International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 91-94
Published Online August 2022 in IJEAST (http://www.ijeast.com)

SENTENCE PREDICTION USING NLP TECHNIQUES

Sakthi Subramaniyam.K, Srivatsan.S, Vignesh Muthu.K, Vishnu.M
Student
Department of Computer Science Engineering
Saranathan College of Engineering
Trichy, India

***Abstract* - In Today's world automation has been playing a significant role. A language model is a set of rules for evaluating the likelihood of sequences of text. Language models have many uses, including generating text by repeatedly answering the question like what word would come next. Sentence Prediction is a technique in which it predicts the possible outcomes of next words. Sentence Prediction using LSTM and BERT technique is an efficient way to predict the word also with web scraping it is easier to look for doubts without switching tabs. The main aim of our project is to predict 5 or more words as quick as possible using BERT and LSTM model.**

***Keywords*— MLM, NLP, NSP.**

I. INTRODUCTION

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. In this project, Next word in the sentence is predicted using LSTM and BERT Model as LSTM is Long short time memory it will understand the past text and predict the words which may be helpful for the user to frame sentences and this technique uses a letter-to-letter prediction means it predicts a character to create a word. BERT or Bidirectional Encoder Representations from Transformers is a language-based model trained on large online text corpus like Wikipedia and movie reviews to make computers aware of the workings of language. Unlike earlier models BERT is deeply bi-directional. Comparison with other models is shown below. GPT is unidirectional where ELMO is shallow bi-directional. Web scrapping deals with collecting web data and information and retrieve it to the user. It mainly gathers information from news gathering, competitive marketing and more. It is very simple and efficient. So, by predicting the next word we can easily draft an essay or a paragraph in short span of time.

II. LITERATURE SURVEY

1 Next Words Prediction Using Recurrent Neural Networks - Sourabh Ambulgekar; Sanket Malewadikar, Raju Garande, and Dr. Bharti Joshi.

The existing system used multi-window convolution (MRNN) algorithm 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. The Authors worked on Bangla Language. They have proposed a novel method for word prediction and word completion. They have proposed N-gram based language model which predicts set of words. They have achieved satisfactory results. The Authors also worked on Assamese language. They have stored transcripted language according to International Phonetic Association (IPA) chart and fed to their model. They created a model for physically challenged people. This model uses Unigram, Bigram, Trigram, based Approach for next word prediction and was found out average to predict the word but accuracy was around 30-40 percentage.

III. EXISTING SYSTEM

We have many deep learning models detecting the text based on available corpus by providing input text and grammar corrections that uses more computation and more storage to store the data. The existing system does not have reference tab. So, we must search in browser and need to switch tab constantly for references.

IV. PROPOSED SYSTEM

The main aim of our project is to predict the sentences as quick as possible using BERT and LSTM model. Models understand the relationship between sentences. It was trained with the masked language modelling (MLM) and next sentence prediction (NSP) objective.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 91-94
Published Online August 2022 in IJEAST (http://www.ijeast.com)

