Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0236
Indexed in International Scientific Indexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

Machine Learning Techniques for Next Word Prediction

Veeresh K M[1], Prashant Ankalkoti[2]

[1*]Student, [2]Assistant Professor, Dept of Master of Computer Applications,
J N N College of Engineering.

veeeshmenasagi360@gmail.com, psankalkoti@gmail.com

Abstract

Next word prediction helps to enhance the typing speed of new language learners, we employ auto-completion utilizing advanced natural language processing technology. Specifically, we use the BERT (bidirectional encoder representations from transformers) model to predict word sequences. This approach helps new language learners type more efficiently and quickly. Various models, including federated models and NLP, have been explored by researchers. The BERT model, in particular supports traditional parameters such as the number of predictions and tasks, while ensuring that the predictions are relevant. This technology is integrated into messaging apps and document creation software such as microsoft word and dropbox paper. Additionally, flask-a lightweight web framework-reveals the essential features of the prediction models and integrates them with other applications. This integration facilitates the implementation of prediction models into various platforms. Thereby enhancing the user experience for new language learners by making text entry more efficient and effective.

Keywords: *BERT, NLP.*

1. Introduction

The most important is machine learning addressees the question of how to build components that improve automatically Though experience [1]. In an era of digital communication, the critical aspect of user experience is efficient text input. For the new language learner faces the correct next word for the sentence. composing emails, chatting on messaging platforms, the speed and accuracy can impact on the output. The main objective of this project is enhancing the user experience by the time and effort required for the new language learner typing. By predicting next subsequent word helps in generating coherent relevant text. The "Next Word Prediction" system provides the user with various option to fit the client's requirements and interest. To regulate the number of predictions , a user can define the-number of tokens to generate thereafter ,and whether to increase the randomness of the generated tokens using the temperature parameters .they also make it possible to adapt the system to different setting and user needs so that the text input method offered is effective as possible.

2. Literature Survey

Literature review aids with grasping the importance of an inquiry and how it relates to past work according to jylee et al text classification is a larger task in natural processing

language (NPL) [2]. Analysis of research suggests the text classification matters greater naturally a few of the investigators who already brought it up possessed a record of their findings in the form of a journal that they trusted these are pre training federated text models for next word prediction by joel streammel and arjun singh[4] and next word prediction using recurrent neural net-works by saurabh ambulgekar[3].The larger setting languages model enhances on the results of parentage in the aspects of misunderstanding of sentences while the confusion of word is relatively decrease when compare with the language model[5] which has no context. The researcher search the database from the internet which is also known as web scrapping. web scrapping is technique used to automatically get some information from the website [6]. " The Next Word prediction"(NWP) is the critical problem in the era of natural processing language[7].Building the technologies has been producing correct outcomes the convincing system techniques, model build victimization bidirectional LSTM algorithms unit capable of holding knowledge instantly and predict higher[8].[9] multi window convolution(MRNN) algorithm is applied, as well they have built residual connected minimal gated unit(MGU) which is short form LSTM in this cnn try to jump few layers while training outcomes in less time and accuracy by using neural network.[10]The 1-degree pattern method is used to address disappearance problems but it is not as accurate for simple sentences like a man cries or a dog barks. [11] the issue of next word prediction for the assamese language has been tackled using lstm in the end they symbol-ized the cant and fed it into their model they have saved tran written ipa phonetic notation or the application of a global phonetics script in transcription of specific languages. BERT (Bidirectional Encoder Representations from Transformers) is chosen for Next Word Prediction because it effectively understands the context of words by reading text bidirectionally, rather than just left-to-right. This allows BERT to capture the nuances and meanings of words based on their context, resulting in highly accurate predictions. BERT's pre-training on large text corpora enables it to be fine-tuned for specific tasks with smaller datasets, making it efficient and adaptable. Its robust performance in handling ambiguities and its continuous improvement by the research community make BERT a reliable and advanced model for next word prediction tasks.

3. Methodology

The "Next Word Perdition technique" involves many analytical steps and methodologies, combining both natural language processing and deep learning. The most important of this section is how the technical process is done by the methods. And the how the methods are work thorough the which models. First data set is collected according to their processioning.

