# MACHINE LEARNING PROJECT

## NEXT WORD PREDICTION USING LSTM

**Submitted By**

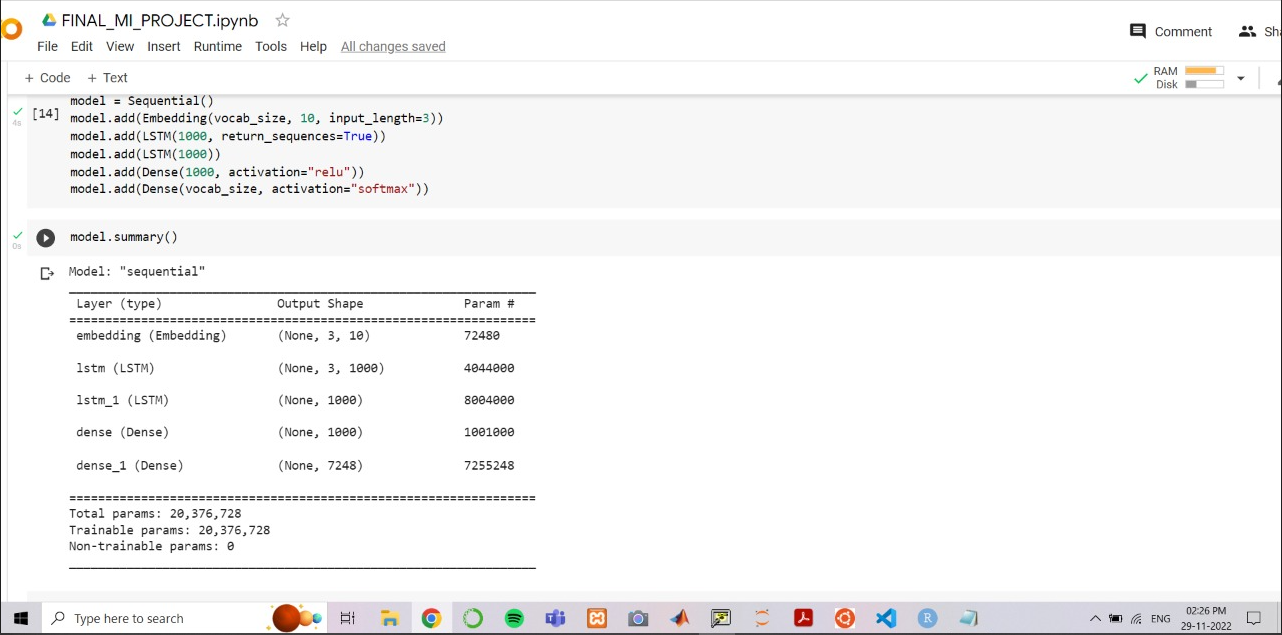
**Venkata Sathvik Adepalli: SRN: PES1UG20CS493**

