e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

TEXT PREDICTION AND AUTOCOMPLETE USING NATURAL LANGUAGE PROCESSING

Isha Shelke[*1], Ritali Waghulde[*2], Rutuja Gurav[*3], Samira Sutar[*4], Prof. Sarika Aundhakar[*5]

[*1,2,3,4]Student, Department of Computer Engineering, Smt.Kashibai Navale College of Engineering, Pune, Maharashtra, India.

[*5]Professor, Department of Computer Engineering, Smt.Kashibai Navale College of Engineering, Pune, Maharashtra, India.

ABSTRACT

This paper aims at improving the timing and accuracy of next word prediction and autocomplete using Facebook's RoBERTa Model. It has the best known accuracy of the NLP models thus far in terms of sentence prediction. We will use the pretrained RoBERTa Model through Pytorch hub for obtaining the predictions.Because the RoBERTa Model has a larger batch size and is extended to accommodate more data, it can predict more words per sentence than the BERT or similar data models. We will be representing it as a web app using Streamlit which is an open-source web app framework.Text prediction has various uses, ranging from allowing users database access to helping users with cognitive impairments.Accurate grammar and spelling are required not only for formal business conversations, but also for formal conversations held on the various different platforms like schools, colleges, etc.The project's goal is to create an intelligent system for optimizing the English language. This system will perform help with Optimization of grammar, spelling check, and Sentence Auto Completion.

Keywords: NLP, Auto-completion, RoBERTa, Prediction, Web.

I. INTRODUCTION

The term text prediction means a set of computer programs and algorithms, which help users edit text with higher efficiency. It can be seen in computer and mobile device applications such as email editors, texting, and web browsers. Auto-completion refers to the completion of a word, or a phrase, as we start typing in a document. The prediction is based upon the selection of the most likely word from a set of frequently used words. The text prediction task consists of editing text with the minimum number of keystrokes feasible. This method is suggesting words that the user intended to write, and the system predicts the next word related to the previous work.

Currently, the natural language used in this method is in English, but the method is not language dependent from any point of view. The system can work on the corpus of any natural language where the tokens are separated by space. This approach can help the differently abled people to write correctly and increase the typing speed. It is a technique which can prevent word error occurrences. The auto-correction (or text replacement) feature completes or replaces previously typed text. It is can be used to either correct words containing grammatical mistakes or to identify the correct word in a dictionary. Although, the process of insertion of a word usually starts after a user finishes typing, it is a type of text prediction.

The project aims at building an intelligent system to optimize the English language. It will perform the following functions:

1. Optimization of Grammar
2. Spelling Check
3. Sentence Auto Completion and prediction

The technology we will be using for this project is RoBERTa. It was developed by Facebook by improving Google's BERT technology. With this technology, we will be able to design machines which react to human language by speaking like humans and passing the Turing test with a considerably fair number of results hence, it is the worth appreciating as far as a man-machine integration is concerned.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

II. RELEVANCE OF WORK

Scope

This project will benefit all users who want to finish their tasks quickly. As we know the internet is an ocean of information and to find correct information in less time, we need smart techniques such as text prediction and autocomplete. Following are some areas for future improvement in this system.

1. To improve web security while applying autocomplete feature.
2. Modify current application for better performance.
3. Add functionality so that specific domain words are suggested.

III. PROPOSED SYSTEM

The proposed system has multiple modules as shown in Figure 1.The editor accepts user input and forwards it to the text extraction unit, which then branches out into the autocorrect and autocomplete engines, which are responsible for correction and prediction, respectively. Then the extracted text is also sent to an analysis unit which extracts keywords from the text. The suggestion unit is responsible for providing real time suggestions ordered by ranks in the user interfaces. All the modules are explained in detail below.

A. Text Extraction Unit

This module acts as a bridge between the user and the system and extracts each individual character typed by the user and passes to the engines..