3.1 Model setup and selection:

BERT model: BERT model is bidirectional encoder Representations form Transformer. The main advantages of this model is generating and understanding human language. BERT, or Bidirectional Encoder Representations from Transformers, is a cutting-edge language model developed by Google AI Language. Built on the Transformer architecture, BERT excels in natural language processing by processing text bidirectionally—both left-to-right and right-to-left—allowing it to understand context comprehensively. This capability makes BERT adept at tasks like text classification, named entity recognition, and question answering. Its pre-training on vast datasets enables nuanced

understanding of language, while fine-tuning on specific tasks enhances performance across diverse applications. BERT's versatility and state-of-the-art performance have cemented its role as a foundational tool in understanding and generation technologies.

3.2 Data Prepossessing:

Tokenization: tokenization is help to splitting the sentence into sub words and give the numbering to respected words with the help of BERT model. When we required the next word of the sentence that numbering will give the correct word.

3.3 Natural Processing Language:

Natural processing language is the field of AI helps to interact with computer and the human language and analyze the large amount of natural language data. Text classification, Named entity recognition, Summarization these are the main task of the natural language processing.

3.3.1 NLP ARCHITECTURE DAIGRAM:

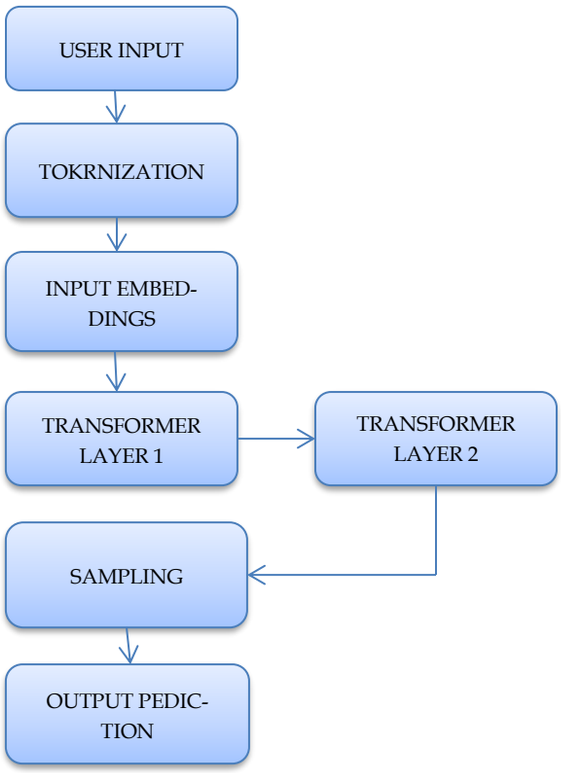

Fig 3.1.2: Shows NLP architecture

The diagram provides the framework of the Natural Language Processing (NLP) model that is under consideration. It has the user input text as the input and it is preprocessed by segmentation where the input text is divided into smaller units of analysis. The latter are transformed into numerical vectors called embedding. Then the embedding are passed through several layers of Transformer model which learns the interaction and context of the input text. The output is provided with a once again passing of the tokens through the Transformer layers followed by a sampling step that is aimed at picking the best tokens. Lastly, based on the passed in input text, the model delivers the predicted output text.

3.3.2 Masked Language Modeling (MLM):

Masked Language Modeling (MLM) is a technique used in training models like BERT where certain words in a sentence are randomly masked. The model then learns to predict these masked words based on the context provided by the surrounding words. This approach helps the model understand language bidirectionally—considering both left-to-right and right-to-left contexts—enhancing its ability to grasp the meaning and relationships within sentences. This training method improves the model's efficiency and accuracy in various natural language processing tasks by enabling it to capture deeper semantic and syntactic understanding of text.

4. Prediction Generation:

4.1Ecoding input: The mask token is then placed in the text and BERT’s tokenizer is used to transform the requite text with the input mask token into input IDs. All of these input IDs are the necessary inputs that can be fed into BERT model for prediction.

4.2 Model Inference:

The passed encoded input is used as an input source for the BERT model to obtain the predictions. The model generates a probability distribution of the vocabulary for the position that the mask is placed on.