**Vrushank G: SRN: PES1UG20CS516**

**Y Srinivas: SRN: PES1UG20CS517**

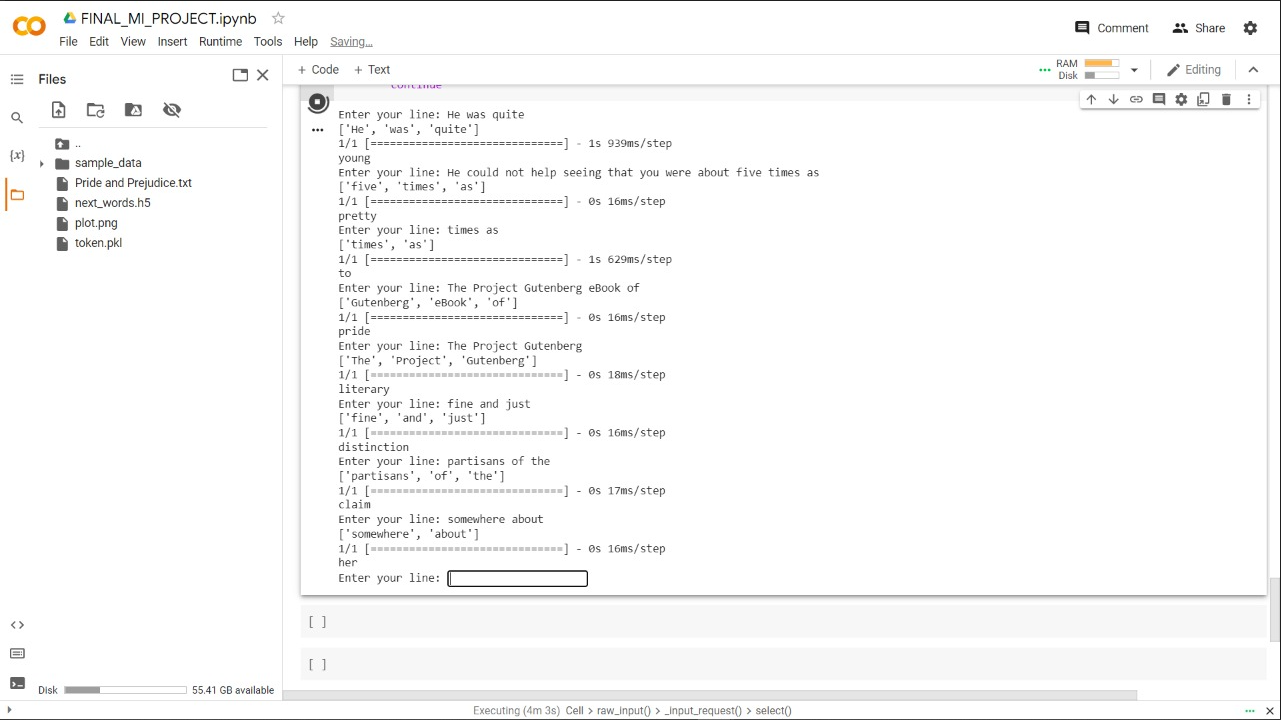
**Prediction:**

For the prediction notebook, we will load the tokenizer file which we have stored in the pickle format. We will then load our next word model which we have saved in our directory. We will use this same tokenizer to perform tokenization on each of the input sentences for which predictions are to be made. After this step, we can proceed to make predictions on the input sentence by using the saved model.



**Observation:**

We are able to develop a high-quality next word prediction for the metamorphosis dataset and reduce the loss significantly in about 150 epochs. The next word prediction model which we have developed is fairly accurate on the provided dataset. The overall quality of the prediction is good. However, certain pre-processing steps and certain changes in the model can be made to improve the prediction of the model.



**LITERATURE SURVEY:**

***PAPER 1: Predicting next Word using RNN and LSTM cells: Stastical Language Modeling***

**Date of Release: 15-17 November 2019**

**Link: https://ieeexplore.ieee.org/document/8985885**

**Synopsis:**

Language Modeling is defined as the operation of predicting next word. It is considered as one of the basic tasks of Natural Language Processing(NLP) and Language Modeling has several applications. In this research paper, the assorted potentialities for the efficient utilization of language models in structured document retrieval are mentioned. A tree-based generative language model for ranking documents and parts has been used here. Nodes within the tree correspond to different document parts like titles, paragraphs and sections. At every node within the document tree, there's a well-defined language model. The language model for a leaf node is predictable directly from the text within the document part related to the node. Inner nodes within the tree are predictable employing a linear interpolation among the various youngster nodes. The paper additionally describes how some common structural queries would be satisfactorily described inside this model.

**Methodology of this Paper:**

Neural-networks in predicting the current or next word has been discussed. Language modeling is one of the important tasks of NLP. First it will make an environment to simulate typing. Then it will save the user's keystrokes that the user types. Next it will use deep layered convolution neural networks to model the language. And lastly, employing a character to word model here can serve the purpose. As of now, it is planned to predict the next word on the basis of 40 preceding characters. As far as autocorrect is concerned, RNNs are used to model grammatical and spelling errors on the basis of training data A recurrent neural network, at its most elementary level, is solely a sort of densely connected neural network. The key distinction to traditional feed forward networks is that the introduction of your time-specially, the output of the hidden layer in an exceedingly recurrent neural network is fed back into itself.

**Conclusion:**

In this report, summarized work in Language Model has been presented. Supported by totally different language Models, the NLM will solve the matter of information exiguity, and that they are prepared to catch the logical data during a range from sub-word-level to corpus-level.

