** **

**DEPARTMENT**

**OF**

**COMPUTER SCIENCE AND ENGINEERING**

**ANNA UNIVERSITY, Chennai**

**CS6030: NATURAL LANGUAGE PROCESSING**

**Documentation**

**Named Entity Recognition for Tanglish Text**

Submitted by :

Mohamed Kani P K

2021103303

Batch : ‘P’

**PROBLEM STATEMENT:**

In multilingual environments such as Tamil Nadu, India, where Tanglish a blend of Tamil and English is commonly used in everyday communication, there exists a significant challenge for natural language processing (NLP) systems. The unique code-mixing nature of Tanglish, characterized by the seamless integration of Tamil and English elements within sentences, poses obstacles for existing NLP tools. These tools struggle to accurately identify and classify named entities within Tanglish text due to the linguistic complexities arising from variations in script, syntax, and cultural nuances. This limitation impacts the effectiveness of NLP applications such as sentiment analysis, information extraction, and language understanding, hindering their ability to process and interpret mixed-language data accurately.

To address this challenge, the project focuses on developing a specialized Named Entity Recognition (NER) system tailored explicitly for Tanglish text. By leveraging advanced NLP techniques and training models on annotated Tanglish datasets, the objective is to enhance the capabilities of NLP systems in accurately identifying and categorizing named entities such as PERSON, LOCATION, ORGANIZATION, and others within Tanglish text. The specialized NER system aims to overcome the code-mixing complexities and linguistic variations inherent in Tanglish, thereby improving the accuracy and performance of NLP applications for multilingual environments.

**INTRODUCTION:**

The project endeavors to overcome the challenges faced by current Natural Language Processing (NLP) systems through the creation of a sophisticated Named Entity Recognition (NER) system tailored specifically for Tanglish text. In NLP, NER plays a pivotal role in identifying and classifying entities within textual data, including but not limited to PERSONs, LOCATIONs, ORGANIZATIONs, PRODUCT, TIME expressions, MONEY denominations, EVENTs, DATEs, and more. However, the inherent complexities of Tanglish—stemming from its amalgamation of Tamil and English—pose significant hurdles for existing NLP tools. These challenges arise from code-mixing intricacies, syntactic variations, and cultural nuances unique to Tanglish.

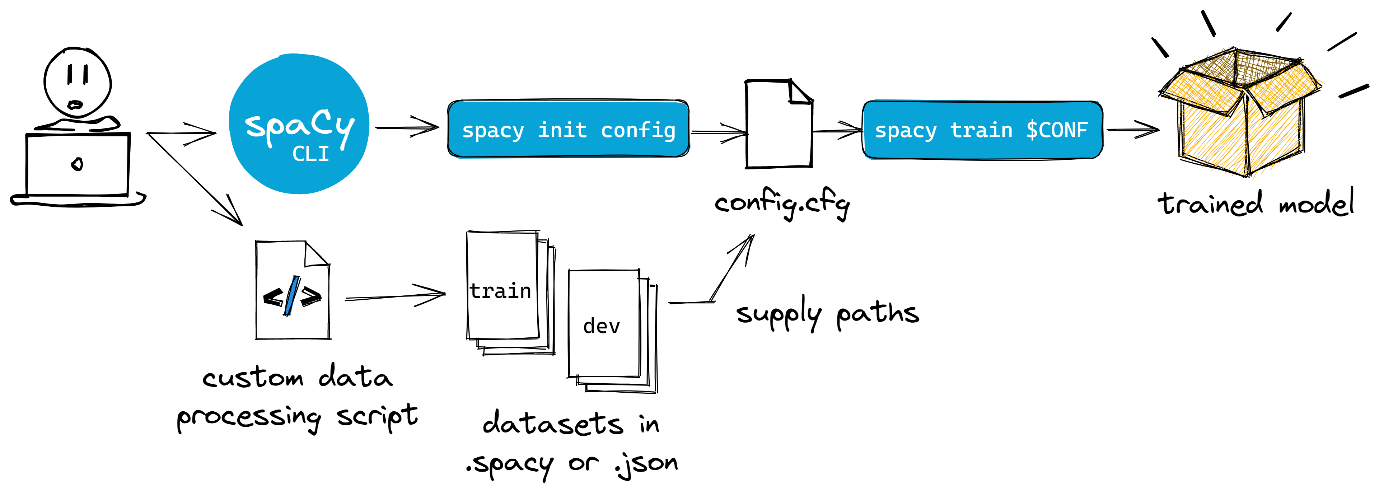
**OBJECTIVE:**

To develop comprehensive annotation guidelines for annotating Tanglish text with named entities. These guidelines will provide clear instructions and standards for annotators to identify and label entities such as PERSON, LOCATION, ORGANIZATION, etc., ensuring consistency and accuracy in the annotated data used for training the NER system.

