NLP SUMMER COURSE

# KNOCK KNOCK WHAT'S NEXT


BY:

HARINI ANAND - PES2UG21CS184

HIMAJA B - PES2UG21CS200

JAWAHAR BALACHANDHER - PES2UG21CS212

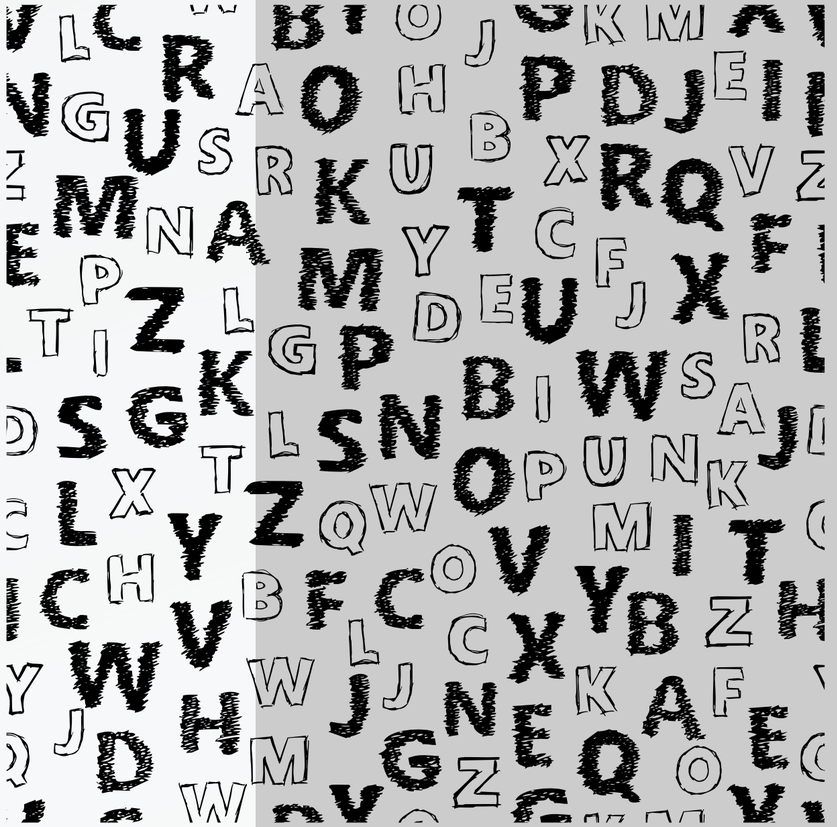
PRANAV DESAI - PES2UG21CS385

# CONTENT

# PROBLEM STATEMENT

89% of the current population faces issues in typing without errors and hence causing delays in extracting search queries.

Current chatbots and text summarization models lack the accuracy needed to provide top-notch results.



# CONCEPTS USED

**Tokenization**: The process of splitting text into smaller units, such as words or subwords, for further analysis.

**Embedding**: A dense representation of words or tokens in a continuous vector space, which is used to capture semantic relationships between words.

**LSTM** (Long Short-Term Memory): A type of recurrent neural network (RNN) architecture designed to capture long-term dependencies in sequential data.

**Dense Layer:** A fully connected layer in a neural network, where each neuron is connected to every neuron in the previous and subsequent layers.

**Sequential Model**: A linear stack of layers in a neural network, where data flows sequentially from the input layer to the output layer.

**Categorical Crossentropy:** A loss function commonly used for multiclass classification problems.

**Adam Optimizer:** An adaptive learning rate optimization algorithm used for training neural networks.

**Preprocessing**: Various data cleaning and transformation steps are performed on the text data to prepare it for modeling.

**Named Entity Recognition (NER):** The process of identifying and classifying named entities in text, such as person names, locations, and organizations.

**Stopwords:** Commonly used words (e.g., "the," "and," "is") that are typically removed from text data as they do not carry significant meaning for many natural language processing tasks.

**Part-of-Speech (POS) Tagging**: The process of assigning grammatical parts of speech to words in a sentence (e.g., noun, verb, adjective).

**FreqDist**: A class from the NLTK library used to compute the frequency distribution of words in a text.

**Callbacks:** In Keras, callbacks are used to perform additional actions during the training process, such as saving model checkpoints or early stopping based on certain conditions.

**NumPy**: A powerful library for numerical computing in Python, used here to manipulate arrays and numerical data.

**Pickle**: A Python module used for serializing and deserializing Python objects, used here to save and load the tokenizer.

**NLTK (Natural Language Toolkit):** A Python library for working with human language data, providing tools for tokenization, POS tagging, and more.

