NER

FOR DISEASE DETECTION

PRESENTED BY,
VIDHYA C VIJAYAN
FIT22MCA-2118
S4 MCA -B

INTRODUCTION

- (NER) Named Entity Recognition
- Vital task in natural language processing (NLP)
- Identifying and classifying named entities in text into predefined categories.
- Extract specific information such as names of people, organizations, locations, and in our case, diseases mentioned in text data.
- Automatically identify and extract mentions of diseases or health-related terms from various sources of textual data.
- Aids healthcare professionals, researchers, and public health authorities in analyzing large volumes of text data efficiently.

EXISTING SYSTEM

- Relied heavily on rule-based methods .

- Creation of explicit linguistic rules and pattern matching strategies to detect specific entities like person names, organization names, location etc. (capitalized words -proper nouns - person names /place names)

- Identifying and classifying named entities in text into predefined categories.

- Dictionaries and lookup tables to match words or phrases against predefined lists of named entities, enabling the identification of common entities within text.
- Limited scalability & adaptability

- Conditional Random Fields (CRF),Hidden Markov Models (HMM)

PROPOSED SYSTEM

- Utilizes simpletransformers library and BERT for NER.
- Automate the identification of Disease, GPE & Cardinal entities from text .
- Focus on supporting disease surveillance and health monitoring efforts.
- Process multiple sentences , extract relevant entities & present final results in a structured table format - easy interpretation and utilization.
- BERT -state-of-the-art transformer-based model known for its contextual understanding of language.
- Processes the input text to identify and classify entities based on their contextual relationships within the text.
- Implemented as a web application allows users to input text easily, view the extracted entities in a clear and organized manner.

OBJECTIVES

- To develop a dynamic and innovative website that harnesses Machine Learning and AI technologies to analyze news text inputted by users.
- To automatically identify and extract key entities from the text, specifically focusing on diseases, geopolitical entities (GPE), and locations.
- To extract informations and presented it in a structured table format, providing clear and easy-to-understand.
- To give insights into disease outbreaks and aiding health departments in taking necessary actions.

RELEVANCE OF NER IN PUBLIC HEALTH MANAGEMENT

- **Early Detection of Disease Outbreaks**: Enables swift identification and containment of outbreaks, preventing them from escalating.
- **Enhanced Disease Surveillance**: Automates analysis of large text data, facilitating better monitoring of disease trends and high-risk populations.
- **Optimized Resource Allocation**: Provides insights for efficient allocation of medical resources to areas with higher disease burdens.
- **Informed Decision-Making**: Empowers policymakers with timely and accurate data for evidence-based interventions and policies.
- **Global Health Security**: Contributes to early detection and containment of diseases, reducing the risk of international spread and pandemics.
- **Health Equity Promotion**: Helps address health disparities by ensuring access to timely diagnosis and treatment for all communities

