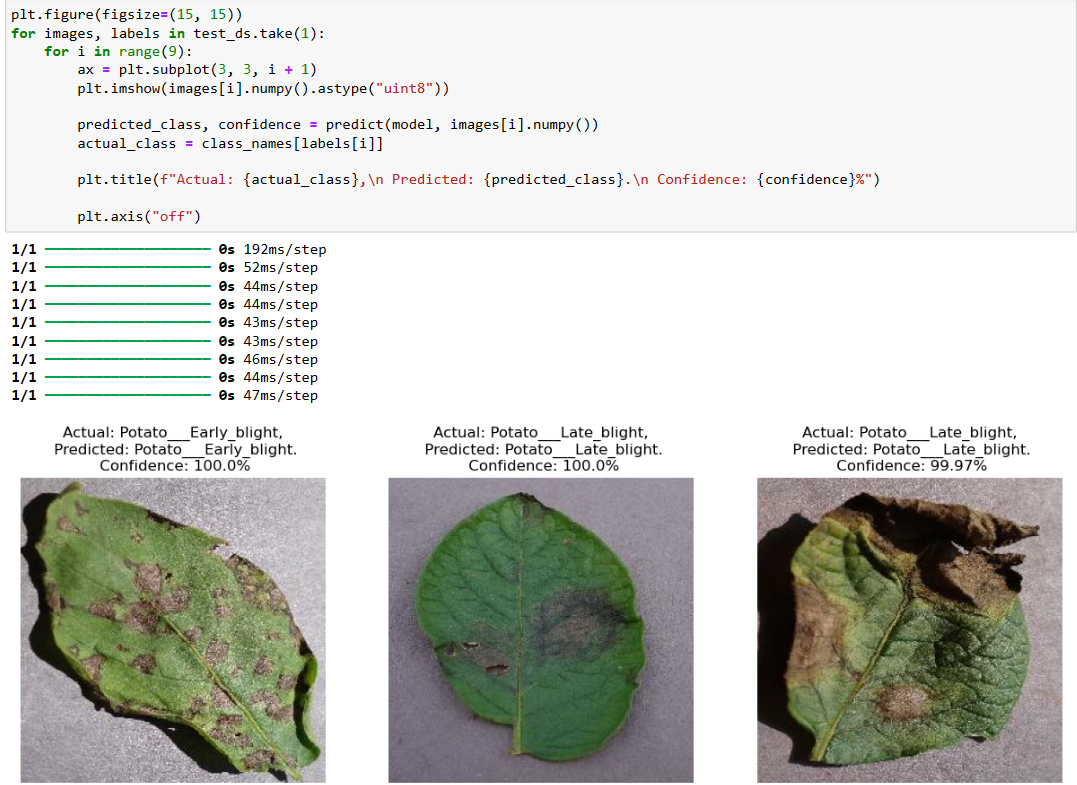
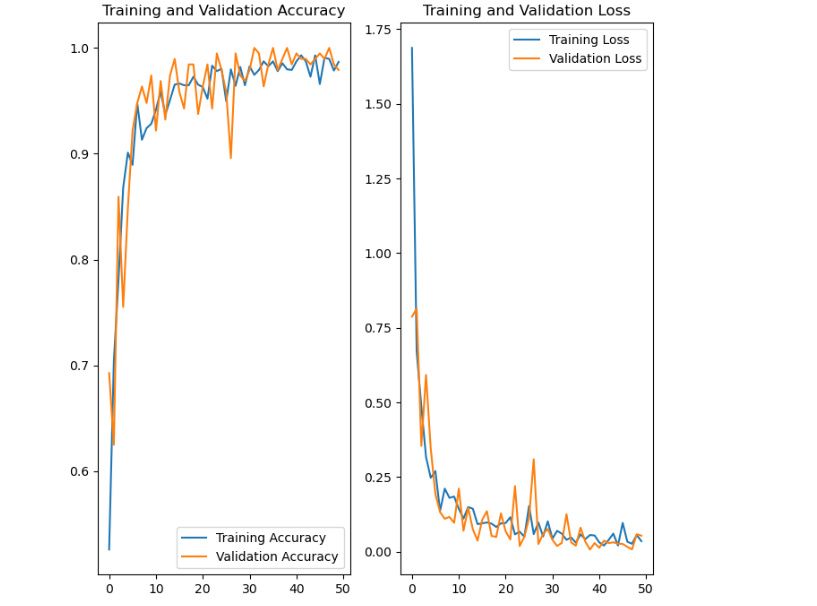
**TESTING**