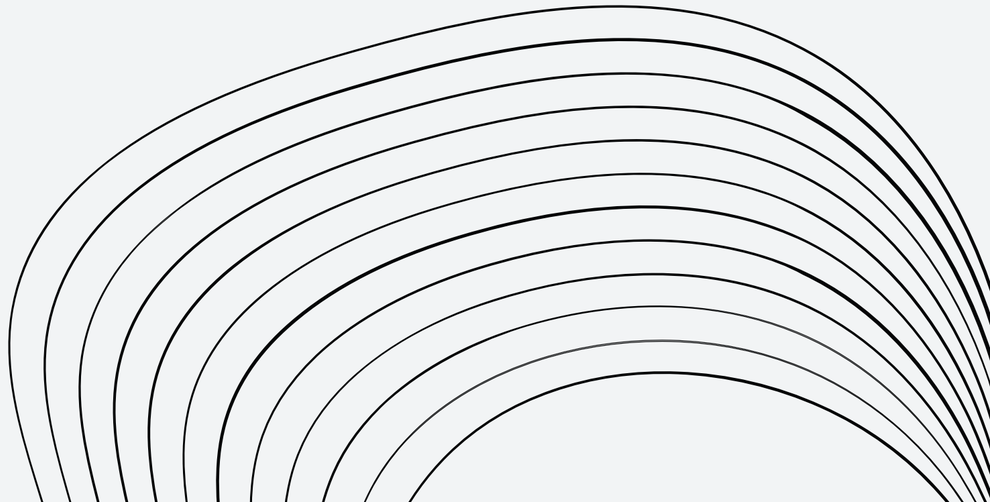
**Spacy**: An open-source library for natural language processing, including capabilities for NER and POS tagging.

Word Tokenization: The process of splitting a sentence into individual words or tokens.

**POS Tagging Visualization:** Displaying the dependency and named entity recognition visualization using SpaCy's displacy.

**Word Token Frequency**: Counting the frequency of each word/token in a given text.

**Model Visualization:** Creating a visualization of the sequential model using Keras' plot_model function.



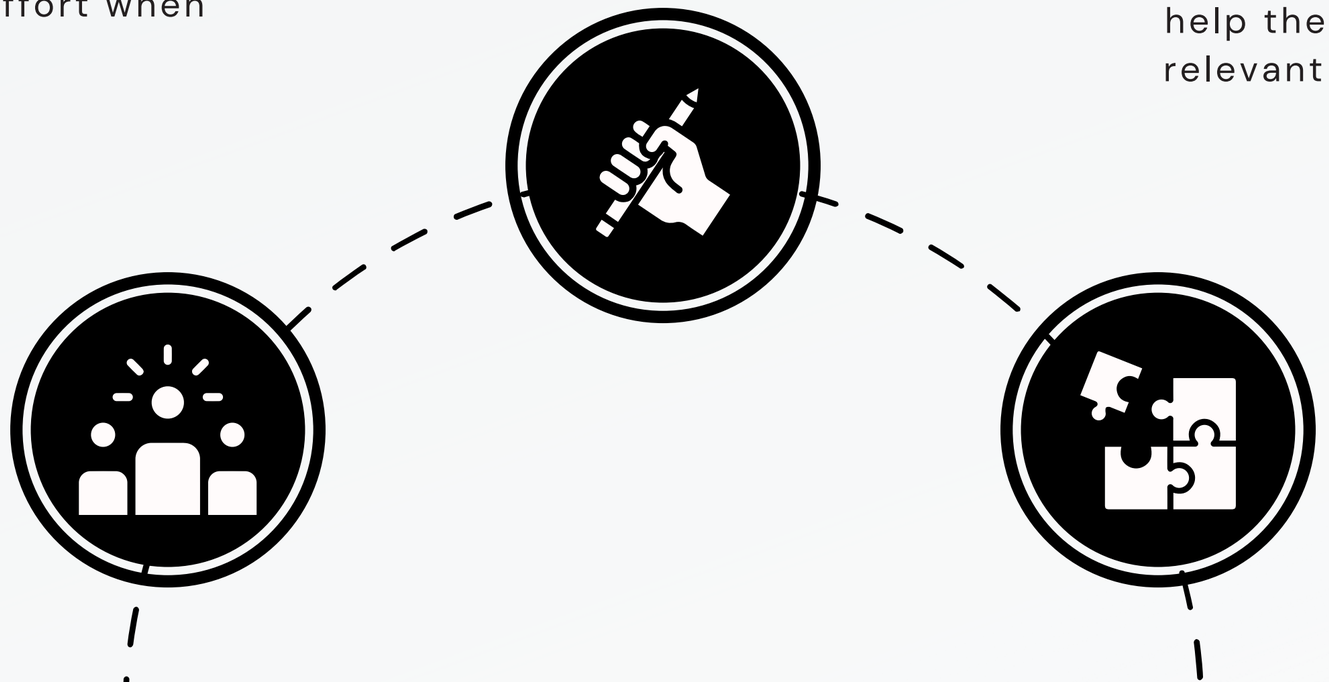
# USE CASES

## Autocomplete Suggestions

Used to provide autocomplete suggestions in text editors and search engines. This can help the user to save time and effort when typing.