**NEED FOR THE PROJECT:**

1. **Overcoming Code-Mixing Challenges:** The project addresses the complexities of Tanglish, a code-mixed language comprising Tamil and English, by developing a specialized Named Entity Recognition (NER) system. This system aims to accurately identify and classify named entities within Tanglish text, overcoming the linguistic intricacies associated with code-mixing.
2. **Enhancing NLP Application Accuracy:** By improving entity recognition in Tanglish, the project enhances the accuracy and effectiveness of Natural Language Processing (NLP) applications such as sentiment analysis, information extraction, and language understanding. This ensures more reliable and context-aware analysis of mixed-language data in multilingual environments.
3. **Catering to Multilingual Communication Needs:** The project caters to the growing need for NLP solutions that can effectively process mixed-language data, particularly in regions like Tamil Nadu, India, where Tanglish is prevalent. Developing a robust NER system for Tanglish text contributes to better linguistic analysis, information retrieval, and communication understanding in Tanglish-speaking communities.

**BLOCK DIAGRAM:**



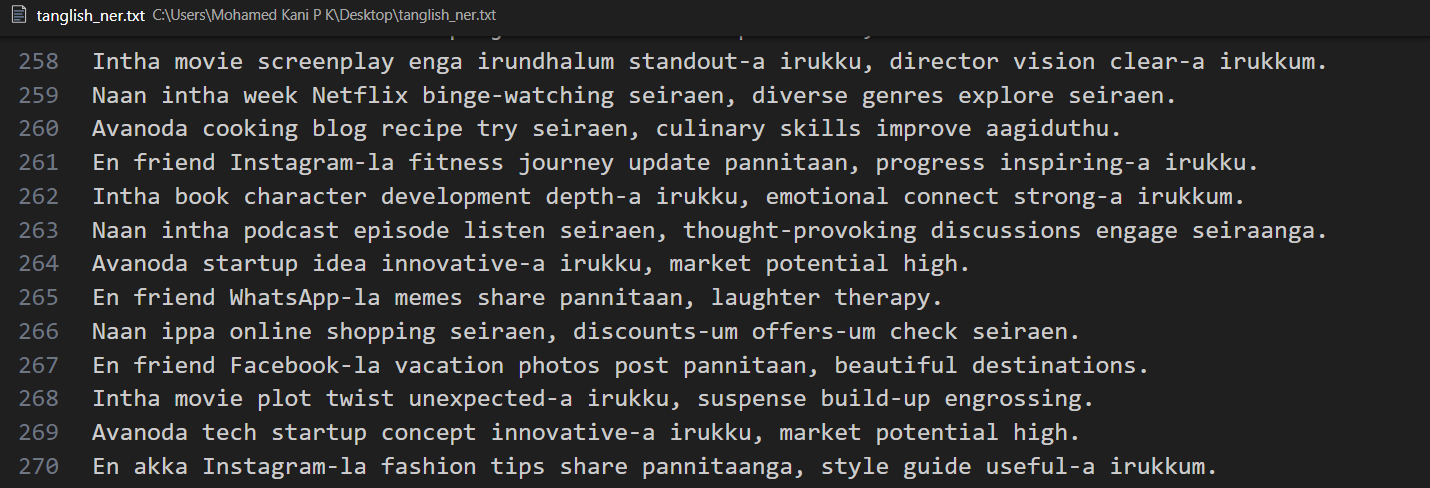
The block diagram illustrates the flow of processes in a Named Entity Recognition (NER) project for Tanglish text. It begins with the input of raw Tanglish text, which is loaded from an annotated JSON file. The data then undergoes processing for training, where it is prepared into suitable training data. This processed data is used to train the NER model using the SpaCy library. The trained model is saved to disk for future use. During text processing, the trained model performs inference on new Tanglish text data to recognize and label entities. The final output includes the Tanglish text with entity labels indicating recognized named entities and their types.

This block diagram encapsulates the essential steps of data loading, processing, model training, inference, and output generation in the NER project for Tanglish text, as described in your project description and code.

