Journal of Theoretical and Applied Information Technology
31st May 2024. Vol.102. No. 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING DISEASE OUTBREAK DETECTION: NAMED ENTITY RECOGNITION WITH FINE-TUNED DISTILBERT

MANJU JOY[1*] , DR. M KRISHNAVENI[2]

[1*]Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

[1*]Corresponding Author E-mail: 19phcsp010@avinuty.ac.in,

E-mail: [2]krishnaveni_cs@avinuty.ac.in

ABSTRACT

Within India's public health arena, Kerala emerges as a pioneer, often at the forefront of detecting and grappling with emerging infectious diseases. With a track record of vigilance and rapid response, Kerala has consistently been the first state in India to report the onset of various infectious outbreaks that have captured global attention. In 2018, Kerala witnessed the emergence of the first Nipah outbreak, followed by the first case of the COVID-19 pandemic in the country in 2020. Kerala's proactive surveillance mechanisms could promptly detect and notify authorities about these disasters. The first case of Monkeypox in India was also reported in Kerala in 2023, and Kerala's unwavering commitment to early detection remains unparalleled. In this scenario, a pioneering research endeavour is underway to harness the power of modern technology and computational algorithms in epidemic surveillance. Even though machines are excellent at extracting information from structured data, understanding human language and mining valuable information from unstructured text is challenging. This research presents a novel approach for detecting outbreak-related entities from amorphous text data utilizing a fine-tuned DistilBERT model with an accuracy of 96%. The model is optimized using the Optuna framework to ascertain optimal hyperparameters and ensure enhanced performance. This study advances NER methodologies in epidemiological surveillance, which is crucial for extracting relevant information from free-form text. We aim to automate the identification of disease names, affected locations, pathogen names, and population sizes from unstructured healthcare texts. Our experimental findings demonstrate that transfer learning techniques surpass baseline methods in NER tasks, mainly when training data are scarce. The proposed model is suitable for deployment in memory-constraint environments due to its capacity to operate with a reduced memory footprint and fewer resources.

Keywords: *Named Entity Recognition, LSTM, Transfer Learning, Fine tuning, DistilBERT*

1. INTRODUCTION

The majority of the world's knowledge is found in the form of text. NER or token classification is indispensable for extracting valuable information from unstructured text data[1],[2]. Analysis of text data from internet sources for outbreak detection is an active area of research. Event-based bio-surveillance systems are famous for their ability to detect outbreaks early. Biocaster, an event-based surveillance system, uses an SRL (Simple Rule Language) rulebook, which includes many rules, a list of words indicating the breadth of vocabulary, and a rich set of patterns for extracting semantic roles and other structured information from sentences. It relies on manually defined regular expressions to guide the analysis of English text data[3]. Bio surveillance systems focusing on a particular disease category might employ keyword-based methods or regular expressions to extract pertinent information. However, utilising an effective Named Entity Recognition (NER) model becomes crucial when the objective is to identify multiple diseases across diverse geographical areas.

Accurately detecting and classifying disease entities is crucial for effective surveillance of

infectious diseases. Traditional approaches, such as Conditional Random Fields (CRF) and Long Short-Term Memory (LSTM) networks, have shown promise but face challenges in terms of adaptability across different domains. To address these challenges, this research explores a novel approach to enhance the adaptability of NER models. Specifically, we investigate how transformer-based architectures, like DistilBERT, can contribute to advancing the state-of-the-art in disease entity detection. Our study aims to answer the following research questions:

1. What novel approach can enhance the adaptability of the Named Entity Recognition (NER) model across different domains?
2. How does the integration of transformer-based architectures, such as DistilBERT, advance the state-of-the-art in disease entity detection from unstructured text data?
3. How to compare existing techniques with transformer based models in terms of performance, scalability, and practical utility?

By addressing these questions, we aim to provide insights that can inform the development of more effective and adaptable NER models for disease detection and surveillance.

Kerala, a state in India, is notable for being the first to report cases of numerous infectious diseases in the country [4]. Therefore, the health sector in Kerala is considered for studying epidemic surveillance of communicable diseases using NLP techniques. NER models trained on general datasets might not excel in specialized domains. Adapting models to the vocabulary of a specific domain and distinctions is essential for efficient information retrieval. The unavailability of benchmark datasets and scarcity of outbreak resources are the major hurdles in the research. Named entities can appear in subject and object positions, making entity recognition difficult. In most NER works, nouns or proper nouns are classified into different classes, such as names of persons, locations, organizations, numbers, etc. In this work, nouns such as death, fatalities, recoveries, outbreak, etc., also should be identified as event entities. Such entities have ambiguous references, which makes it challenging to identify their boundaries and types correctly, and their overlapping nature poses challenges in distinguishing and labelling them correctly. Disease names and locations in Kerala should also be identified as entities. This work explores the performance of supervised Machine Learning, Deep Learning techniques and Large Language Models (LLMs) on entity recognition relevant to the health domain and their efficiency based on various metrics. NER models based on handcrafted features or limited contextual information may struggle to recognize out-of-vocabulary words in the sample data, hindering their ability to adapt to emerging trends or entities. So, a computationally efficient transformer-based DistilBERT model is proposed for token classification tasks, which can capture global contextual dependencies and relationships within the input data.

The remaining sections of the paper are organized as follows: The related literature is covered in the second session, and the third section elaborates on the proposed methodology and its operational procedure. Section four discusses the evaluation of the proposed NER model with respect to baseline methods and a comparison with previous works reported in the same domain. Section five of the paper gives an idea of the experimental setup. Section 6 provides a conclusion and future work recommendations.