V. ARCHITECTURAL DESIGN

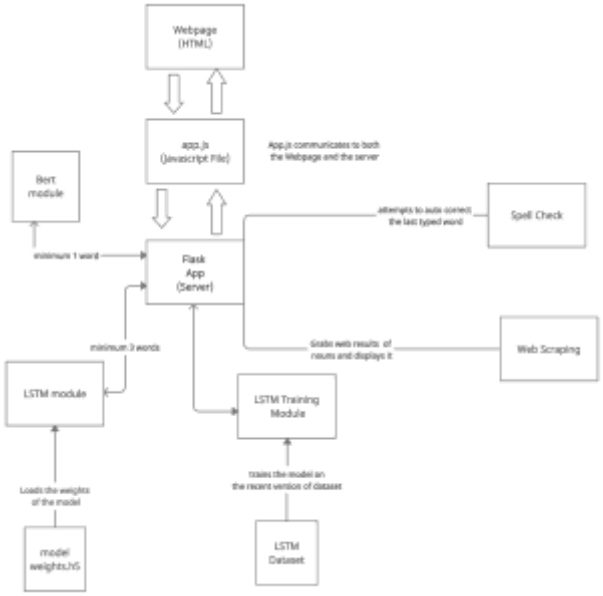

Fig.1 Architecture Diagram

Fig 1shows the Architectural Diagram of how our system works. First the input will be passed only when the user hits the space on the text. Then the program it checks the spelling if any mistakes found it has been rectified and it goes to noun check function, if the given word is noun it calls the web scrap function and shows the relevant information about the word. Then the program checks whether the number of words in the text box is less than three or not. If the number of words in the text box is less than three then the text is sent to the BERT model alone. And the BERT model will give the predictions of the next word and the predictions are displayed in a list. If the number of words is greater than three then the input from the text box, which is the last three words from the text, is sent to the LSTM model and the BERT model. Here both the models will give predictions about the next word. And both the results will be displayed. And the result from our LSTM model is highlighted in the list of words to indicate it is from the LSTM model.

VI. MODULE DESCRIPTION

A. SPELL CHECK

The spell check module handles the input text by detecting and rectifying spelling mistakes and shows the correct spelling above the character. The spell check is implemented using the library Text blob. Some of the tasks, were done better using Text blob over other Python libraries, which are spelling correction, part of speech tagging, and text classification. But it can be used for various NLP tasks like The Spell Check also checks whether the word is Noun or not. If the word is noun, then it is sent to the web scraping module.

B.WEB SCRAPPING

Beautiful Soup is used to get the data from the internet. Here, Scraping is done on the Wikipedia and Google search engine's output. Web scraping is done only when the last typed word is noun or the last typed word cannot be recognized by Text Blob. Now the last word is used to create Google search results and the relevant information present in the top search result or Wikipedia is displayed.

C. PREDICTION

The prediction module contains the functions to do the predictions based on the input. Here we have two models BERT and LSTM. Each model has its own weight file. The models are loaded at the first prediction call. The models which should be used for the prediction is based on the no of words given as input.

1) BERT Module: It stands for Bidirectional Encoder Representation Transformers, it is a pretrained model to predict Top k words where top_k<=10
2) LSTM Module: If the input is minimum of 3 words, then the LSTM start to predict words. The last 3 words of sentence is taken as input and predict the fourth word as next word prediction

VII. IMPLEMENTATION

Here the input is sequence of words. If the input is single word, then the output will come from LSTM module. If the input is minimum of 3 words, then the output will come from pre-trained Bert module. If it is not noun then spell check will be processed and if it is a noun the web search output of particular noun will be displayed.

A) App.py

The application will run on flask server. In front end we use CSS, HTML, bootstrap and the backend will be app.js and python. In app.js the data will be transmitted from client to server using AJAX in j Query and the client data will be transmitted to app.py in flask server. Where in every time user type space the data start to transmit from app.js to app.py. In app.py consist of modules that would be designed to run the application like main.py, predict.py, spelling.py, webscrap.py. Each and every module will be called in app.py.

B) BERT Module

It stands for Bidirectional Encoder Representations from Transformers, BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Before feeding word sequences into BERT, 15% of the words in each

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 91-94
Published Online August 2022 in IJEAST (http://www.ijeast.com)

sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words.
For example:
How are _____.
Here we are going to append mask tag '<mask>'
Then the input sentence is like: How are <mask>

Here the word tag <mask> is replaced by list of predicted words. That is "How are you". The BERT module requires 2 module one will be input sentence and other will be k<=10 to predict k number of words. In python here using module called transformers, from transformers import Bert Tokenizer. Inside Bert Tokenizer we call from pretrained () function. Here we step into two major processes call encoder and decoder. In encoder function will processed by torch module to find out input_id and masked_id. With the help of input id and masked id to perform decoding. Input id consist of matrix with id of unique words and mask id is the position of [MASK] tag. In decode function can process with input id and mask id to predict top k words.