**IMPLEMENTATION:**

**Data Collection:**

* Gather data from social media platforms, such as Twitter, Facebook, or Instagram. This data may include user comments, posts, or conversations.
* Collect product reviews from e-commerce websites or review platforms. This data can provide insights into customer sentiments and opinions.
* Use ChatGPT or similar conversational AI platforms to generate text data. This data can simulate natural language conversations and interactions.



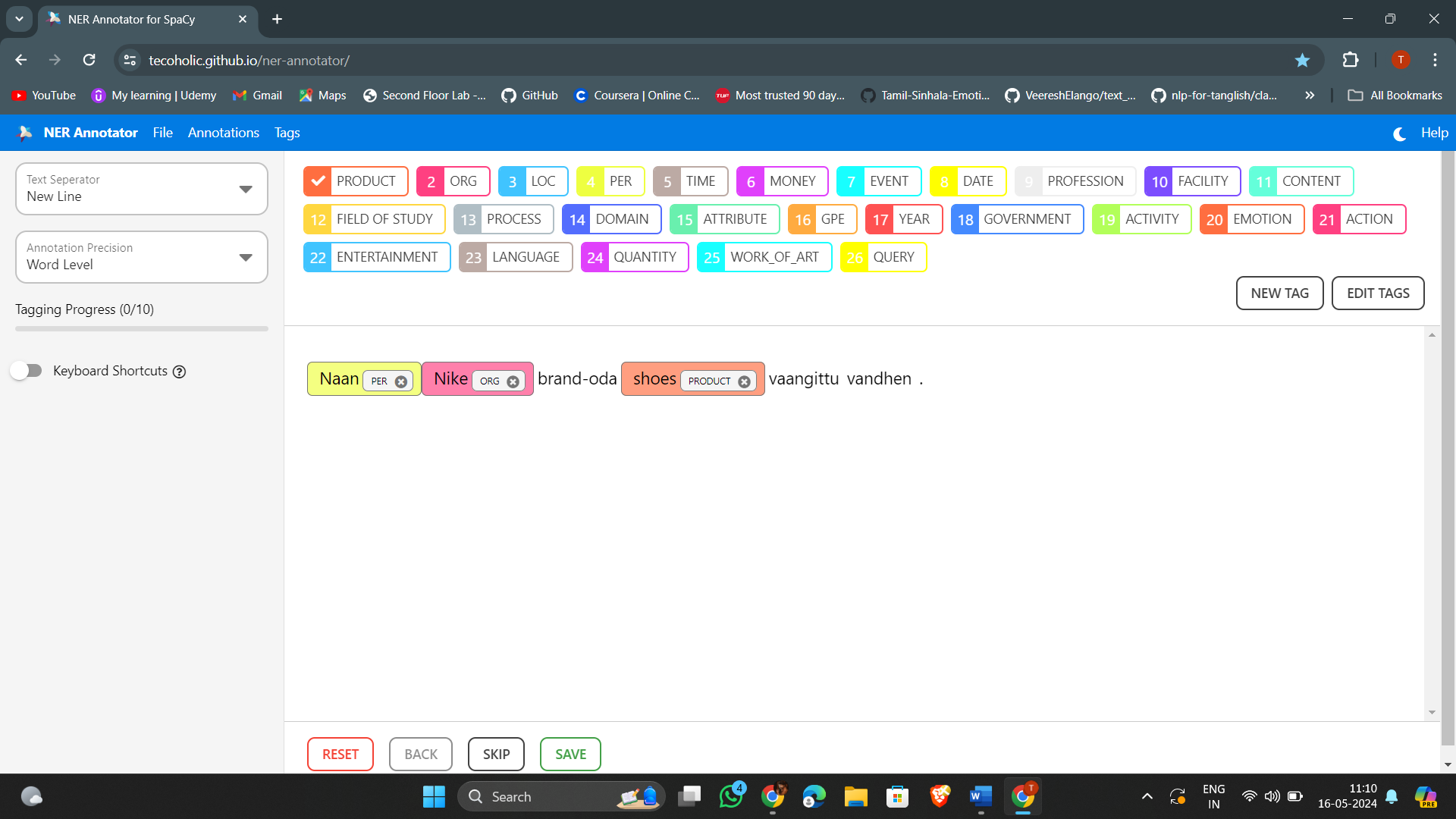
*Fig. 1 Sample dataset*

**Data Annotation**:

Annotate the collected dataset with entity labels using a Named Entity Recognition (NER) annotator tool. Label entities such as PERSON, LOCATION, ORGANIZATION, PRODUCT, TIME, MONEY, EVENT, DATE, etc.