## Language Translation

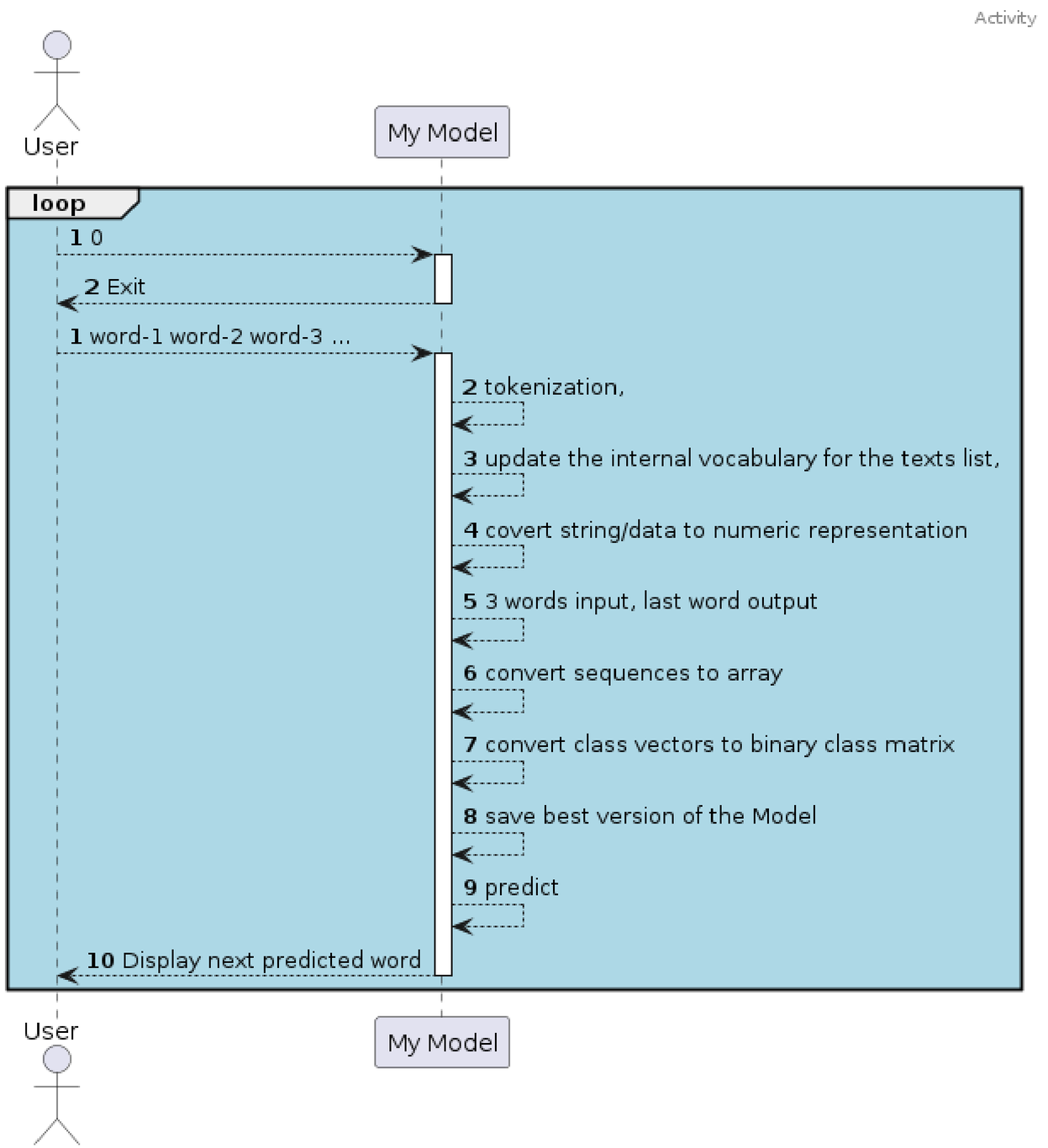
Used to improve the accuracy of language translation models



## Chatbots

As the chatbot is interacting with a user, it can use next-word prediction to predict the user's next question or statement. and help the chatbot to provide more relevant and helpful responses to the user.

# HIGH LEVEL ARCHITECTURE




Jawahar Balachandher, Pranav Ajay Desai, Harini Anand, Himaja B
B.Tech CSE, 2025 Batch
PES University, Electronics City Campus, Bangalore

# DATASET USED

- Used text of classic fables like Romeo and Juliet which is an uncleaned dataset containing textual extracts of the fable.

