

NEXT WORD PREDICTION

PROJECT REPORT

18CSE484T – DEEP LEARNING

(2018 Regulation)

III Year/ VI Semester

Academic Year: 2022 -2023

By

BOTCHA SANDHYA SRI (RA2011026010281)

VARSHA S (RA2011026010286)

URVI HIRANI(RA2011026010293)

DEEKSHA(RA2011026010301)

Under the guidance of

Dr. M. KIRUTHIKA

Assistant Professor Department of Computational

Intelligence



FACULTY OF ENGINEERING AND TECHNOLOGY

SCHOOL OF COMPUTING

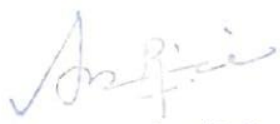
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR, KANCHEEPURAM

MAY 2023

BONAFIDE

This is to certify that **18CSE484T–DEEP LEARNING project report** titled “**NEXT WORD PREDICTION**” is the bonafide work of BOTCHA SANDHYA SRI(RA2011026010281), VARSHA(RA2011026010286), URVI HIRANI(RA2011026010293), AKKATI DEEKSHA(RA2011026010301) who undertook the task of completing the project within the allotted time.



Signature of HOD
Dr. R. Annie Uthra
Professor and Head
Department of CINTEL
SRM Institute of Science
and Technology



Signature of the Guide
Dr. M. KIRUTHIKA
Assistant Professor
Department of CINTEL
SRM Institute of Science and Technology

ABSTRACT

Numerous technologies are used to make countless word predicting applications that make typing easier. These technologies that also facilitates typing on a mobile device by suggesting words the end user may wish to insert in a text field. It also increases writing fluency, allowing students to generate more writing skills. It also helps the system for typing of free text. It also helps in forming the sequency of well-structured language. It is used to develop highly recommended applications like Grammarly etc. As well as it is used in user enters that letter of required word, The system displays a list of the most probable words that could appear in the position. It can also predict words various language like Hindi, Spanish etc. The main objective is to predict the next word in a sentence. This study involves N-gram modelling, convolution neural network, recurrent neural networks and some of deep learning techniques which enables the feature information towards predicting the next word more fully. This work also includes results, analysis and techniques. By studying all these we can easily predict the next coming word. Next word prediction is a significant task in natural language processing that aims to anticipate the most probable word to follow a given context. It plays a vital role in various applications such as text generation, auto-completion, and virtual assistants. In recent years, language models based on deep learning techniques, particularly transformer-based models like GPT-3, have demonstrated remarkable advancements in next word prediction tasks. Moreover, we discuss the evaluation methodologies employed to assess the performance of next word prediction models, including perplexity and accuracy measures. Additionally, we examine the impact of various factors on next word prediction performance, such as context window size, training corpus size, and domain-specific fine-tuning. We highlight the significance of diverse and representative training data to enhance the model's ability to handle various language patterns and domains effectively. Next word prediction is a text generation task that involves predicting the most probable word that will follow a given sequence of words. This task has numerous practical applications in fields such as natural language processing, speech recognition, and machine translation. To achieve accurate next word prediction, various techniques such as n-gram models, neural language models, and contextual embedding models have been developed. These models utilize statistical, neural network, and machine learning techniques to learn from large amounts of text data and generate predictions based on the learned patterns. Effective next word prediction can significantly improve the user experience in various text-based applications such as search engines, virtual assistants, and chatbots. By understanding the current advancements, challenges, and potential applications of next word prediction models, this paper aims to provide a comprehensive foundation for researchers, developers, and practitioners interested in leveraging language models to improve text completion and user experience in nlp tasks.

INTRODUCTION

Natural Language Processing (NLP) and Deep Learning have enabled the development to the powerful applications for next-word prediction. This technology has the potential to revolutionize the way people interact with computers and machines. The goal of this project is explore the possibilities of using NLP and Deep Learning to create a next word prediction system that can accurately predict the next word in a sentence .Word prediction tools were developed the which might facilitate to speak and additionally to assist the individuals with less speed while writing. during this paper, a language model based mostly framework for fast electronic communication, which will predict probable next word given a group of current words are briefed. Word prediction technique will the task of guesswork the preceding word that's probably to continue with few initial text fragments. Our goal is to facilitate the task of instant electronic communication by suggesting relevant words to the use. The advancements in deep learning and the availability of vast amounts of text data have significantly improved the accuracy and usability of next word prediction systems. They have become an essential component in many applications, aiding users in faster and more efficient text generation, suggesting relevant search queries, and facilitating seamless human-machine interactions. In recent years, neural network-based approaches have gained significant popularity in next word prediction. These models, such as recurrent neural networks (RNNs) and transformers, can learn complex patterns and capture contextual information effectively. By training on large text corpora, these models can learn word embeddings and generate more accurate predictions based on the learned representations. Next word prediction involves analyzing the previous words in a given text and using this analysis to predict the most probable word that should follow. This task can be challenging due to the complexity and variability of human language, where the same word can have multiple meanings depending on the context and the order in which words appear in a sentence can significantly influence the meaning of the sentence.