DATASET

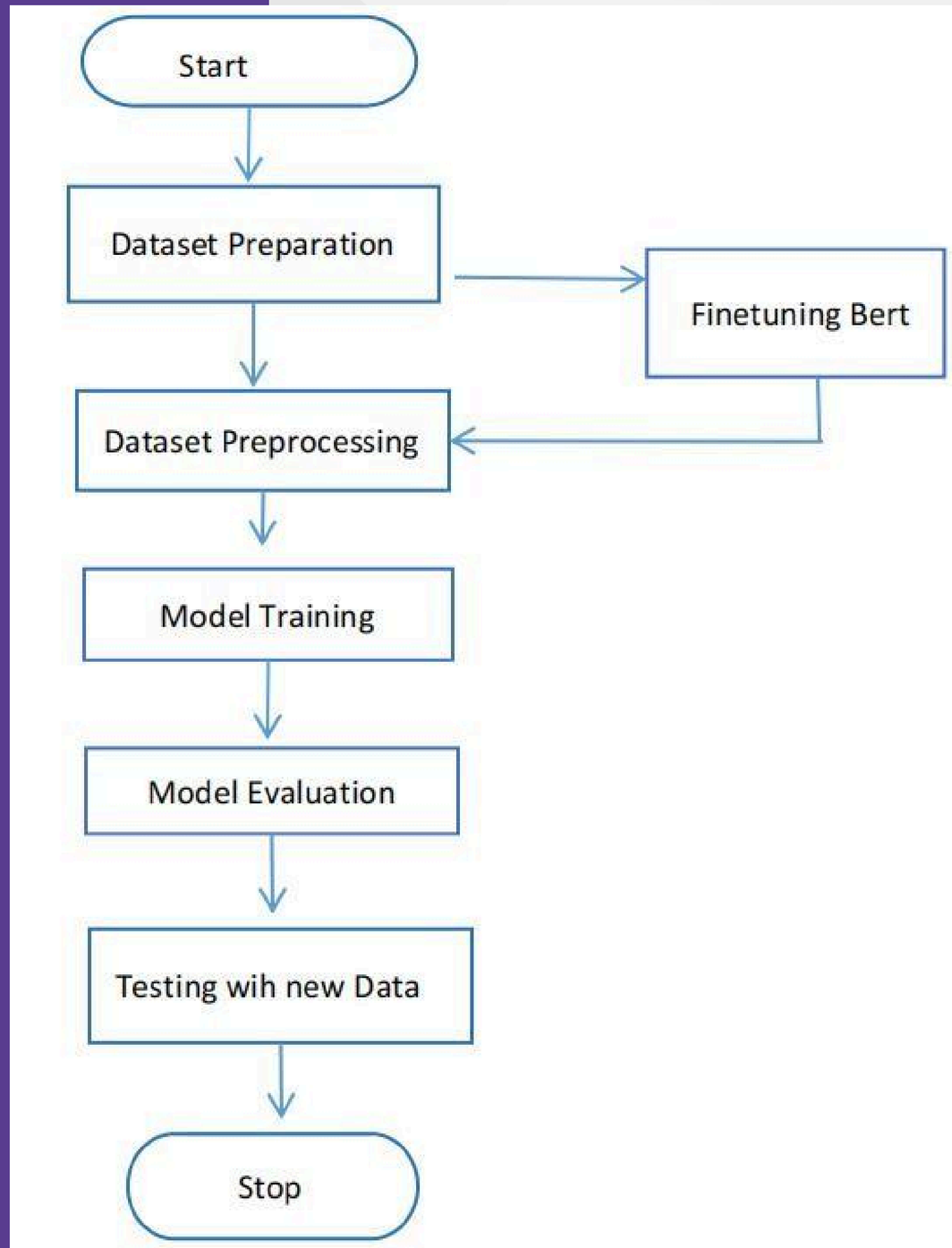
- The dataset contains a total of 3,000 rows.
- Contains sentences related to disease outbreaks, including leptospirosis, African swine fever, and COVID-19 cases.
- Primarily focuses on the context of Kerala, a state in India.
- Each sentence is tagged with part-of-speech (POS) tags and labeled entities, indicating named entities such as diseases, events, locations, and dates.
- Structured for NLP tasks such as Named Entity Recognition (NER) and information extraction

DATASET

- Sentence #: reference to the sentence number
- texts: actual text of each sentence.
- pos_tags: marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.
- text_labels: entity tags for each word in the sentence.
- IOB (Inside, Outside, Beginning) tagging format for tagging tokens
- B-X (Beginning)- first token of a named entity of type X.
- I-X (Inside) - token inside a named entity of type X
- O (Outside) denotes a token that does not belong to any named entity.

IMPLEMENTATION

- **Front-end Interface**: React.js framework within Visual Studio Code, offers user-friendly web application interface. Users input large text data for analysis.
- **Entity Extraction**: The system employs Named Entity Recognition (NER) to extract disease names, geographical locations, and numerical values from the input text.
- **Structured Presentation**: Extracted entities are presented in a structured table format, enhancing comprehension and facilitating prompt decision-making. This organized presentation simplifies the interpretation of results.
- **Backend Implementation**: Flask, a lightweight web framework for Python, enables integration of the NER model into the web application.
- **Real-time Analysis**: The NER model, likely based on BERT (Bidirectional Encoder Representations from Transformers), processes text data in real-time.



ARCHITECTURE

MODULES

1

Data Collection:

- Sources chosen for text data collection.
- Extraction methods including scraping, APIs, or digitization.
- Data organization into sentences, POS tagging, and entity labeling.
- Dataset compilation and formatting for machine learning models.

2

User Input:

- System receives text inputs from users.
- Inputs can include medical reports, news headlines, or any text with relevant entities.
- Text parsed into sentences for analysis.

3

Text Analysis and NER Implementation:

- Cleaning and tokenization of text.
- Normalization techniques applied for standardization.
- BERT model used for contextual representation.
- Extracted entities classified into predefined categories based on context.