Ensure consistency and accuracy in the annotation process to improve the quality of the training data.

* We used tool named as NER Annotator Tool to annotate the entities.



* The Entities used for annotate the data are

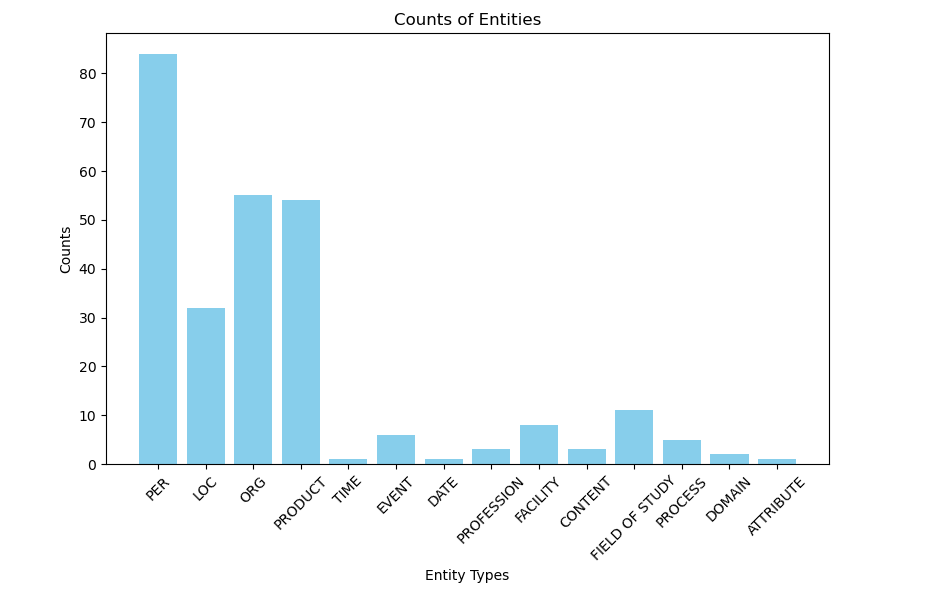
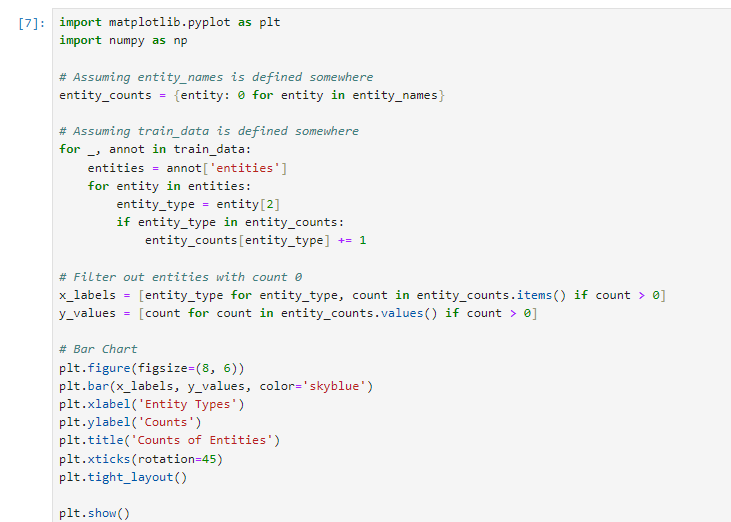
"classes":["PER","LOC","DATE","TIME","GPE","EVENT","MONEY","LANGUAGE","LAW","PRODUCT","QUANTITY","WORK\_OF\_ART","ORG","CARDINAL","FACILITY"]

* The annotated data is saved as .json format.