C) LSTM Module
If the input is minimum of 3 words, then the LSTM start to predict words. The last 3 words of sentence and predict the fourth word as next word prediction. Suppose, we have sentence like **"I am Yash"** and this will be converted into a sequence with their respective tokens **{'I': 1,'am': 2,'Yash': 3}**. Thus, output will be **['1','2','3'].**Here in python module for LSTM is Tensor Flow to train the customized dataset called "dataset.txt". To convert the lines of dataset to string to process the string by removing the special character and new line. And to tokenize the sentence in words and dump the content in pickle file. To convert the content into array by using Num Py module. We create two list called X and Y where x consists of sequence of 3 words and y consist of 4^{th} word of sequence. To train the sequence of data using LSTM function where to store the trained data after 70 epochs in next_words.h5 file. NOTE: Here to use LSTM model because to train the customized dataset.

D) Spell check
Before moving into this function first the program have to check whether the n^{th} word is noun or not. If it is noun then the n^{th} word not process into spell check function because the noun not need spell check. Where to process this by using NLTK module. If it is not noun, then the n^{th} word process into spell check function. Here using predefined module called "Text blob". Then to call the correct () function to correct the spelling if the n^{th} input word is same as output, then the given word corrects but if the input word is different from output, then the n^{th} word maybe spelling mistake and popup in output window.

E) Web Scrapping
Web scrapping is the automatic method to obtain large amount of data from website. Most of this data is unstructured data in an HTML format which is then converted into structured data. Here in input data to find the n^{th} word is noun or not. If the given word is noun, then the n^{th} word is pass as the input in web scrapping module and get result from Wikipedia. Here we import module called beautiful soup and by using request. get() function we process URL of Google. Then we get all <a> tag of page of input noun and get all link in list. Then we choose the Wikipedia link and get search of input content. Inside Wikipedia page we select the first heading class and using beautifulsoup module to select <p> tag after the heading to get the first paragraph of Wikipedia.

VIII. RESULTS

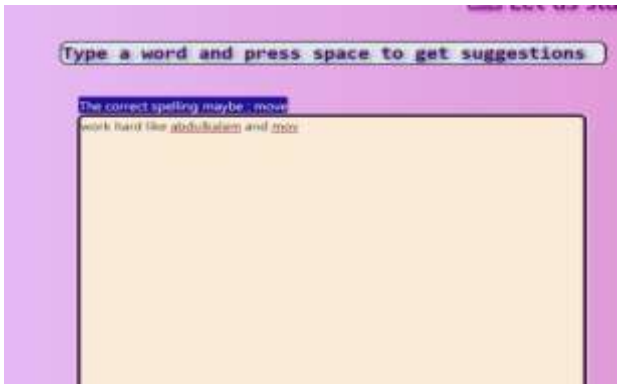

Fig.2 Spell check

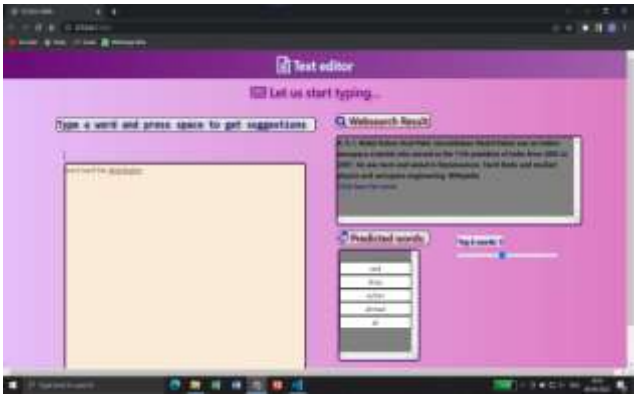

Fig 3 Web scrapping

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 91-94
Published Online August 2022 in IJEAST (http://www.ijeast.com)

