EXO Next Word Prediction Using Machine Learning

Dr. S. Rajakumar
Professor(s), Dr.MGR Educational and Research Institute, rajakumar.subramanian@drmgrdu.ac.in

Dr. V. Rameshbabu
Professor(s), Dr.MGR Educational and Research Institute, rameshbabu.cse@drmgrdu.ac.in

Dr. D. Usha
Professor(s), Dr.MGR Educational and Research Institute, usha.cse@drmgrdu.ac.in

Nikila K
Final year B.Teh CSE, Dr.MGR Educational and Research Institute, nikilareddy2k1@gmail.com

Ramya Shree B
Final year B.Teh CSE, Dr.MGR Educational and Research Institute, bramyashree2002@gmail.com

Sakthi Priya R
Final year B.Teh CSE, Dr.MGR Educational and Research Institute, sakthiravi05062001@gmail.com

Abstract

It is never simple to write quickly without making mistakes. It's not difficult to type quickly and properly on a desktop computer, but many individuals find it challenging to type on compact devices like mobile phones. By simply guessing the word that will come next in a sentence, the next word prediction project makes it simpler for you to type on small devices. Due to the algorithm's ability to predict the following word and drastically reduce errors, you are not required to finish sentences. The creation of a language that is understandable by humans and sounds natural is the aim of natural language generation (NLG). This study suggests a novel technique for anticipating the following word in an English phrase. By assuming the next word in a series, the user can reduce the number of keystrokes they make. Two deep learning algorithms, Long Short Term Memory (LSTM) and Bi-LSTM were used to analyze the problem of predicting the next word. Accuracy values for LSTM and Bi-LSTM were 59.46% and 81.07%. Many NLG tasks, including sentence and narrative autocompletion, among others, can be accomplished using this technique.

They're still assuming the coming word grounded on the words that came ahead it’s essential to have clear-headed language models of the loftiest class, in fact, recent work on large-scale language models

has shown that RNNs perform inadequately on their own but well when used in confluence with n-gram models this is because their advantages over n-gram models are fully different circles live in intermittent neural network networks they enable the preservation of information. Assists in reducing user keystrokes while typing.

- Assists users in saving typing time.
- Assists in reducing user spelling errors. Assists non-native speakers in learning the language by suggesting new and correct

Next Word Prediction, English, LSTM, and Bi- LSTM are some of the index terms.

Keywords: *LSTM, Bi-LSTM, NLG, NLP*.

I. INTRODUCTION

Word prediction relies on computers simulating the thinking section of the human brain to suggest the following word as you type a phrase. It requires less manual labor, improves the syntax and syntactic organization of the sentence, and creates the appearance that it is shorter. We may effectively implement this concept while adhering to its original intent by "predicting the next nice word" in practice. Recurrent neural networks, deep learning, and strong conditional independence have considerably enhanced language modeling analysis since they allowed researchers to examine several tasks. Area assumptions fail to realize even if simpler models, like N-grams, only rely on a brief history.

words, thereby expanding their vocabulary. English is a widely used language in India. The great majority of individuals living in the world will benefit from the next word prediction in English. It won't be difficult to find the correct spellings. As a result, processing language at the word level produces better outcomes and requires less effort.

A part of artificial intelligence (AI) called natural language processing (NLP) helps develop effective methods to connect with people and gain from their interactions. One such commitment is to deliver anticipated words to portable clients. Long short-term memory, or LSTM for short, is able to comprehend prior information and

anticipate words that may be helpful for the user to create sentences as they write within applications. This can support message conveyance by allowing the client to select a suggested word rather than producing it. This method predicts a character to construct a word using letter- to-letter prediction. It will aid end-users in framing significant paragraph components and allow users to concentrate on the subject rather than wasting time on what to enter next because composing an essay and structuring a lengthy paragraph requires time.

Auto-complete features will be created or imitated using LSTM. The majority of software employs a variety of techniques for this task, including standard neural networks and NLP. With LSTM and the Default Nietzsche text file as our training data, we will be experimenting with this issue. The task of anticipating the following word is known as a next-word prediction or language modeling. It is an essential NLP activity with a wide range of applications.

The document is structured as shown below.