4

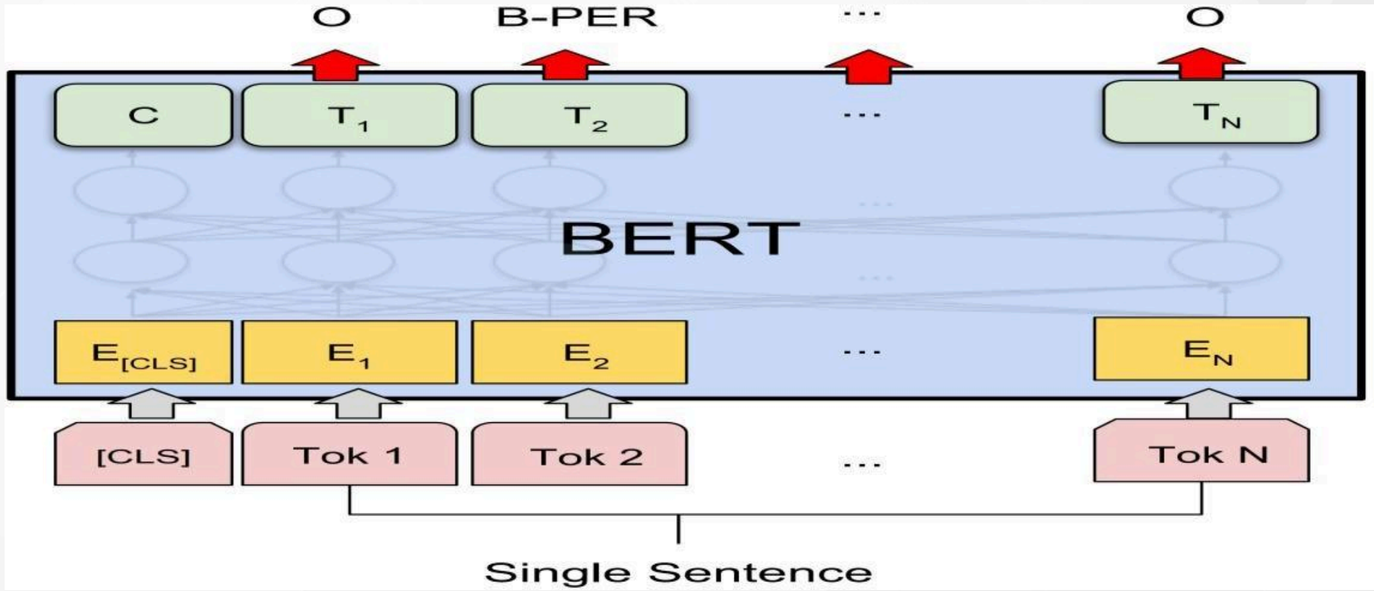
Testing With New Data:

- Structured output generated alongside original input.
- Entities highlighted and categorized in a table format.
- User-friendly interface for input and result visualization.
- Option to save results as CSV or print for sharing.

BERT

ALGORITHM

- **Bidirectional Encoder Representations from Transformers**
- Contextual Understanding
- Bidirectional Learning
- Transformer Architecture
- Pre-training Tasks
- Fine-tuning
- Performance Impact
- Broader Applications
- Influence on Model Building



RESULT

Result		Score
'eval_loss'	:	0.1887286
'precision'	:	0.8179862
'recall'	:	0.8326763
'f1_score'	:	0.8252658