***PAPER 2: Next Word Prediction Using Deep Learning: A Comparative Study***

**Date of Release: 27-28 January 2022**

**Link:** [**https://ieeexplore.ieee.org/document/9734151**](https://ieeexplore.ieee.org/document/9734151)

**Synopsis:**

Deep learning is a subclass of machine learning, it mimics the functionality of the human brain the way it processes and creates pattern in the facts for choice making. It is basically an AI function that has networks capable of learning unsupervised data that is shapeless. The succeeding word forecast is performed on dataset consisting of texts. Next Word Prediction is an application of NLP (Natural Language Processing). It is also known as Language Modelling. Basically it is the process of predicting the next word in a sentence. It has many applications which are used by most of us such as auto-correct which is mostly used in Emails / Messages; it has also it’s usage in MS Word or google search where forecasts the next word based on our search history. In this work we have studied NLP, different deep learning techniques such as LSTM, BiLSTM and performed a comparative study. We found satisfactory results in BiLSTM and LSTM. The accuracy received using BiLSTM and LSTM are: 66.1% and 58.27% respectively.

**Methodology of this Paper:**

LSTM is a version of RNN which is a sequential network that allows information to be persistent. It solves the problem of vanishing gradient in RNN. Basically RNN is used for persistent memory. As humans remember what is going to happen further when watching a video or while reading a book they know what would be happening further; similarly RNNs also work in the same way and remember the previous information to use it for processing the current input. the major drawback of PNN is that because of the vanishing gradient they cannot remember long term dependencies. and so does lstm is designed to avoid this problem.

BiLSTM or bidirectional LSTM is version of RNN and are improved version of LSTM. It is designed to improve the model’s performance in sequence classification problems On the input sequence, it trains 19 two LSTMs rather than one LSTM. The input sequence is used for the first training, and a reversed duplicate of the input sequence is used for the second. It leads to 60 quicker and more thorough learning of the problem, as well as providing new context to the network. The primary difference between BiLSTM and LSTM is that LSTM will maintain information from input that has previously gone through its hidden state, whereas bidirectional will run the inputs in two directions, one from the past to the future and the other from the future to the past. In a backwards-running LSTM, information from the future is only preserved, but by combining the two hidden states, information from the past and future may be preserved at any point in time.

**Conclusion:**

There is high overfitting. Hence it also tells that taking more datasets might solve this problem so that training is done on a large number of datasets.

***Paper 3: Next Word Prediction in Hindi Using Deep Learning Techniques***

**Date of Release: 26th September 2019**

**Link:https://ieeexplore.ieee.org/document/8971796**

**Synopsis:**

Natural Language Generation (NLG) focuses on the generation of natural, human-interpretable language. This study proposes a novel methodology to predict the next word in a Hindi sentence. By predicting the next word in a sequence, the number of keystrokes of the user can be reduced. Two deep learning techniques namely Long Short Term Memory (LSTM) and Bi-LSTM have been explored for the task of predicting next word and accuracy of 59.46% and 81.07% was observed for LSTM and Bi-LSTM respectively. This approach may be used for various NLG tasks like story auto-completion, sentence autocompletion, etc.

**Methodology of this Paper:**

The process starts with the cleaning of the dataset. Then, the dictionary of unique words is created by splitting the words in the dataset and filtering the unique words which constitute to a dictionary. The input file is parsed with the iterator, and unique words are collected. Succeeding, the unique words in the dictionary are mapped to indices. This indicates that the words are not easily processed by neural networks in machine learning, and it is important to map it to the indices, which are easy for neural networks to process.

**Conclusion:**

A model has been proposed in this paper to predict the next word in Hindi, given a sequence of at least six words using LSTM and Bi-LSTM Neural Networks. For both the models softmax activation function and categorical cross entropy loss function was used. Softmax activation function is a mathematical function that is used to obtain the probability that a word will be predicted on the basis of the score for each word obtained from the neural network model thus obtaining the probability distribution of all the words present in the dictionary

***PAPER 4: Predicting the next word using the Markov Chain Model according to profiling personality***

**Date of Release: 29th July 2022**

**Link:https://assets.researchsquare.com/files/rs1879234/v1\_covered.pdf?c=1659108973**

**Synopsis:**

Understanding human data has been the focus of philosophers and scientists. Social media platforms encourage people to be creative and share their personal information. By analyzing data, we will be able to identify people’s personalities and information that is also important to specific profils. The aim of this paper is to propose an approach that predicts the next word during writing a sentence based on the user’s personality.