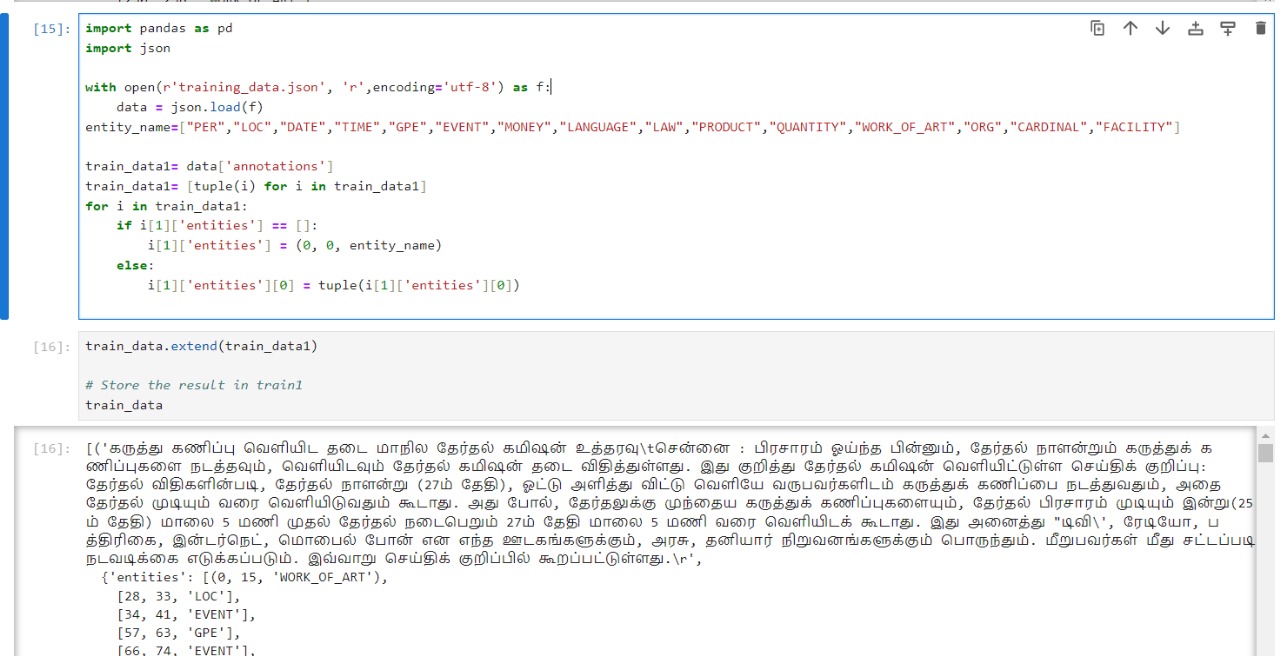


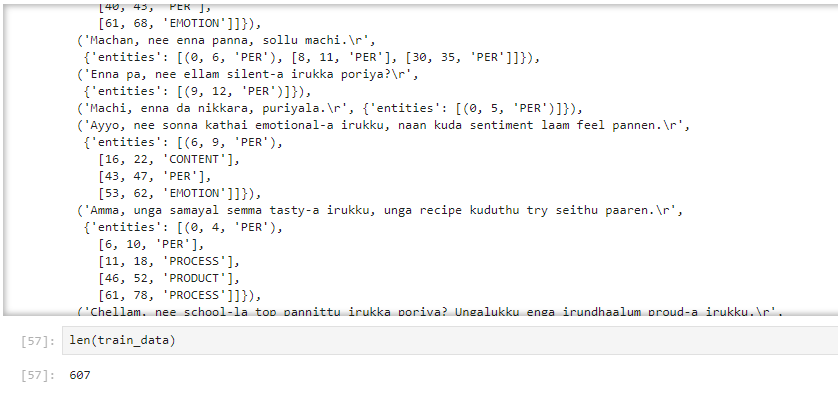
1. **Model Development**: Develop an NER model using Spacy v3 that can predict entities based on the context of the text.