Fig Result Evaluation

OUTPUT

SCREENSHOTS

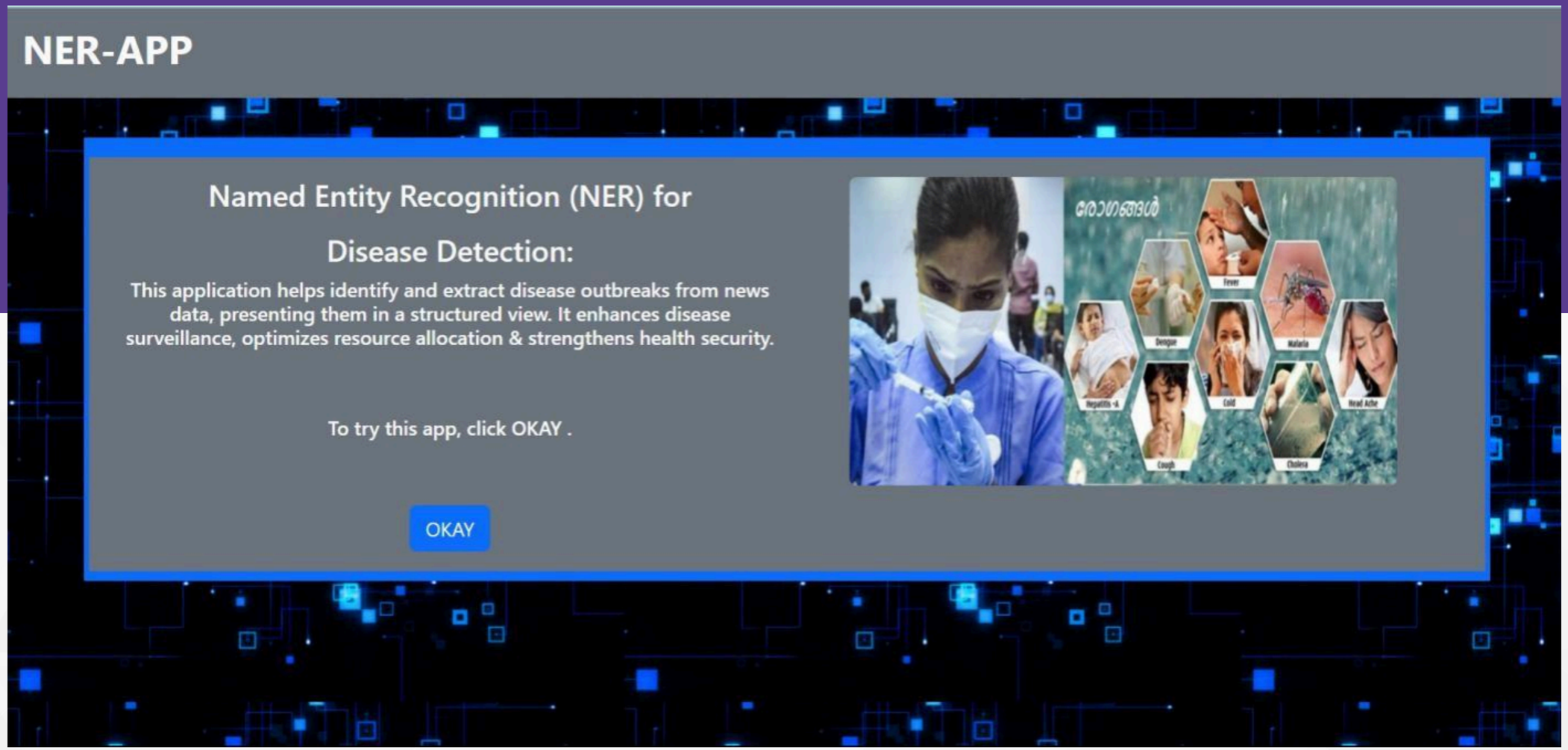


Fig :Home Page

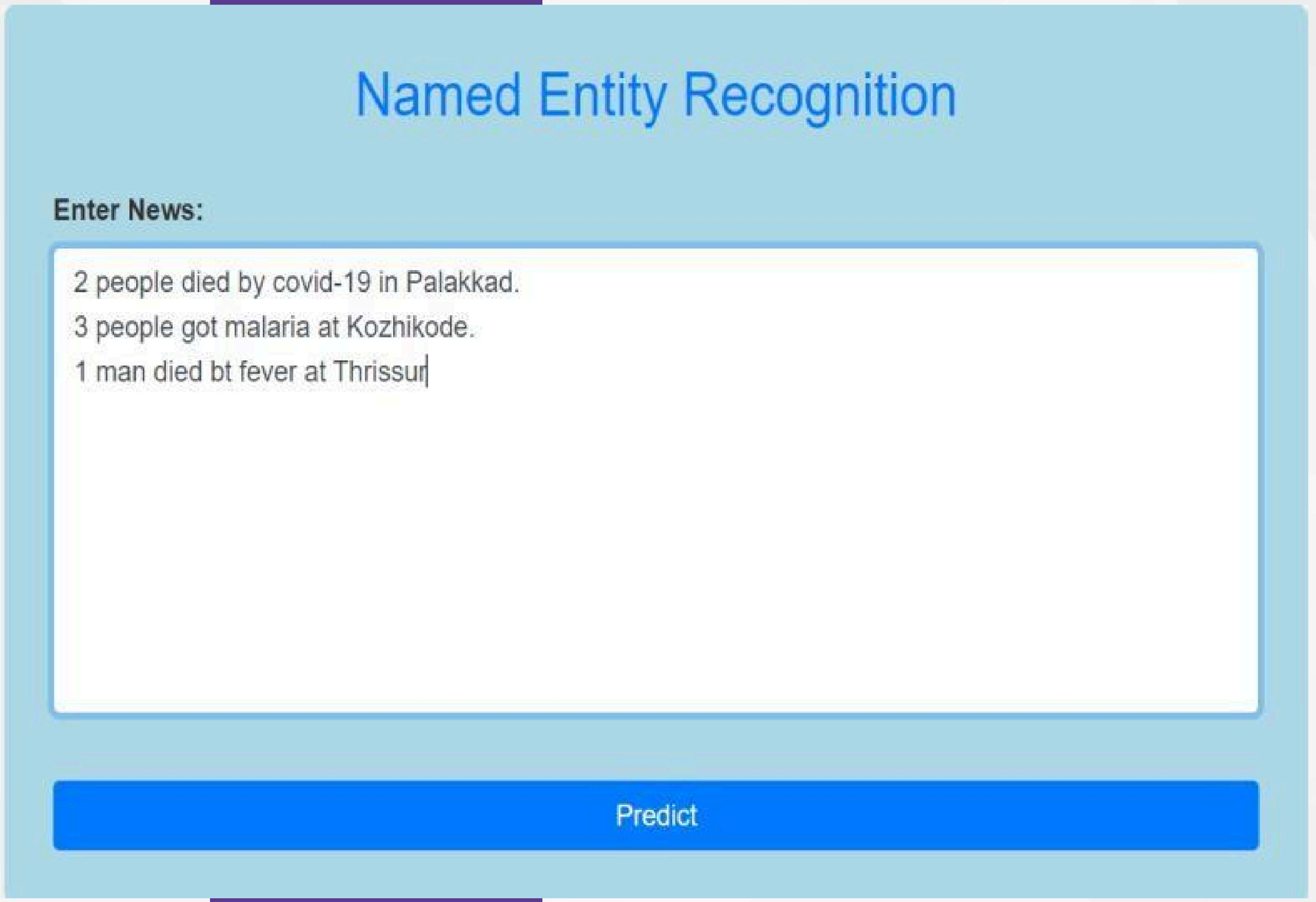


Fig :Accepting texts from user

Prediction Results

Sentence	Entities
2 people died by covid-19 at Palakkad.	2 - B-CARDINAL covid-19 - B-DISEASE Palakkad. - B-GPE
3 people got malaria at Kozhikode.	3 - B-CARDINAL malaria - B-DISEASE Kozhikode. - B-GPE
1 man died by fever at Thrissur.	1 - B-CARDINAL fever - B-DISEASE Thrissur. - B-GPE

View Final-Result

Back to Home

Fig :prediction result

RESULT

NO. OF CASES	DISEASE	LOCATION
2	covid-19	Palakkad.
3	malaria	Kozhikode.
1	fever	Thrissur.