PROBLEM STATEMENT

The problem to be solved is to develop a system that can accurately predict the next word in a sentence given a set of input words. The system should be able to take into account the context of the sentence and the surrounding words to make an accurate prediction. The ultimate goal of the problem statement is to create a next word prediction model that enhances user experience, improves productivity, and provides accurate suggestions in various text-based applications

METHODOLOGY

DATA PRE-PROCESSING

These area unit easy clean-up procedures that makes it easier to use the information in sequent steps. This method is administered with the assistance of Tensor flow library. The subsequent area unit few pre-processing steps typically done:-

1. Marking white areas
2. Lower-case conversions
3. Removing numbers
4. Removing punctuation
5. Removing unwanted words
6. Removing non-English words

TEXT ANALYSIS

To know the speed of occurrence of terms, Term Document Matrix operate was accustomed to produce term matrixes to achieve the account of term frequencies.

TOKENSIZATION

One in every of the vital social control strategies is named tokenization. It's merely segmenting the continual running text into individual segments of words. One terribly easy approach would be to separate inputs over each house associate degree assign an identifier to every word.

PAD SEQUENCE

When changing sentences to numerical values, there's still a difficulty of providing equal length inputs to our neural networks. Not each sentence is constant length. pad sequences operate is employed for artefact the shorter sentences with zeroes, and truncating a number of the longer sequences to be shorter. Additionally, because it is often specified whether or not to pad and truncate from either the start or ending, relying upon the pre settings and post settings for the padding arguments and truncating arguments. By default, artefact arguments and truncation can happen from the start of the sequence

LITERATURE SURVEY

[1] Hamarashid, H. K., Saeed, S. A., & Rashid, T. A. (2021). Next word prediction based on the N-gram model for Kurdish Sorani and Kurmanji. *Neural Computing and Applications*, 33(9), 4547-4566.

This author includes the main algorithm and their accuracies which performed on predicting Next Word Prediction. In this paper, author proposed Ngram language model. Language models assign probabilities to a series of words or a sentence or the probability of the next word given a preceding group of words. In the Stupid Back Off algorithm, the attempt to build language models by distributing true probabilities. These models can be useful in a variety of fields, such as spell correction, speech recognition, machine learning, etc. N-gram modelling is utilized to suggest text accurately. The Ngram model has been used for next word prediction to reduce the amount of time. The model is 96.3% accurate.

[2] Yang, J., Wang, H., & Guo, K. (2020). Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network. *IEEE Access*, 8, 188036-188043.

Multi-window convolution and residual-connected minimal gated unit (MGU) network for the natural language word prediction. The convolution kernels with different sizes are used to extract the local feature information of different graininess between the word sequences. CNN extracts sequence feature information with different granularity by using different window sizes of convolution kernels. The overall experimental results show that the proposed MCNN-ReMGU significantly improves the performance of the word prediction task over the traditional methods. N-gram language model has been used to improve the text input rate in work. It is revealed that 33.36% reduction in typing time and 73.53% reduction in keystroke. The designed system reduced the time of typing free text which might be an approach for EHRs improvement in terms of documentation.

[3] Stremmel, J., & Singh, A. (2021, April). Pretraining federated text models for next word prediction. In *Future of Information and Communication Conference* (pp. 477-488). Springer, Cham.

In this paper, author proposed LSTM-RNN language model for the Next Word Prediction. Federated Averaging Algorithm is used which averages the model parameters. After applying the gradient to all models of database. Federated training on Stack Overflow without any pretraining which yields the two learning curves that exhibit the lowest levels of train and validation accuracy respectively. Dimensionality reduction approach is useful for federated training in which we are constrained by model size. As this paper has achieved the level of accuracy around 22.1%, and

22.2%. Federated learning is a decentralized approach for training models on distributed devices, by summarizing local changes and sending aggregate parameters from local models to the cloud rather than the data itself.