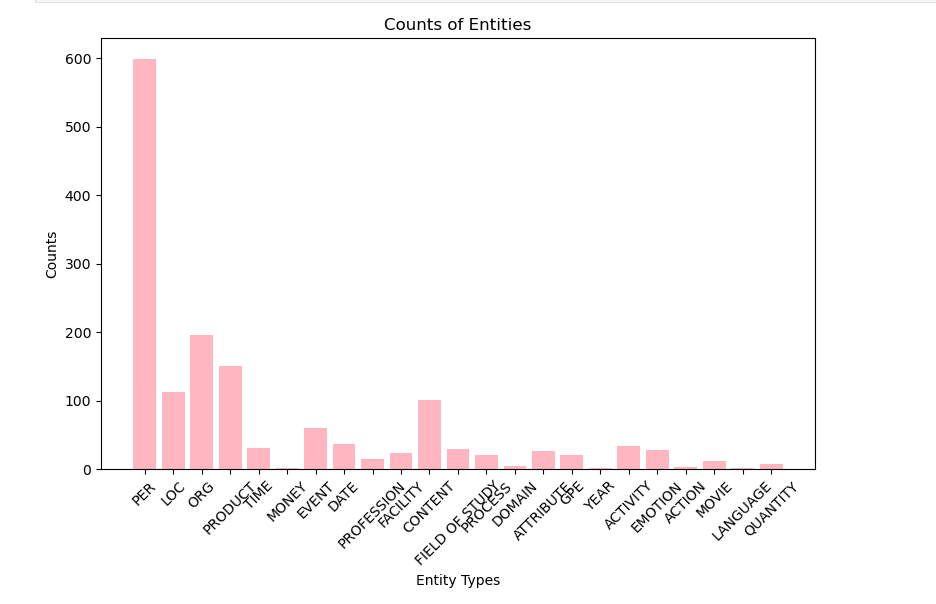


* Append the another .json file in the already existing model.





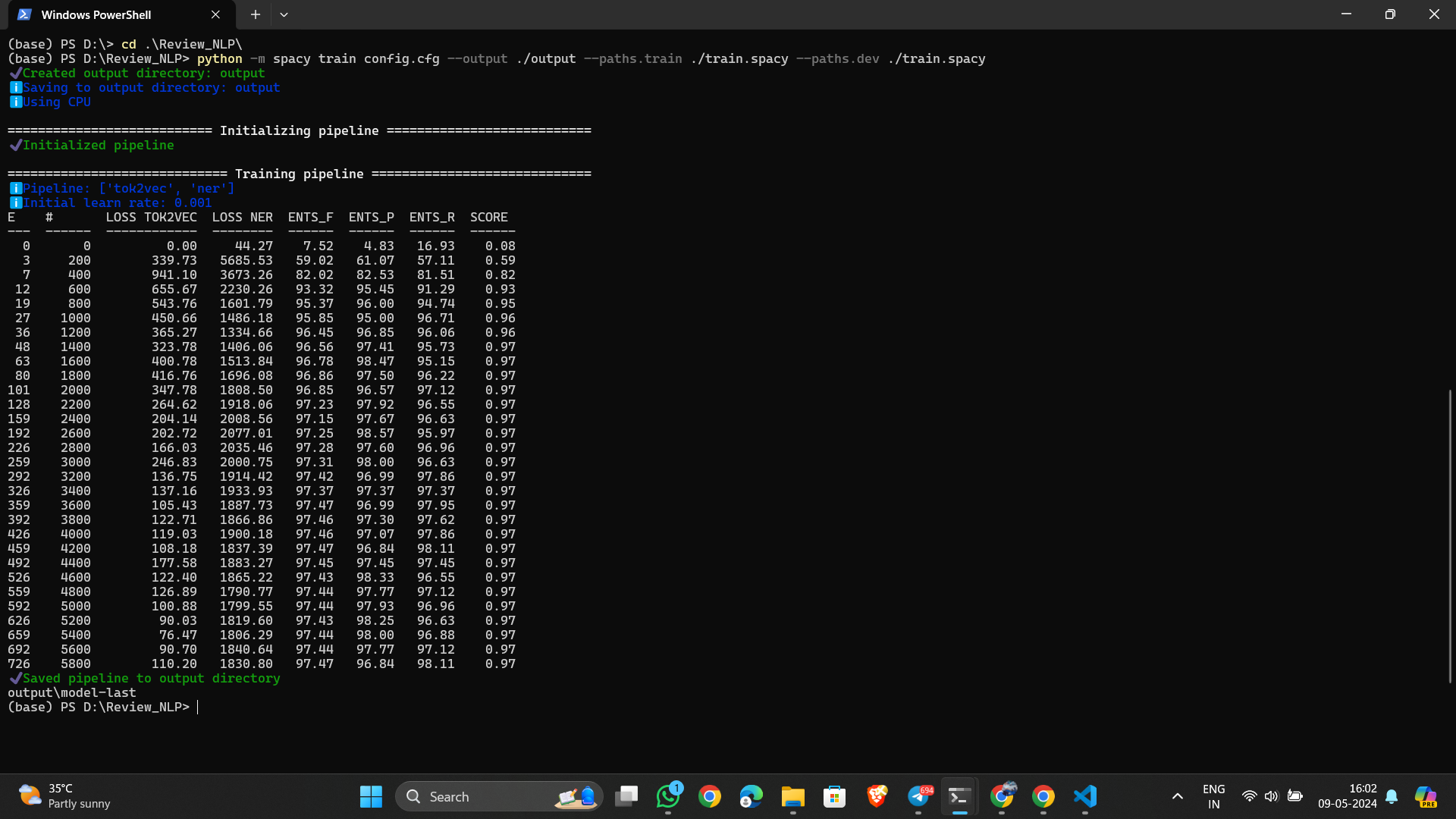


1. **Model Training and Evaluation**: Train the model on the annotated dataset and evaluate its performance using metrics such as Precision, Recall, and F1 Score.