The literature review conducted on next- word prediction and the English language is discussed in Section 2. The proposed methodology is discussed in Section 3, along with the details of the dataset employed. The outcomes are illustrated in Section 4. Section

5 addresses the findings and the recommended course of action.

II. LITERATURE SURVEY

Existing systems predict the next word that will be supported by their word using a word prediction model. These systems rely on machine learning algorithms, which are limited in their ability to form correct syntax. They must also create a residual connected lowest gated unit (MGU), which is a condensed version of the LSTM, according to the multi-window convolution (MRN N) formula. Attempt to skip a few

layers throughout this CNN to reduce coaching time while maintaining reasonable accuracy. When predicting n words, using multiple layers of neural networks will result in latency. We introduce latency to predict n numbers by sacrificing multiple layers of neural networks.

They outlined the restrictions because the larger context Language Model doesn't depend on the notion of shared autonomy of sentences in an extremely large corpus. By depicting a conditional probability using words from a related sentence, it illustrates the influence of the context. However, the setting is also made up of a variety of previous sentences of varying lengths. In keeping with this, this model looks into a different aspect of the context- dependent recurrent Language Model. The larger setting Language Model improves sentence misunderstanding while significantly reducing word confusion in comparison to the Language Model [6] without context.

1. Text classification is a crucial task in natural language processing, according to

J.Y. Lee et al. [7].

Many approaches to classification have been developed, including B. SVM (Support Vector Machine), Naive Bayes, and others.

2. Authors J. Goodman, S. F. Chen, J. T. Great man, R. Kneser, and H. Ney

Set the following restrictions:

Despite the smoothing method described above there are practical applications for the N-gram language model.

The limitations are listed as follows by the authors (J. Goodman[3], S.F. Chen[2] and

J.T. Greatman[2], R. Kneser[1] and H. Ney[1]): Despite the smoothing strategies mentioned above and the reasonable usability of the N-gram Language Model, the curse of dimensionality (concerning a few different learning algorithms) is very strong here because there is a vast array of different blends of input variable values

that must be For LM, this is frequently the wide range of plausible word arrangements; for instance, there are 1050 possible groupings with a progression of ten words picked from a lexicon of 100,000.

III. METHODOLOGY

The first dataset is gathered, subjected to preprocessing, and used in the following ways using LSTM. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network that may learn order dependence in sequence prediction problems. The current RNN step uses the output from the previous step as its input. The LSTM was developed by Hochreiter & Schmidhuber. It addressed the problem of RNN long-term reliance, in which the RNN can predict words based on recent data but cannot predict words kept in long- term memory. As the gap length increases, RNN's performance is inefficient. By default, the LSTM may hold data for a very long time. It is utilized for the processing, forecasting, and classification of time-series data.

1. LSTM
2. Bi-LSTM
3. Data Preprocessing
4. Data Input, Training and testing
5. Tensorflow

1. LSTM

Long Short-Term Memory (LSTM) approach: Using a neural language model, Long Short-Term Memory (LSTM) is a more sophisticated method (LSTM). Unlike conventional neural networks and other models, the LSTM model uses Deep Learning in conjunction with a network of synthetic "cells" that control memory, making it more effective for text prediction.

2. Bi-LSTM

A Bi-LSTM, also known as a bidirectional LSTM, is a sequence processing model made up of two LSTMs, one of which processes data in a forward way and the other in a backward direction. By successfully expanding the network's information pool, Bi-LSTMs enhance the context that the algorithm may access (e.g. For instance, understanding the words in a sentence that come before and after a word).

By running your inputs in two directions—from the present to the future and from the future to the present—bidirectional processing differs from unidirectional processing in that future information is preserved in the LSTM that runs backward, and by combining the two hidden states, you can preserve both past and future data at any given time.

3. Data Preprocessing

The adjustments made to our data before feeding it to the algorithm are referred to as pre-processing. Data preparation is a crucial step in the data processing process since it changes or transforms the existing data into usable data. In other words, anytime data is acquired from various sources, it is done so in a raw manner that makes analysis impossible.

4. Data input, Training and testing

We all make mistakes, and some of these mistakes, if left unchecked, might result in failures or defects that can be quite expensive to fix. Testing our code enables us to identify these errors or stop them from ever entering the production environment. Thus, testing is crucial to the creation of software. Effective testing aids in finding problems, ensuring product quality, and confirming that the software performs as intended.