The Project Gutenberg eBook of Romeo and Juliet, by William Shakespeare

This eBook is for the use of anyone anywhere in the United States and most other parts of the world at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org. If you are not located in the United States, you will have to check the laws of the country where you are located before using this eBook.

Title: Romeo and Juliet

Author: William Shakespeare

Release Date: November, 1998 [eBook #1513]
[Most recently updated: June 27, 2023]

Language: English

Produced by: the PG Shakespeare Team, a team of about twenty Project Gutenberg volunteers

*** START OF THE PROJECT GUTENBERG EBOOK ROMEO AND JULIET ***

# EXPECTED OUTPUT

The expected output of a next-word prediction model would be a list of the most likely next words, ranked in order of probability.

Entering in input words or a string of words

THE CAT SAT ON THE

The accuracy of the model's predictions would depend on the size and quality of the training data and the complexity of the mode

MODEL PREDICTION

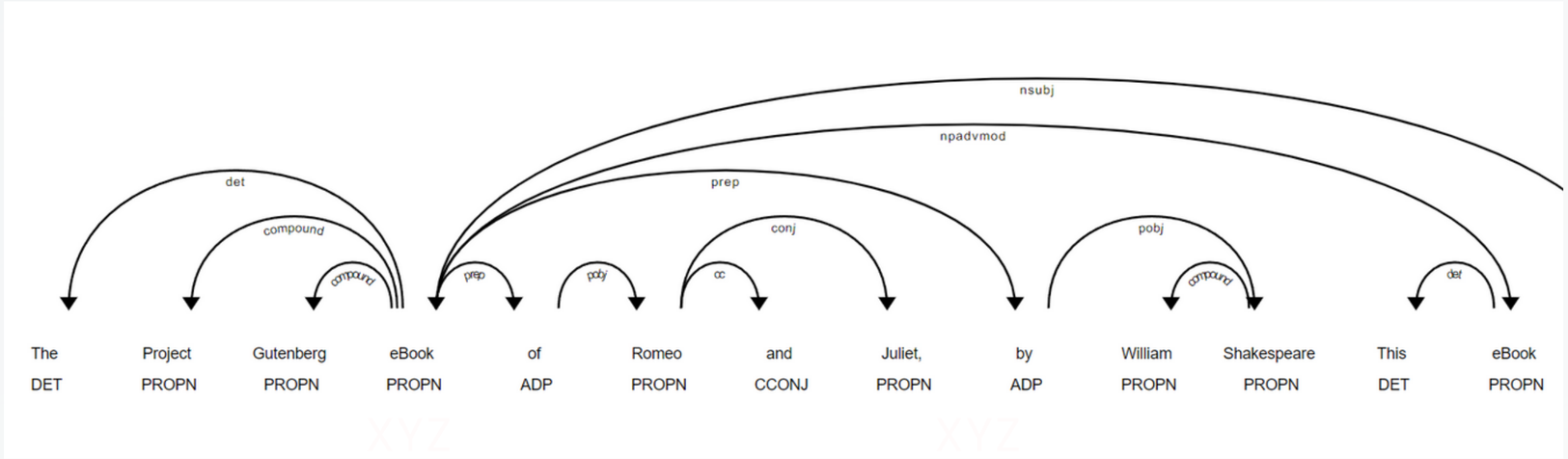
The expected output can be a range of words related to the context

"MAT", "FLOOR", "RUG"

# DEMO

# PROJECT OUTPUTS


wordcloud




POS


DisplaCy

# Success Test case

```
    print("Exit")
    break

  else:
    try:
      text = text.split(" ")
      text = text[-3:]
      print(text)

      Predict_Next_Words(model,tokenizer,text)


    except Exception as e:
      print("Error occurrred: ",e)
      continue
```

```
Enter: where you are
['where', 'you', 'are']
1/1 [==============================] - 1s 704ms/step
located
Enter:
```

romeo_&_juliet.txt

File Edit View

The Project Gutenberg eBook of Romeo and Juliet, by William Shakespeare

This eBook is for the use of anyone anywhere in the United States and most other parts of the world at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org. If you are not located in the United States, you will have to check the laws of the country where you are located before using this eBook.

# Termination Test case

```
while(True):
  text=input("Enter: ")

  if text=="0":
    print("Exit")
    break

  else:
    try:
      text = text.split(" ")
      text = text[-3:]
      print(text)

      Predict_Next_Words(model,tokenizer,text)


    except Exception as e:
      print("Error occurrred: ",e)
      continue
```

```
Enter: 0
Exit
```

# FUTURE SCOPE

Next word prediction is a valuable tool that can be used to improve a variety of tasks.
It is a powerful NLP technique that has the potential to benefit a wide range of users.

Potential Applications:

- Writing assistants
- Machine translation
- Productivity tools

# REFERENCES

https://en.wikipedia.org/wiki/Natural_language_processing

https://plantuml.com/

# THANK YOU!