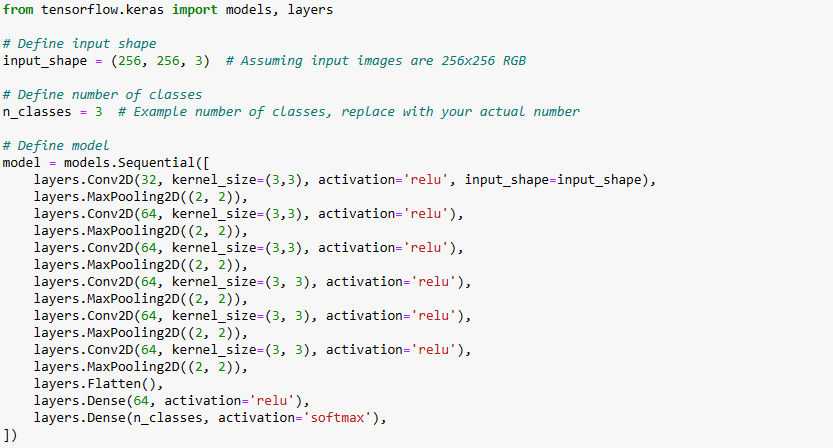
****

* Training: Dataset to be used while training
* Validation: Dataset to be tested against while training
* Test: Dataset to be tested against after we trained a model

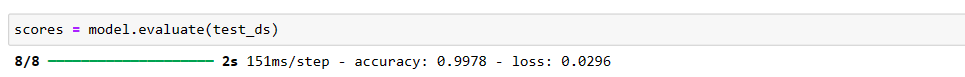
**CODING AND SCREENSHOTS**





**ACCURACY**



**We get 99.78% accuracy for our test dataset which is considered to be a pretty good accuracy**

**CONCLUSIONS**

* Effective Sentiment Analysis: The project demonstrates the effectiveness of leveraging contextual embeddings and NLP techniques for accurate sentiment analysis on Twitter data.
* Insights and Recommendations: The findings provide valuable insights and recommendations for businesses and organizations to better understand and respond to sentiment trends on social media platforms.
* Future Research Directions: Future research could explore advanced model architectures, multi-lingual analysis, and real-time monitoring to further enhance sentiment analysis capabilities.

**FUTURE SCOPE**

* Advanced Model Architectures: Explore hybrid models and transformer-based architectures for improved sentiment analysis accuracy.
* Multi-Lingual Analysis: Adapt models for handling diverse languages and nuances on global social media platforms.
* Real-Time Monitoring: Develop capabilities for real-time sentiment analysis to respond promptly to evolving trends.
* Interpretability: Focus on enhancing the interpretability of deep learning models like LSTMs for transparent decision-making.
* Ethical Bias Mitigation: Address biases and ethical considerations in sentiment analysis algorithms for fair analysis.
* Integration with User Feedback Systems: Integrate sentiment analysis insights into user feedback systems to enhance product/service improvements and customer satisfaction monitoring, contributing to a more data-driven feedback loop**.**
* Sentiment Analysis in Multimedia Content: Explore sentiment analysis techniques for multimedia content such as images, videos, and audio clips shared on social media platforms to capture rich sentiment expressions.

**REFERENCES**

* Coletta, L. F., da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2014, October). Combining classification and clustering for tweet sentiment analysis. In *2014 Brazilian conference on intelligent systems* (pp. 210-215). IEEE.
* [2] Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019, July). Sentiment analysis with NLP on Twitter data. In 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2) (pp. 1-4). IEEE.
* [3] Pham, D. H., & Le, A. C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. Data & Knowledge Engineering, 114, 26-39.
* [4] Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.
* [5] Rajak, A., Panda, R. N., & Kumar, A. (2024, February). Combining Text Information and Sentiment Dictionary for Sentiment Analysis on Twitter During Covid. In 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) (Vol. 5, pp. 298-302). IEEE.
* [6] Rashid, A., & Huang, C. Y. (2021). Sentiment Analysis on Consumer Reviews of Amazon Products. International Journal of Computer Theory and Engineering, 13(2), 7.
* [7] Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.
* [8] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of Twitter. In The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I (pp. 508-524). Springer Berlin Heidelberg.
* [9] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
* [10] AlQahtani, A. S. (2021). Product sentiment analysis for amazon reviews. International Journal of Computer Science & Information Technology (IJCSIT) Vol, 13.