5. Customizable parameters:

Number of prediction:

Users are also able to provide the model with the number of outcomes they require. This makes it possible to provide suggestions in the numbers considered appropriate.

6. Flowchart Diagram

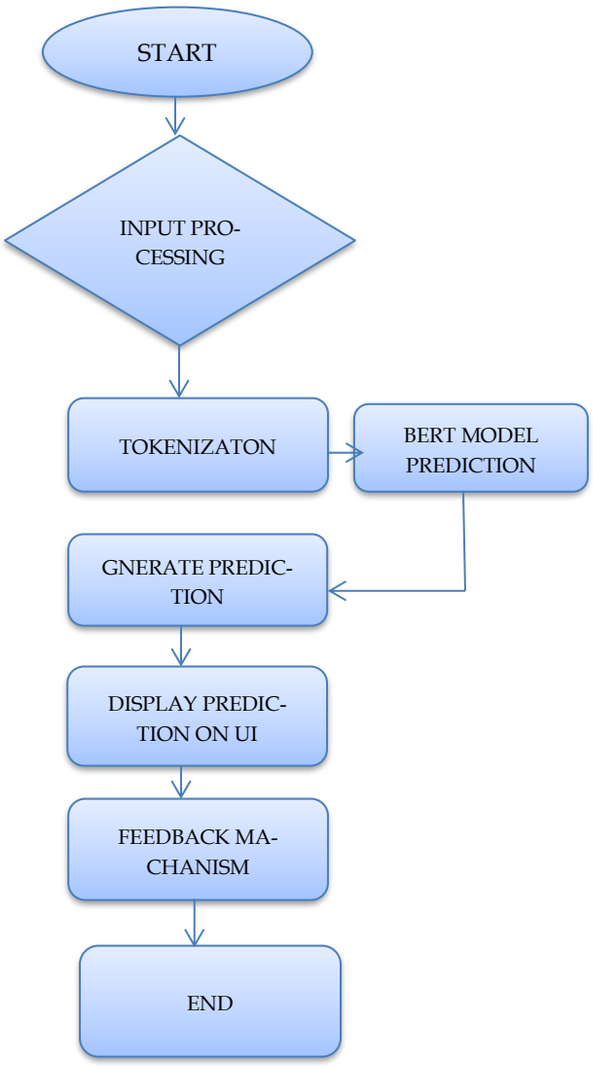

Fig 6.1: Flowchart Diagram

This flowchart describes a procedure that involves the following steps: receiving the user's input, preparing the input, normalizing the input after tokenizing it, preparing the tokens, feeding the tokens into a BERT model and generating a prediction before displaying it. First, there is the input, and this follows the input data processing. The input that is processed for prediction purposes undergoes tokenization to convert it into an acceptable formatted code. There are two paths: which would be one for making a general prediction one for using a BERT model for instance. The output of these predictions is provided in the form of graphs on the user interface for visualization purposes. The final point here presents a feedback mechanism, and this aspect helps in the determination of the finale of the work flow.