5. Tensorflow

Tensorflow is an open-source, Python-compatible toolkit for numerical computation that accelerates and simplifies the creation of neural networks and machine learning algorithms. Dataflow graphs—structures that depict how data flows across a graph, or a collection of processing nodes—can be created by developers using TensorFlow. A mathematical operation is represented by each node in the graph, and each edge between nodes is a multidimensional data array or tensor.

Algorithm- Intermittent neural networks(RNNs) are a type of neural network where the affair of one step is used as the input for the coming step. The neural network's inputs and labors are independent of one another, but in cases where it's important to prognosticate the coming word in a judgment, it's needed to flash back the words that came ahead. RNN was created as a result, and it used a retired Subcaste to attack this issue. The main and most important property of RNNs is the retired state, which keeps some information about a sequence.

TEST STRATEGY AND APPROACH

Hand testing will be used in the field, and feature-written practical tests will be used.

 Each field entry must function correctly. Starting a page requires using the recognised link.

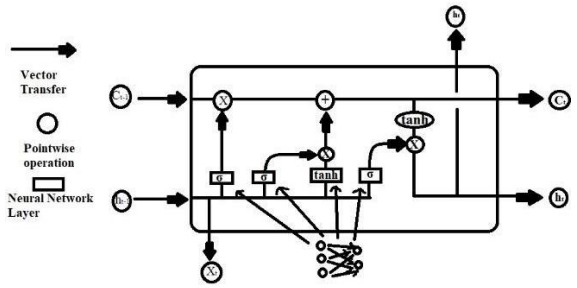
 You must not alter the access screen, messages, or responses. Verify the features.

 Confirm that the accesses are set up correctly.

 No replacement passes should be permitted.

Every link must direct users to the correct page.

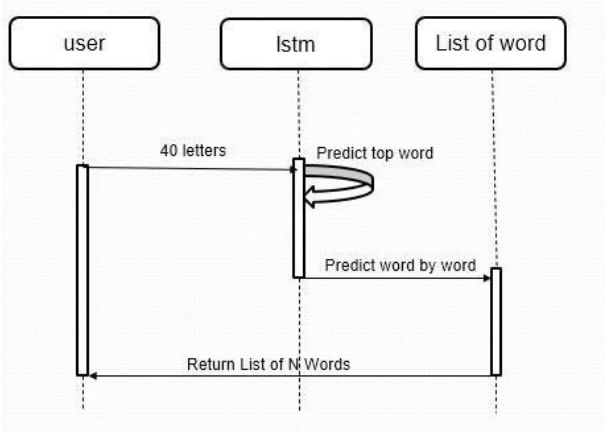
Fig 1.1 LSTM Network Architecture



LSTM: There are four layers will be there in the below illustration. The whole vector is carried out by the affair of a single knot followed by the inputs of the remaining. Thus the shape of the circle will represent the pointwise Operations and that will do some operations like is. Thus the boxes represent the neural network layers. also the resemblant lines denote the consecution, thus those separating symbols imply contents that were copied and transferred into colorful positions.

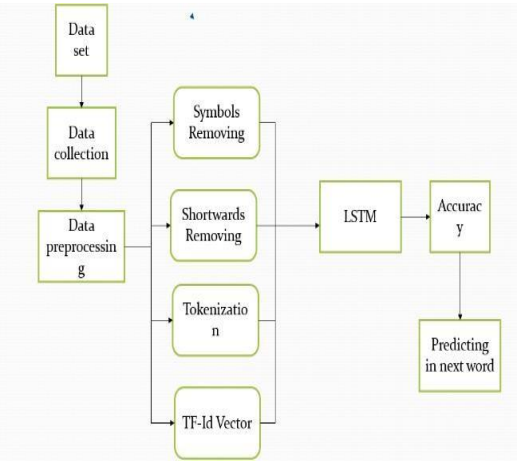
Fig 1.2 LSTM Sequence Diagram

As seen in the above graphic, the user's task is



to provide The LSTM Neural Network with 40 input words. LSTM: LSTM predicts a top word letter by letter after being trained on the default next- word prediction text file. Term list This will save every word and, in response to the user's request for N top works, count N words and provide the user with a list of words.