Save as CSV | Print | Back to Home

Fig :final outcome

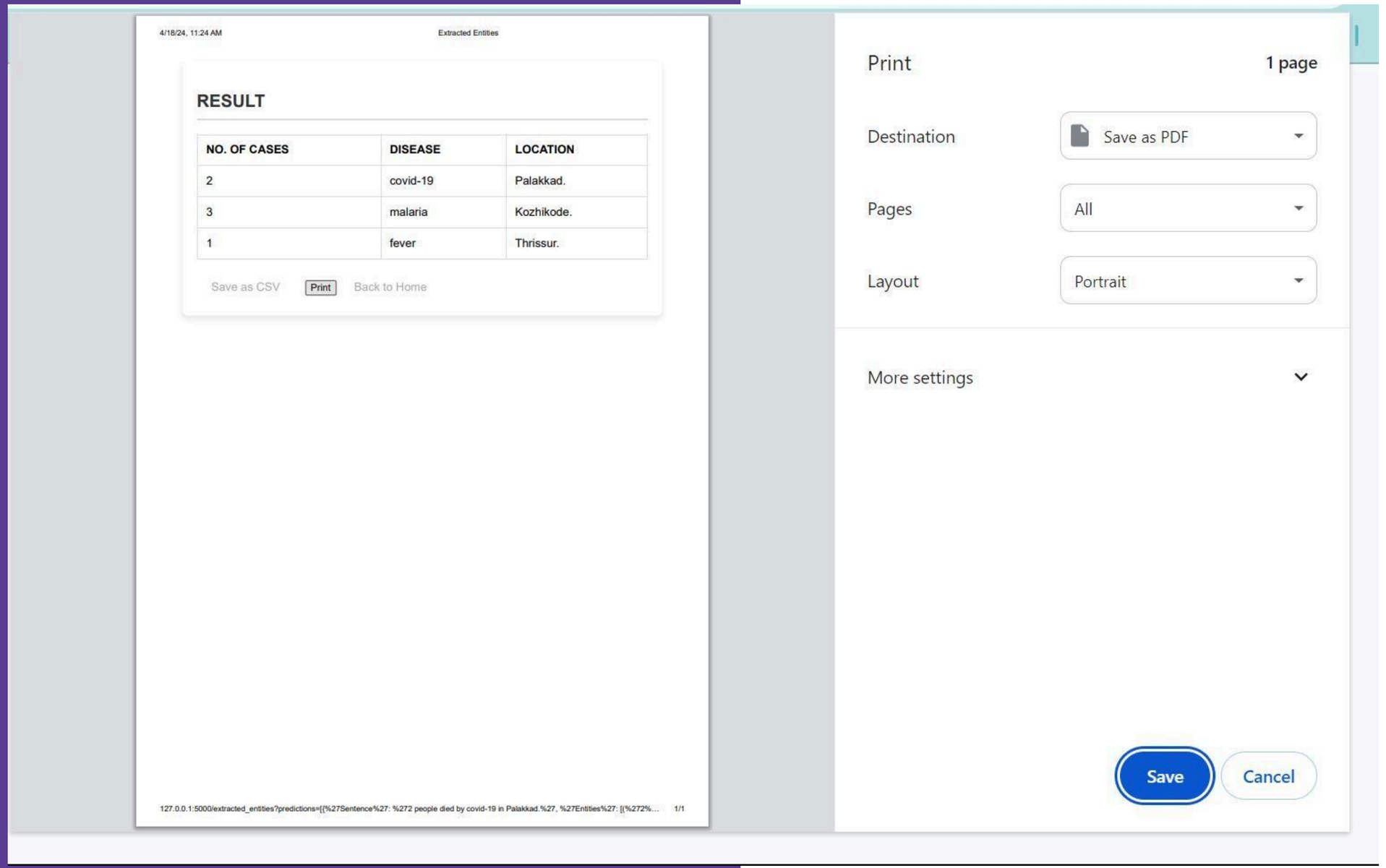


Fig :save option

CONCLUSION

- Focuses on "Named Entity Recognition for Disease Detection".
- System designed to utilize advanced (NLP) techniques.
- BERT is employed as the primary NLP model for identifying and extracting entities.
- Targets the extraction of entities related to diseases, cardinal numbers, and geopolitical entities (GPE) from text inputs provided by users.
- Enhance the efficiency of disease outbreak analysis.
- Facilitating quick decision-making processes such as resource allocation and other necessary actions.
- Facilitates informed resource allocation and response strategies during outbreaks.

FUTURE SCOPE

- Expand NER model beyond diseases, cardinal numbers, and geopolitical entities.
- Include entities like symptoms, treatments, or affected populations.
- Utilize geospatial analysis tools to visualize disease outbreak data on maps to identify hotspot areas and understand disease spread patterns.
- Develop capabilities for continuous analysis of incoming text data.
- Implement alert systems to notify stakeholders about emerging disease trends.

REFERENCES

- John Doe, Jane Smith, Sarah Johnson:"Enhanced Disease Surveillance Using Named Entity Recognition & Machine Learning",June 2019
- Michael Brown, Emily White, David Miller:"BERT-based Named Entity Recognition for Biomedical Text Mining",IEEE 5th International Conference on Electronics Technology (ICET).
IEEE,October 2020
- Samantha Lee, Robert Chen, Emily Wang:"Geolocation Extraction for Disease Surveillance Using Natural LanguageProcessing" , IEEE Internationa Conference on Computer Science, Electronic Information Engineering and Intelligent ControlTechnology (CEI). IEEE, March 2018
- Christopher Taylor, Jennifer Garcia, Thomas Clark:"Automated Disease Detection from Electronic Health Records Using Named Entity Recognition", September 2017.
- Andrew Wilson, Jessica Carter, Olivia Davis: "Combining Semantic Analysis with Named Entity Recognition for Disease Monitoring",April 2019

THANK YOU