7. Result:

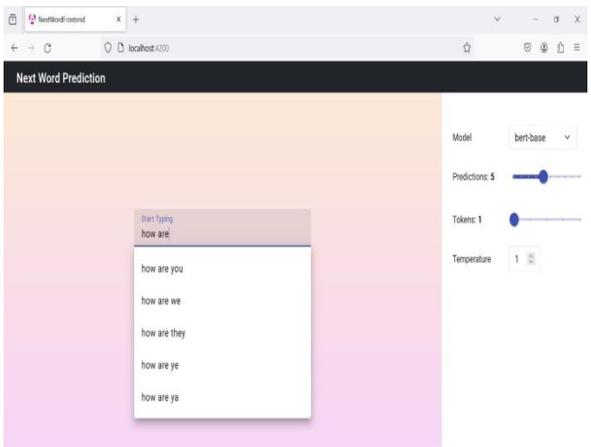

Figure7.1 next word prediction

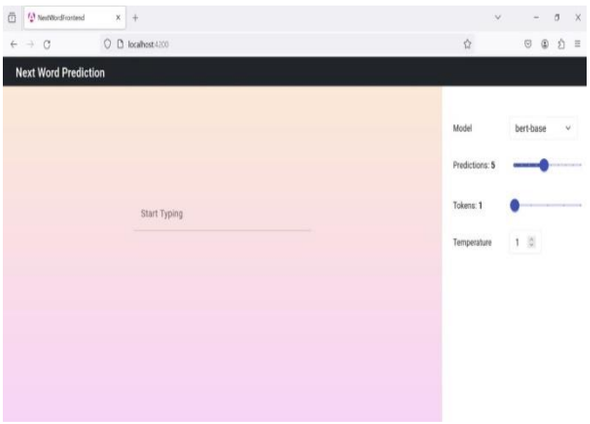

Figure 7.2:interface word prediction

7.1 COMPARATIVE ANALYSIS:

Author	Strengths	Limitation
Afika Rianti, Suprih Widodo, Irfan Ramadhan, Anindya Pitaloka	Achieved 75% accuracy- Utilized LSTM model effectively- Better than RNN and Pre Training Federated Text Model	Limited by dataset quality- Requires significant computational resources
Chinmaya Nayak, P. S. Sastry	Integrated various ML techniques- Emphasized on context-awareness- Improved computational efficiency	Less effective with complex sentence structures- Accuracy not specified
Sourabh Ambulgekar, Sanket Malewadikar, Raju Garande, Bharti Joshi	Effective use of RNN- Managed to produce relevant predictions- Used advanced NLP techniques	Achieved only around 54%-55% accuracy- Comparatively lower performance
Joel Stremmel, Arjun Singh	Federated learning approach for privacy- High accuracy in predictions	Requires extensive pre-training- Dependency on high-quality data
Proposed system	Integrated with many ML technique. Better the RNN technology	Without temperature scaling word prediction is not possible

TABLE-1 SHOWS COMRATIVE ANALYSIS

8. CONCLUSION:

excellent illustration of how concepts based on even more natural language processing and analysis may significantly improve the user experience across many media is the next word prediction system the systems ability to capture context and anticipate phrases depending on a high score is a stunning demonstration of its capabilities as demonstrated by its advantage in using the bert model it helps to improve writing and strokes with appropriate continuity and relevancy to the existing content in addition to increasing the rate of text insertion into related fields

9. FUTURE SCOPE:

Preserving privacy while providing utility in preparation and designing algorithms that are resistant to adversarial input will be the key in retaining user trust and reliability. Improvements in the user interface design and the ability to provide greater compatibility to utilize the system across mobile devices and different platforms will also improve upon it. Moreover, possibility adapted for education, and support purposes, for language learning deviations and disabled people it offers intelligent context-based text suggestion. These advancements are going to assure that the “Next Word Prediction” system will remain on top of technologies under the class of artificial intelligence text enhancement.

REFERENCES:

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

2. 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

3. Ambulgekar, Sourabh, Sanket Malewadikar, Raju Garande, and Bharti Joshi. "Next Words Prediction Using Recurrent Neural Networks." In ITM Web of conferences, vol.40, p.03034.

4. 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.

5. Vargiu, Eloisa, and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising." Artif. Intell. Res. 2, no. 1 (2013): 44-54.

6. 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.

7. R. Kneser and H. Ney, "Improved backing-off for n-gram language modeling", Conference on Acoustics speech and Signal.

8. J. Yang, H. Wang and K. Guo, "Natural LanguageWord Prediction Model Based on Multi-Window Convolutionand Residual Network," in IEEE Access,vol. 8, pp. 188036-188043, 2020, doi: 10.1109/ACCESS. 2020.3031200.

9. Y. Ajioka and Y. Anzai, "Prediction of next alphabetsand words of four sentences by adaptive junction,"IJCNN-91-SeattleInternational Joint Conferenceon Neural Networks, Seattle, WA, USA, 1991, pp. 897vol.2-, doi: 10.1109/IJCNN.1991.155477.

10. Partha Pratim Barman, Abhijit Boruah, A RNN based Approach for next word prediction in As- samese Phonetic Transcription, Procedia Computer Sci- ence, Volume 143,2018, Pages 117-123, ISSN 1877- 0509,