B. Autocorrect Engine

This engine is in charge of ensuring that the user typed word is correct.

C. Autocomplete Engine

This engine is used for predicting the remaining part of an incompletely typed word and for guessing which word is likely to come next to the typed word fragment. This is divided into 2 submodules: Current Word Prediction and Next Word Prediction.

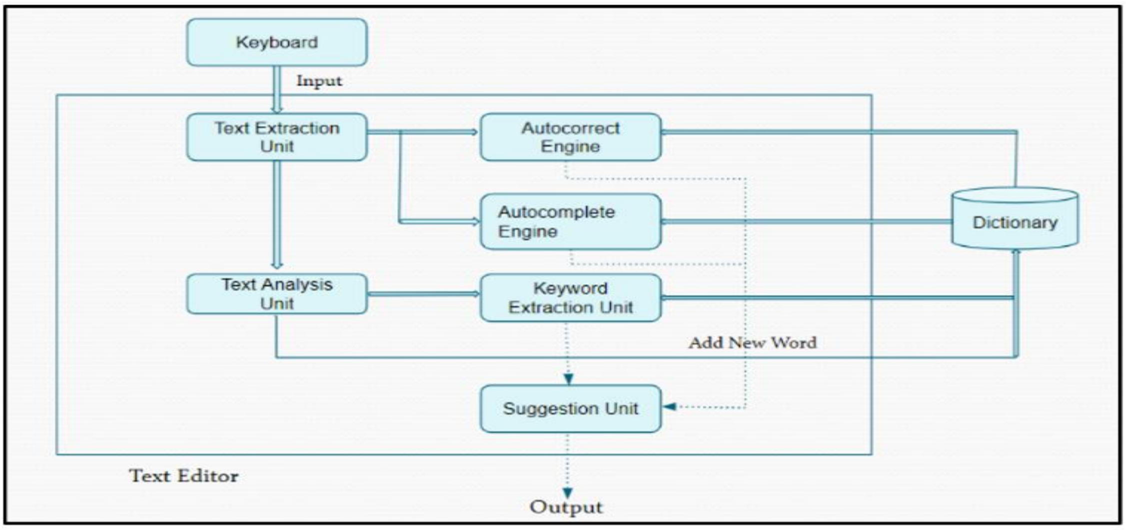


Figure 1. Proposed System

D. Dictionary

It is the main database used in the system. First, a linear list of words is compiled. The words are stored alphabetically and the frequencies of words is updated. Then the dictionary is created using this list to form a structure which looks like tree. The dictionary can be adapted as per the user.

E. Text Analysis and Keyword Extraction Unit

This unit extracts keywords from the entire text which is entered by the user. Based on analysis of these keywords the semantically and relevant words are suggested to the user.

F. Suggestion Unit

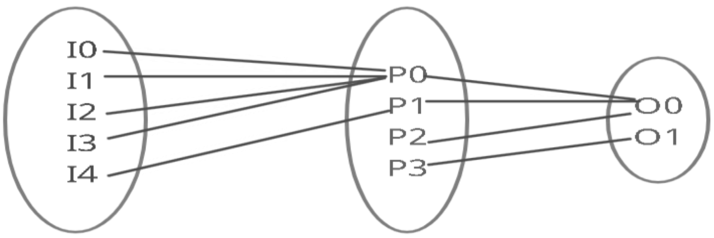
This unit is responsible for compiling the suggestions given by all units and then it displays the suggestions by ranking them.

Mathematical Model

Problem Statement : To predict/autocomplete next few words in an user document using RoBERTa technology.

Let S be the whole system.

- S ={I, P, O}
  I = Input
  P = Process
  O = Output
- I = {I0, I1, I2, I3, I4}
  I0 = User Registration
  I1 = User Email
  I2 = User Password
  I3 = User Login
  I4 = Word Input from User
- P = {P0, P1, P2, P3}
  P = Get word inputs from user.
  P1 = Provide next word autocomplete and autocorrect suggestions.
  P2 = Repeat suggestions until user stops or exits manually.
  P3 = Save the document.
- O= {O0, O1}
  O0 = User gets suggestions and saves keystrokes.
  O1 = Error-free and complete document is generated.