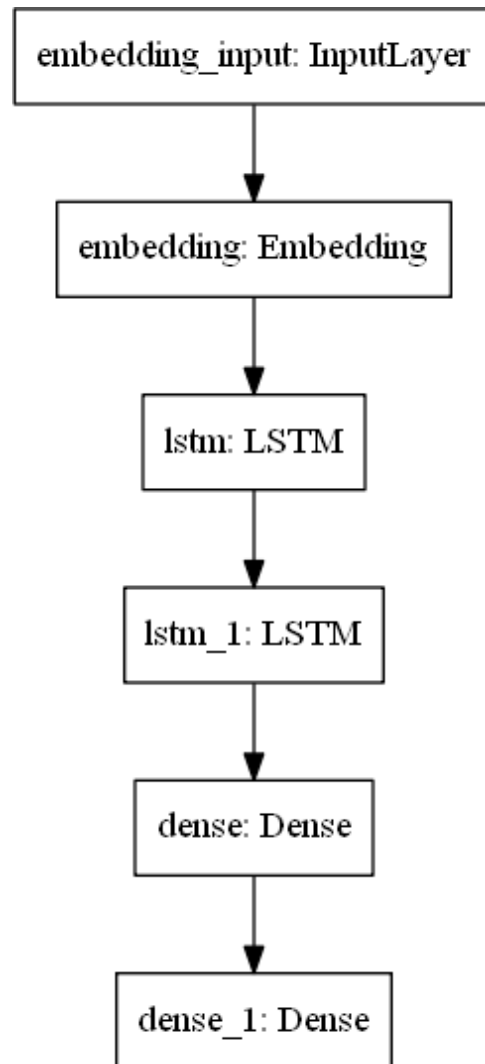
[4] Hamarashid, H. K., Saeed, S. A., & Rashid, T. A. (2022). A comprehensive review and evaluation on text predictive and entertainment systems. *Soft Computing*, 1-22.

This paper has used memory-based learning for next word prediction using Recurrent Neural Network and Long short-term memory. The goal of NLP is to learn and analyze the difficulties of automated generation and comprehending of languages of human beings. RNN has the capability of learning to utilize the previous information when the space between appropriate information and the position that is necessary is tiny. LSTM suffers from a large number of parameters, but it resolves the problem of memory. This model has reached up to the accuracy of 44.2%. Using hybrid-based technique such as Naive Bayes and Latent Semantic Analysis (LSA) model. The probabilistic method Naive Bayes is utilized in NLP like N gram. This model has produced accuracy of 88.2%.

[5] Barman, P. P., & Boruah, A. (2020). A RNN based Approach for next word prediction in Assamese Phonetic Transcription. *Procedia computer science*, 143, 117-123.

Next word prediction is a highly discussed topic in current domain of Natural Language Processing research. Recurrent Neural Network based language model to improve prediction of next word in sequential data. LSTM to generate complex long-range structured sequences. The Next Word using LSTM with an accuracy of 88.02% for Assamese text and 72.10% for phonetically transcript Assamese language

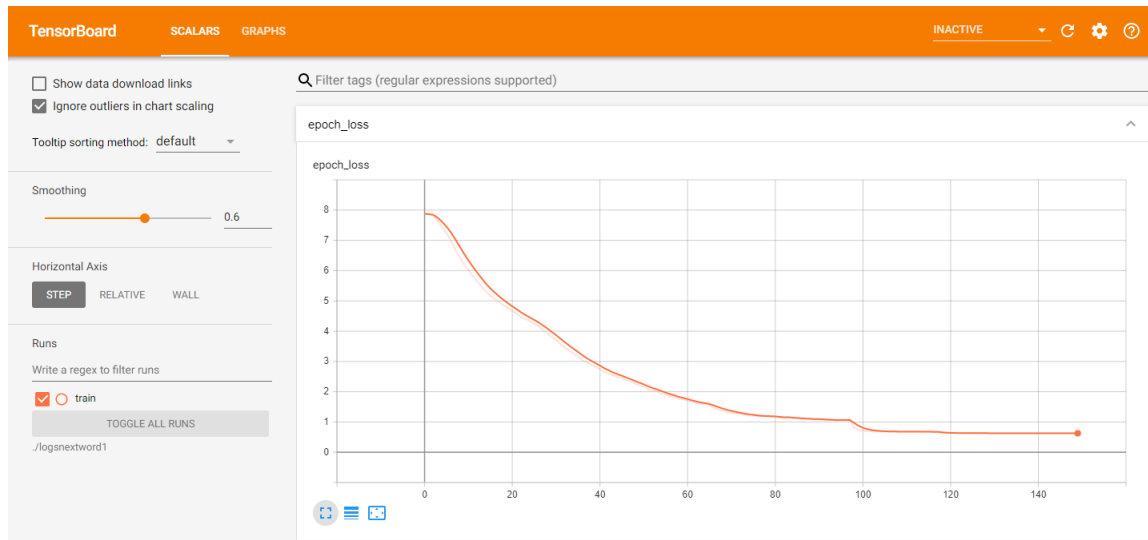
MODEL DESIGN



Fig(1.1) : The model plot

IMPLEMENTATION

GRAPH:



CONCLUSION

The subsequent word prediction model that was developed is fairly correct on the provided dataset. NLP requires applying various types of pattern discovery approaches aimed at eliminating noisy data. The loss was considerably reduced in concerning a hundred epochs. Files or dataset that are large to process need still some optimizations. However, bound pre-processing steps and bound changes within the model are often created to boost the prediction of the model.