* Training the model using pipeline.

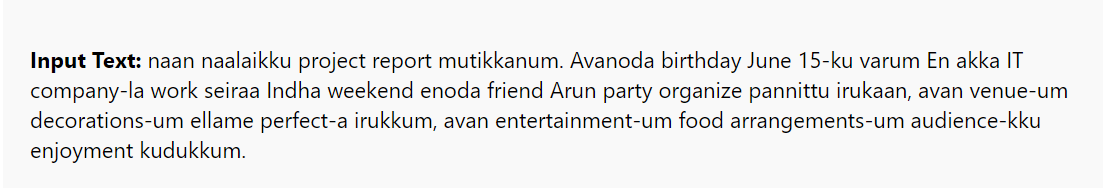


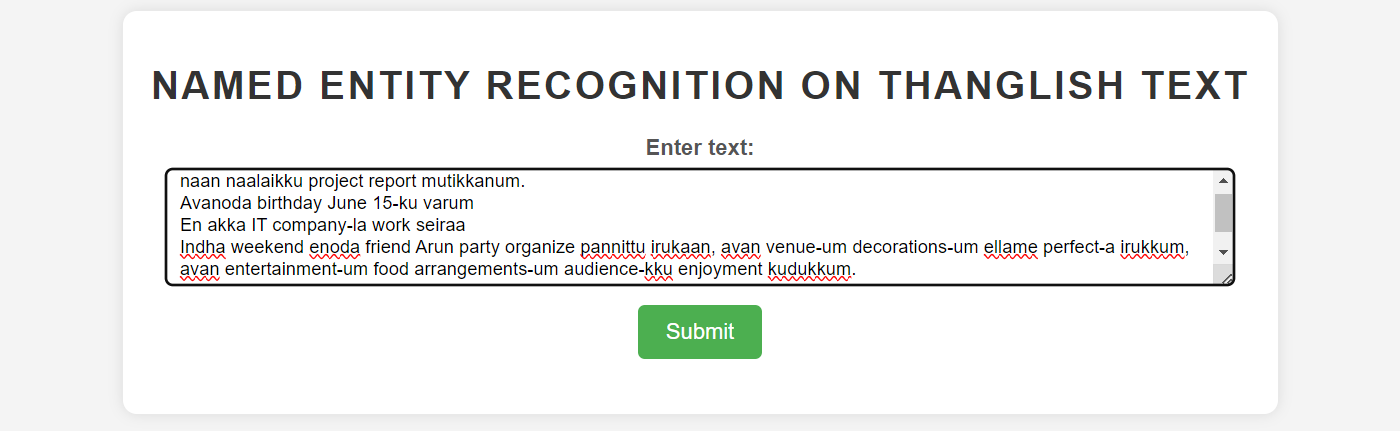


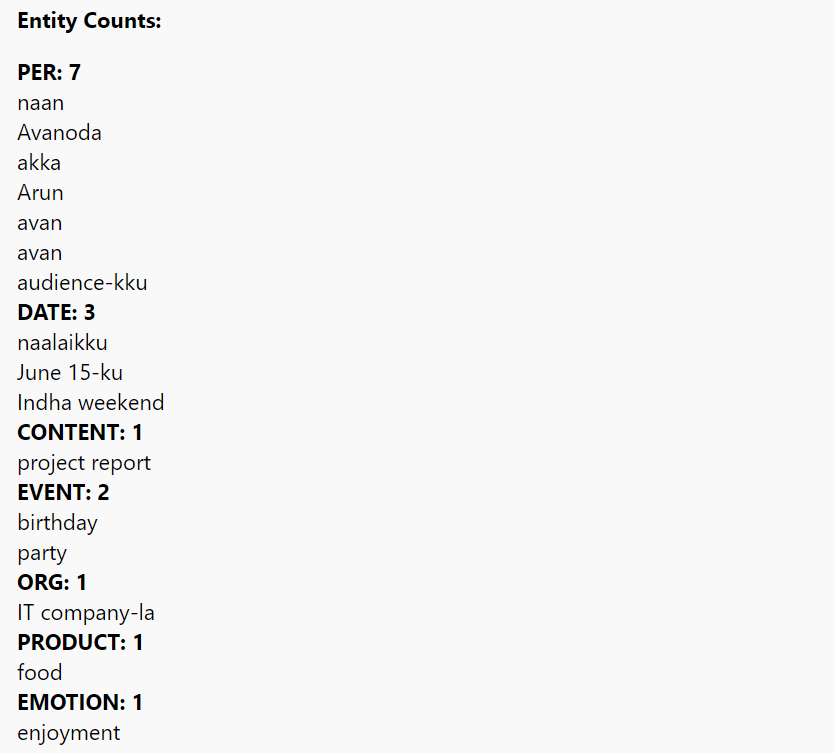
1. **Integration and Deployment**: Integrate the trained NER model into a suitable platform for testing and deploy it for real-time tagging of tanglish text.