IV. PROPOSED ALGORITHM

RoBERTa (Robustly Optimized BERT) Algorithm -

Step1:Load Model and Tokenizer

First, Load RoBERTa model. Model will be in eval mode because we are not training model but using pre-trained model. Then, load tokenizer to tokenize the text.

Step2:Add mask at the end

As we want to predict the word in given text, so it is required to add a mask token at the end of input text because RoBERTa requires input to be pre-processed in this way.

Step3:Encode and Decode:

We use tokenizer to encode the input in this step. After encoding, the final step is to decode it.

Step4:Wrapper function for encoder and decoder:

Both encoding and Decoding functions we wrapped in wrapper function.

V. APPLICATIONS

A. In Web Browsers

In web browsers, we use auto-complete in the search and address bar. Autocomplete for web addresses is particularly convenient because the full addresses are often long and difficult to type correctly.

B. In email programs

In e-mail programs autocomplete is typically used to fill in the e-mail addresses of the intended recipients. Web addresses and e-mail addresses are often lengthy and difficult to remember and type, hence they seem inconvenient.

C. In source code editors

When a software senses a variable or string that is recognizable, it displays a menu to the programmer that contains the complete name of the identified variable or the methods applicable to the detected class, and the programmer makes a choice with her or his mouse or keyboard arrow keys.

D. In search engines

Autocomplete user interface features in search engines provide users with suggested queries or results as they type their query in the search box.

E. In Word Processor

In processing word programs, we can save keystrokes by using auto-complete on repetitive words and phrases.

F. In command-line interpreter

In command-line interpreter, autocomplete of command names and file names can be made possible by tracking of all the possible things a user may need. Auto-complete in command-line interpreter usually happens when the user presses Tab ⭾ key after typing the first few letters of a word.

VI. CONCLUSION

Text prediction and autocomplete can help people to increase their writing speed by predicting the relevant words. The model has proved efficient in prediction and completion of text, with RoBERTa attaining the highest recognition accuracy. The precision, f1 scores, and recall of the RoBERTa model demonstrate its superiority to other models. In the near future, various components of the RoBERTa model can be used to enhance the performance of auto-completion unit. Also, strategies to inculcate common sense knowledge into the model would be considered to improve its generalization ability.

VII. FUTURE WORK

With the shift to remote and digital literacy, a combined platform that meets students' effective literacy needs is required. Algorithm Visualizer is a combined platform that's a comprehensive result for educators and students to educate and learn online effectively. It has a heavy emphasis on "visualization of algorithms" ,which allows for a better insight of its flow and functions.

ACKNOWLEDGEMENTS

The success and outcome of this research required a lot of assistance, and we are extremely privileged to have got it all. All that we have accomplished is due to the excellent supervision and assistance we have received, for which we are grateful. We appreciate and thank Prof. Sarika Aundhakar, for providing us with knowledge and skill that greatly assisted the research. We are extremely grateful to her for her kind support and guidance.

VIII. REFERENCES

[1]. Text Optimization or Text Summarizer by using NLP by IRJET.
[2]. Predictive text computer simplified keyboard with word and phrase auto-completion' by David Gikandi.
[3]. Ani Nenkova, Kathleen McKeown – 'A Survey of Text Summarization Techniques'.
[4]. Ward D., Hahn J., and Feist K. (2012). Autocomplete as a research tool: a study on providing search suggestions. Information Technology and Libraries, 31(4), 6.
[5]. Berget G. and Sandnes F.E. (2015). The Journal of the Association for Information Science and Technology– 67(10), 2320-2328.
[6]. https://arxiv.org/abs/1907.11692