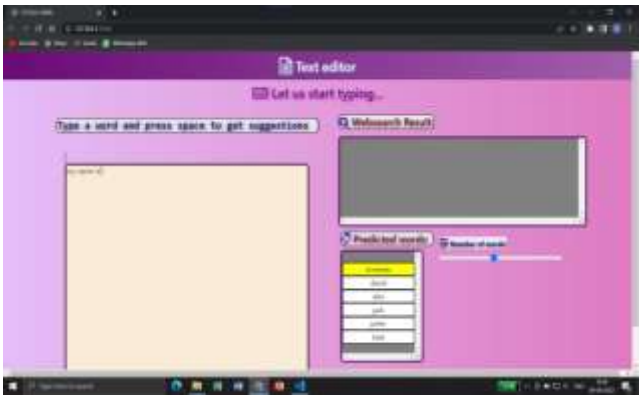

Fig 4 LSTM Prediction

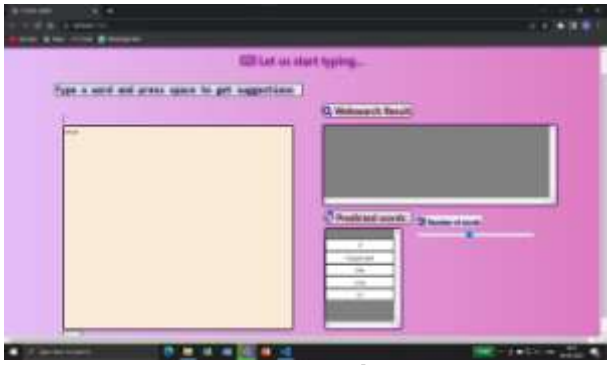

Fig 5 BERT Prediction

IX. CONCLUSION

In this project, the next sentence has been predicted using LSTM and BERT model where Bert is created by Google which is already a pretrained model is used to predict minimum one word and LSTM model is used to predict if the sentence has more than of three words. Web Scrapping is also used if the given word is noun so, it retrieves relevant data and provides the link from where it has been scrapped. Spell check is also used to correct the words if there is any error. The Project has an accuracy of about 75% in BERT and 79% in LSTM (depending upon the dataset). The project has been completed and the sentences has been predicted successfully.

X. REFERENCES

[1] Sun, Y., Zheng, Y., Hao, C., &Qiu, H. (2021). NSP-BERT: A Prompt-based Zero-Shot Learner Through an Original Pre-training Task-Next Sentence Prediction. ArXiv, abs/2109.03564.

[2] Khare, A., Gupta, A., Mittal, A., Jyothi, A. (2021). Text Sequence Prediction Using Recurrent Neural Network, in Advances and Applications in Mathematical Sciences, (Pg377-382).

[3] Algan, A. C. (2021). Prediction of words in Turkish sentences by LSTM-based language modeling [M.S. - Master of Science]. Middle East Technical University.

[4] Karunarathne, M.S., Nanayakkara, L., &Ponnamperuma, K. (2013). Sentence Prediction on SMS in Sinhala Language.

[5] Christian, H., Suhartono, D., Chowanda, A., &Zamli, K.Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8, 1-20.

[6] Wang, B., &Kuo, C.J. (2020). SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2146-2157.

[7] Sarker, S., Islam, M.E., Saurav, J.R., & Nahid, M.M. (2020). Word Completion and Sequence Prediction in Bangla Language Using Trie and a Hybrid Approach of Sequential LSTM and N-gram. 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), 162-167.

[8] Moura, G.B., &Feltrim, V.D. (2018). Using LSTM Encoder-Decoder for Rhetorical Structure Prediction. BRACIS.

[9] Shi, W., &Demberg, V. (2019). Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. EMNLP.

[10] Khera, S., & Kumar, M. (2020). THE COMPARATIVE ANALYSIS WITH BERT AND ELMOMETHODS FORMOVIE REVIEWS PREDICTION USING NLP.

[11] Widmoser, M., Pacheco, M.L., Honorio, J., & Goldwasser, D. (2021). Randomized Deep Structured Prediction for Discourse-Level Processing. EACL.

[12] Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., & Heck, L. (2016). Contextual LSTM (CLSTM) models for Large scale NLP tasks. ArXiv, abs/1602.06291.

[13] Soam, M., & Thakur, S. (2022). Next Word Prediction Using Deep Learning: A Comparative Study. 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 653-658.

[14] Brahma, S. (2018). Improved Sentence Modeling using Suffix Bidirectional LSTM. arXiv: Learning.