Fig 1.3 Process Diagram



IV. PROBLEM IDENTIFICATION

We tend to present condensed work in the language model in this paper. The NLM will address the issue of information adequacy with the help of completely distinct language models, and they are ready to capture the logical data at many levels, from sub-word to

corps level. The biggest disadvantages, however, are an extremely lengthy preparation period and a substantial amount of named training material.

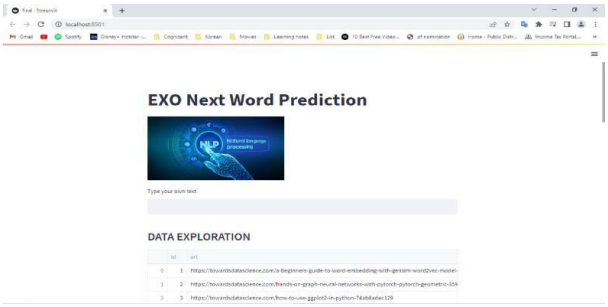
Using artificial intelligence to understand the passage RNN-style algorithms can help decipher and organize complete sentences and narratives. Songwriting is one of the most significant professions one can pursue. The method's end users can utilize it to make future predictions. When the model is trained to listen to a music data set of lyrics, a phrase is utilized in songs.

As more data is collected, the model that will determine utilize the weights to comprehend the fundamental aspects of para-favorable results is predicted by graphs or phrases.

To paraphrase is to craft someone else's thoughts and own words, You must rephrase a source to rework a passage without altering its intended meaning for our algorithms to estimate more numbers from the original text terms that take into account a particular sentence and aid users in frame n sentences.

Data Preprocessing, Text analysis, Pad sequence, Tokenization

V. RESULT AND DISCUSSIONS



For word prediction in English, we employed Data Preprocessing, Text Analysis, Pad Sequencing, and Tokenization in this project. In addition, we used Long Short Term Memory followed by Bi-LSTM to predict the word quickly and easily, saving the user's time while increasing accuracy to 82%.

VI. FUTURE ENHANCEMENT

By using various machine learning algorithms like RNN, LSTM, and Bi- LSTM to predict words in additional languages like Tamil, Arabic, Japanese, Chinese, Urdu, and so on, we may create artificialintelligence.

We can forecast the correct words and meanings in English from a movie's audio by identifying the audio of that particular movie, cartoon, television show, etc. People may not speak all the languages that are subtitled in movies for understanding purposes.

VII. CONCLUSION

Using the RNN algorithm, LSTM, Bi- LSTM, gathered datasets, preprocessing modules, and tokenization, we have improved the accuracy of word prediction in this study. The method has predicted the sequence of words and the right meanings of those specific words that the user will input in English. Because of the accuracy, we have included in our project, it is therefore highly useful for the users to speak with the people they want to and it also saves time for the user.

Reference

[1] J. Brownlee, How to Develop a Bidirectional LSTM For Sequence Classification,2019 (accessed April 10, 2019).

[2] Joel Stremmel, Arjun Singh. (2020). Pretraining Federated Text Models for Next Word Prediction using GPT2.

[3] J. Yang, H. Wang and K. Guo, "Natural Language Word Prediction Model Based on Multi-Window Convolution and Residual Network," in IEEE Access, vol. 8. pp.188036-188043, 2020, DOI: 10.1109/ ACCESS.20202.3031200.

[4] Milind Soam; Sanjeev Thakur Next Word Prediction Using Deep Learning: A Comparative Study Published in 2022

12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2022

[5] R. Sharma, N. Goel, N. Aggarwal, P. Kaur, and C. Prakash, "Next Word Prediction in Hindi Using Deep Learning Techniques", 2019 International Conference on Data Science and Engineering (ICDSE), pp. 55-60, Sep. 2019.

[6] ThanakomSangnetra,"glove.6B.300d.txt ", 05 2021.

[7] Y. Wang, K. Kim, B. Lee and H. Y. Youn, "Word clustering based on pos feature for efficient twitter sentiment analysis", Human-centric Computing and Information Sciences, vol. 8, pp. 17, Jun 2018.

[8] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single Layered Bi-LSTM Model", IEEE Access, vol. 8, pp. 73992-74001, 2020.