**Methodology of this Paper:**

To completely understand their users’ activity, several research initiatives are working on gathering metadata from their products and platforms. Its recommendations are based on past research done by other users and even by their regions.

Our approach has two main points. First point (1: Extract Big Five Personalities): focus on a real database from one of the social networks, then extract all the posts filtered by their users. Through data profiling we need to extract an approximate personality score of a user from an existing library and store the results in our repository. Second point (2: calculate distance & predict next word), predict the next word for the new post based on the user’s personality, and the algorithm can be adapted to suggest words from another personality.

**Conclusion:**

In this paper, we profile user personalities based on the Big Five Model to predict the next word when writing a new tweet using the Markov Chain Model. Through our approach, we found that there is a big difference between predicting the next word with and without profiling the user’s personality.

***PAPER 5: Next Words Prediction Using Recurrent NeuralNetworks***

**Date of Release: August,2021**

**Link:https://www.itmconferences.org/articles/itmconf/pdf/2021/05/itmconf\_icacc2021\_03034.pdf**

**Synopsis:**

Natural Language Processing (NLP) is a significant part of artificial Intelligence, which incorporates AI, which contributes to finding productive approaches to speak with people and gain from the associations with them

**Methodology of this Paper:**

This paper represents an exploration of the contributions that have already been made in the academic field. [1] multi-window convolution(MRNN) algorithm is implemented, also they have created residual-connected minimal gated unit(MGU) which is short version of LSTM in this CNN try to skip few layers while training result in less training time and they have good accuracy by far using multiple layers of neural networks can cause latency for predicting n numbers of words. This paper used RNN algorithm and also they have used GRU (another form of RNN) for code completion problem as RNN helps to predict next code syntax for users

**Conclusion:**

Understanding the paragraph using machine learning algorithms like RNN can help soon to understand and frame paragraphs and stories on their own. As more data we can train the model which will reevaluate the weights to understand the core features of paragraphs/sentences to predict good results.

***Paper 6: NEXT WORD PREDICTION AND CORRECTION SYSTEM USING TENSORFLOW***

**Date of Release: May 2017**

**Link:https://www.academia.edu/45187683/NEXT\_WORD\_PREDICTION\_AND\_CORRECTION\_SYSTEM\_USING\_TENSORFLOW**

**Synopsis:**

As per the title, the paper presents the concept on language processing. Natural language processing is a field of science and engineering where humans and the computers are interacted. With respect to the computer system, An Artificial Intelligence is also a field of science and technology where the computer system should act giving their feedback/reviews/ comments in social media or other medium with shortcuts and spelling mistakes. So goal is to predict that mis-spelt word and correct it with word to vector and recurrent neural networks models using tensorflow.

**Methodology of this Paper:**

The proposed design consists of four blocks namely Database, Computer system, Tensorflow model, Training System. The database is having thousands of sentences which need to be processed at a faster rate so that the performance of the system should increase. Second block is computer system where the system should support the requirements for the TensorFlow model to communicate. Training model invokes tensorflow whenever the it calls by the computer system. Main goal of this proposed design is to achieve misspelt prediction and correction in a sentence.

**Conclusion:**

This paper presents how the system is predicting and correcting the next/target words using some mechanisms and using TensorFlow closed loop system, the scalability of trained system can be increased and using perplexity concept the system will decide that the sentence is having more mis-spelt words and the performance of the system can be increased

**REFERENCES:**

[1] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.

[2] Sahoo, Abhaya Kumar, Chittaranjan Pradhan, and Himansu Das. "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making." In Nature inspired computing for data science, pp. 201- 212. Springer, Cham, 2020.

[3] Prajapati, Gend Lal, and Rekha Saha. "REEDS: Relevance and enhanced entropy based Dempster Shafer approach for next word prediction using language model." Journal of Computational Science 35 (2019): 1-11.

[4] Ambulgekar, Sourabh, Sanket Malewadikar, Raju Garande, and Bharti Joshi. "Next Words Prediction Using Recurrent NeuralNetworks." In ITM Web of Conferences, vol. 40, p. 03034. EDP Sciences, 2021.

[5] Stremmel, Joel, and Arjun Singh. "Pretraining federated text models for next word prediction." In Future of Information and Communication Conference, pp. 477-488. Springer, Cham, 2021.