IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# NEXT WORD PREDICTION

[1]Keerthana N, [2]Harikrishnan S, [3]Konsaha Buji M, [4]Jona J B

[1]Student, [2]Student, [3]Student, [4]Associate Professor

[1]Department of Computer Applications,

[1]Coimbatore Institute of Technology, Coimbatore, India

***Abstract*:** Writing long sentences is bit boring, however with text prediction within the keyboard technology has created this easy. Next Word Prediction is in addition referred to as Language Modeling. It's the endeavor of predicting what word comes straightaway. It's one in every of the key assignments of human language technology and has various applications. Long short time memory formula can perceive past text and predict the words which can be useful for the user to border sentences and this method uses letter to letter prediction suggests that it predict a letter when letter to form a word.

***Index Terms - NLP, LSTM, RNN, Next Word.***

## I. INTRODUCTION

Word prediction tools were developed which might facilitate to speak and additionally to assist the individuals with less speed 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 user.

## II. LITERATURE REVIEW

Existing systems work on word prediction model, that suggests subsequent immediate word supported this out their word[2]. These systems work victimization machine learning algorithms that has limitation to form correct syntax. Multi-window convolution (MRN N) formula is enforced, additionally they need created residual-connected lowest gated unit(MGU) that is brief version of LSTM during this cnn try and skip few layers whereas coaching end in less coaching time and that they have sensible accuracy out and away victimization multiple layers of neural networks will cause latency for predicting n numbers of words .Developing technologies has been manufacturing additional correct outcomes than the prevailing system technologies, models developed victimization bidirectional LSTM algorithms area unit capable of handling additional knowledge expeditiously and predicts higher[2].

## III. THE ALGORITHM

### 3.1 LSTM

Vanishing gradient descent may be a downside featured by neural networks once considering back propagation. It's Brobdingnag Ian impact and also the weight update method is wide affected and also the model became useless. thus LSTM that incorporates a hidden state and a memory cell with 3 gates that area unit forgotten, scan and input gate.

The forget gate is principally accustomed get sensible management of what data has to be removed that isn't necessary. Input gate makes positive that newer data is additional to the cell and output makes positive what elements of the cell area unit output to subsequent hidden state. The sigmoid operate utilized in every gate equation makes positive we will bring down the worth to either a zero or one.

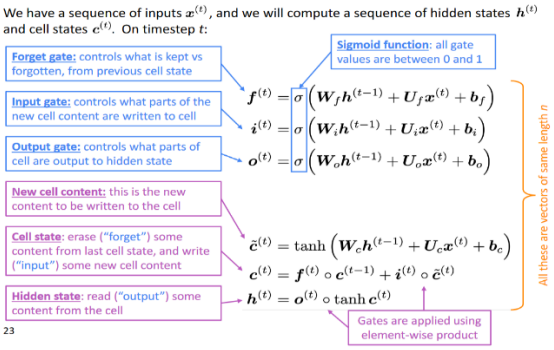


**Fig -1 LSTM Architecture**

The below Fig-2 is the Illustration of basic LSTM prediction model for our downside. The model encodes the input word sequence describing previous sub events into associate degree embedding and decodes a word sequence describing the doable future sub event out of the embedding.

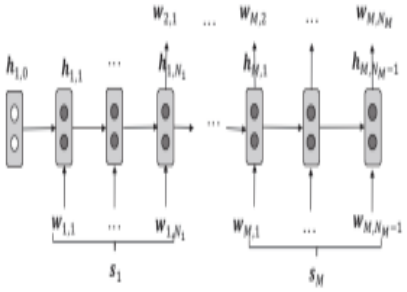

**Fig-2 LSTM Prediction Model**

### 3.2 RNN

Continual neural networks area unit generalizations of an instantaneous transmission neural network that has internal memory. The RNN is repetitive in nature as a result of it performs constant operate for every knowledge entry, however at constant time, this output depends on the previous calculation. the choice relies on associate degree analysis of this input and also the output from the previous input. RNNs will use their internal state (memory) to method input sequences once direct communication neural networks cannot. All RNN inputs area unit interconnected.

This state formula is that the following :-

$$h_f = f(h_{t-1}, x_t)$$

Application of the activation operate :-

$$H_t = \tanh(W_{hh} ht\text{-}1 + W_{xh} x_t)$$

## IV. PROPOSED METHODOLOGY

### 4.1 Data Preprocessing

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

### 4.2 Text Analysis

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

### 4.3 Tokenization

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 degreed assign an identifier to every word.

### 4.4 Pad Sequences

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.

## V. IMPLEMENTATION AND RESULT

The implementation and results obtained for this project are shortly delineate during this section.