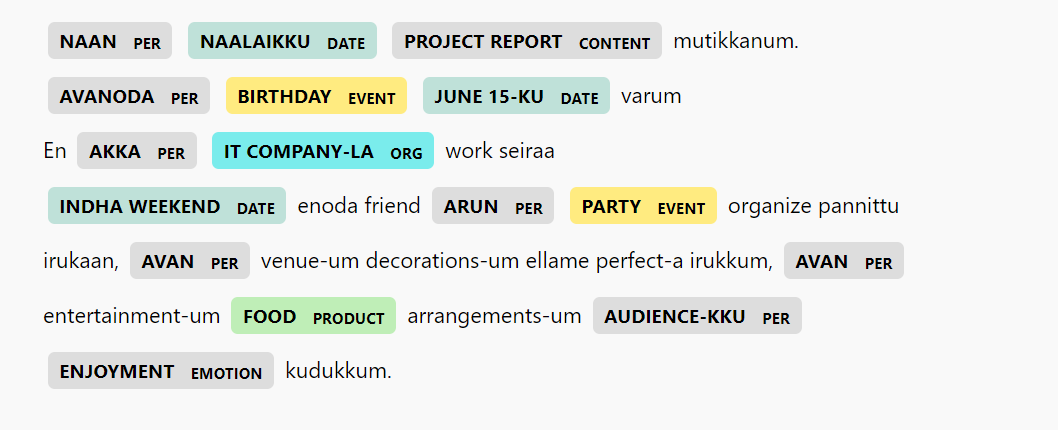
**RESULT:**

****

****

****

**OUTPUT:**

****

**FUTURE SCOPE:**

The successful completion of the NER project opens up several avenues for future research and development in the field of NLP and multilingual text processing. Some potential areas for future exploration and enhancement include:

* Domain-Specific NER: Extend the NER model to recognize domain-specific entities relevant to specific industries or applications, such as healthcare, finance, legal, or technology.
* Fine-Tuning and Optimization: Continuously fine-tune and optimize the NER model to improve its accuracy, speed, and performance on diverse Tanglish text datasets.
* Multimodal NER: Explore the integration of multimodal data sources, such as images, videos, or audio, to enhance entity recognition capabilities and provide a more comprehensive understanding of text content.
* Named Entity Linking: Extend the NER system to perform named entity linking, which involves linking recognized entities to knowledge bases or databases for additional context and information retrieval.

**CONCLUSION:**

The completion of the Named Entity Recognition (NER) project for Tanglish text marks a significant milestone in leveraging natural language processing (NLP) technologies for multilingual environments. Through the development and training of a robust NER model, the project has successfully demonstrated the ability to identify and classify entities such as PERSON, LOCATION, ORGANIZATION, PRODUCT, TIME, MONEY, EVENT, DATE, and others within Tanglish text.

The project's implementation involved collecting a diverse dataset from social media, product reviews, and ChatGPT, annotating the dataset with entity labels, preprocessing the data, training the NER model using the SpaCy library, and evaluating the model's performance. The trained NER model can now accurately recognize named entities in Tanglish text, thereby enhancing the accuracy and usability of various NLP applications in multilingual contexts.

**REFERENCES:**

1. Chiu, Jason, and Eric Nichols. "Named Entity Recognition with Bidirectional LSTM-CNNs." Transactions of the Association for Computational Linguistics 4 (2016): 357-370.
2. SpaCy Documentation. https://spacy.io/api/doc.