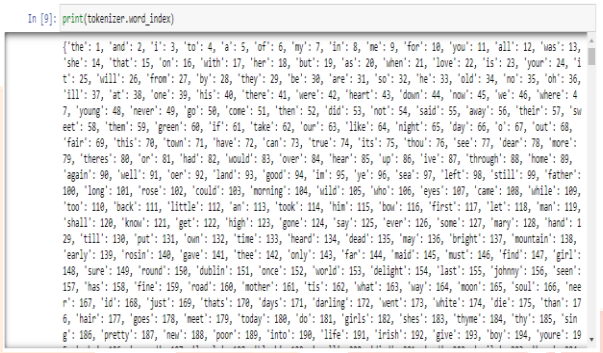


**Fig – 3 Tokenization**

In Fig-3 the full dataset is separated into individual phase of words. When segmenting a word index is formed with specified distinctive variety of words in irish-lyrics-eof.txt by distribution fact to every of them.

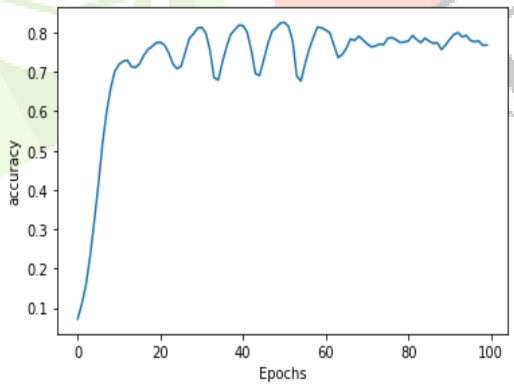


**Fig – 4 Accuracy**

In Fig-4 when training the model, the accuracy is found to be 0.80 in a hundred epochs.

```
In [13]: seed_text = "I see a beautiful river"
         next_words = int(input("Word Count : "))

         for _ in range(next_words):
             token_list = tokenizer.texts_to_sequences([seed_text])[0]
             token_list = pad_sequences([token_list], maxlen=max_sequence_len-1, padding='pre')
             predicted = np.argmax(model.predict(token_list), axis=-1)
             output_word = ""
             for word, index in tokenizer.word_index.items():
                 if index == predicted:
                     output_word = word
                     break
             seed_text += " " + output_word
         print(seed_text)

         Word Count : 2
         I see a beautiful river my love
```

**Fig – 5 Output**

In Fig-5 associate degree input sentence "I see a stunning river" is given. variety of words to be expected when this words is got as input from the user. Because the word count is given as two "my love" is that the expected words.

## VI. 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.

## REFERENCES


**[1]** Bengio, Y., Simard, P., Frasconi, P., 1994.Learning long dependencies with gradient descent is troublesome. IEEE transactions on neural networks five, 157– 166.

**[2]** R. Kneser and H. Ney, "Improved backing-off for n-gram language modeling", *Conference on Acoustics speech and Signal Process*, pp. 181-184, 1995.

**[3]** Mohd. Majid and Piyush Kumar, *Language Modelling: Next word Prediction*, 2019.

**[4]** Bengio, Y., Simard, P., Frasconi, P., 1994., Learning NLP with gradient descent is troublesome. IEEE transactions on neural networks five, 157–166.

**[5]** Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016. Building end-to-end dialogue mistreatments generative stratified neural network models. In Proceedings of the 30th Conference on Artificial Intelligence. AAAI.

**[6]** J. Yang, H. Wang and K. Guo, "Natural language Word Prediction Model supported Multi-Window Convolution and Residual Network," in IEEE Access, vol. 8, pp. 188036-188043, 2020, doi: 10.1109/ACCESS.2020.3031200.

**[7]** M. K. Sharma, S. Sarcar, P. K. Saha and D.Samanta, "Visual clue: Associate approach to predict and highlight next character," 2012 fourth International Conference on Intelligent Human pc Interaction (IHCI), Kharagpur, 2012, pp. 1-7, doi:10.1109 /IHCI.2012.6481820.

**[8]** Sukhbaatar, S., Weston, J., Fergus, R., et al., 2015. End-to-end memory networks, in: Advances in neural information processing systems, pp. 2440–2448.

**[9]** Zhou, C., Sun, C., Liu, Z., Lau, F., 2015. A c-lstm neural network for text classification. arXiv preprint arXiv:1511.08630.

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

**[11]** S. M. Sarwar and Abdullah-Al-Mamun, "Next word prediction for phonetic typing by grouping language models," 2016 2nd International Conference on Information Management (ICIM), London, 2016, pp. 73-76, doi: 10.1109/ INFOMAN.7477536.

**[12]** 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.