2. LITERATURE SURVEY

Named entities are characterized as individual elements or words within the text that fit into predefined categories, encompassing designations like names of individuals, organizations, geographic locations, temporal expressions, quantities, monetary figures, percentages, and so forth. The concept of NER was introduced through the Sixth Message Understanding Conference (MUC-6) held in 1995 [5],[6]. Proper nouns or nouns present in the sentences are usually identified as entities, and the type and relevance of entities will vary from domain to domain. So, NER models tailored to specific domains often do not perform equally well in different domains[7]. The efficacy of Named Entity Recognition (NER) techniques varies across languages, as the NER system developed for English is not suitable for Hindi or Chinese. The majority of the NER studies in English, German, Arabic, Chinese, Bengali, Hindi, etc., are confined to identifying general entities like names of places, persons and organizations. An ample amount of work in NER has been done for the open domain, but minimal work has been done for domain-specific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

named entity extraction. Only the biomedical domain has been considered for domain-specific NER to extract the biomedical entities such as Genes, proteins, RNA, DNA, cell types, enzymes, etc. An agricultural NER is proposed in the literature to identify agriculture-related phrases such as crop name, soil name, and names of crop diseases[8]. Even if the health domain has similarities with bio medical domain, naming conventions used and specialized terminology make it a different and challenging task. Identifying entities within the health domain poses a greater complexity than general domains, involving tasks like discriminating names of persons, diseases and pathogens.

Different methods for NER model creation proposed in the literature are rule-based, machine learning or deep learning-based, hybrid, and transfer learning-based methods [2]. Earlier approaches for NER tasks focused on lexical and syntax analysis of sentences to identify their semantics. In the rule-based approach, handcrafted rules based on specific patterns of words, part-of-speech tags, or other linguistic features that specify the presence of named entities are used to detect and extract named entities from text. These rules are typically regular expressions or simple if-then rules to categorize entities. A dictionary containing all types of entities is employed in this method, and after the tokenization process, a search in the dictionary is carried out to identify entities present in the corpus [7]. This method was very popular in earlier research studies [29],[30],[31]. Most frameworks today are moving away from a rules-based approach to NLP in favour of machine learning or deep learning-based approaches[9].

In machine learning-based NER, developers use statistical methods to teach a computer system. The hidden Markov model, Conditional Random Field (CRF)- a probabilistic framework for sequence labelling, and the Maximum Entropy Markov Model are popular approaches. Certain NER studies show that a hybrid approach, by combining ML and rule-based systems, shows better accuracy than individual statistical methods[9]. Deep Learning models for processing sequential data are RNN, LSTM and GRU. They are helpful for sequence modelling tasks such as speech recognition, handwriting and gesture recognition, Time Series Analysis, etc. LSTMs can be used for several NLP tasks, including sentiment analysis, part-of-speech tagging and named entity recognition [11]. The existing pre-trained NLP library, spaCy, can detect 18 different entities but not disease-related entities. Even if spaCy can detect geo-political entities (GPE), it failed to detect GPEs in Kerala, as illustrated in the following example.

Input: “The Alappuzha district registered 2,168 dengue cases on Saturday. The Health Department has not released the test positivity rate for Covid-19. Of the fresh cases, 2,087 people contracted Covid-19 through local transmission."

In the spaCy output shown in the above example, ‘Alappuzha’ should be identified as a DISTRICT. spaCy does not even identify COVID-19 as an entity. So, NER models which can adapt to domain-specific vocabulary and nuances are essential for efficient information retrieval. An overview of the Named Entity Recognition (NER) works documented in the literature is given in Table 1.

Table 1: An Overview Of NER Works

Reference	Domain /Language	Named Entities	Algorithm(s) used	Dataset(s) used
Khan W. et al. (2022) [7]	Urdu Language	The name of the Person, Organization, Location, Date, Designation, and Number are the entities.	Conditional Random Field algorithm is used.	IJCNLP-Urdu (benchmark dataset) and articles collected from the BBC Urdu website named **UNER-1**

				(self-created dataset) are used.
Veena Gangadharan and Deepa Gupta (2020)[8]	Agriculture Domain	Crop names, Types of Soil, Pathogen names, Crop Diseases and Fertilizers are the entities.	Agriculture vocabulary-AGROVOC- is used to identify crop names. Latent Dirichlet Allocation based topic modelling algorithm is used.	3000 sentences related to the agriculture domain of Kerala State collected from reputed agriculture websites are used.
Shaker A et.al. (2021)[11]	Arabic language	Person, Location, geopolitical, time, profession, organization, disease, geography, and miscellaneous.	2 NER models based on LSTM and GRU are proposed. LSTM gave better results.	The data set is not mentioned clearly.
Azizi S et.al. (2022)[12]	Neurological signs and symptoms	Symptoms of neurological disorders are tagged based on the length of the word.	A NER model is proposed to recognize neurologic symptoms. CNN and BERT are used to develop NER. The performance of the BERT model was better than CNN for the three corpora considered.	Neurological case antiquities from five textbooks, Clinical Synopses of Nerve Diseases from OMIM Corpus, physician notes from the electronic health record, etc., are used for training and testing.
Islamaj R et. al. (2021)[13]	Bio-Medical Domain	Chemical and Drug names	BlueBERT with multi-terminology candidate resolution normalization architecture (MTCR) is proposed.	NLM-Chem Corpus with 150 full-text articles is used.
Macarious Abadeer (2020)[14]	Medical Text	Name of Doctor/Patient, Hospital, Date, Organization, Medical record etc.	A fine-tuned DistilBERT cased model is used.	I2b2b2010 and i2b2 2014 are the two datasets used for the study.
Lee J et.al. (2020)[15]	Bio-Medical Domain	Drug/Chemical, Gene/ Protein, Diseases, Species	Fine tuning of BioBERT is done.	PubMed abstracts, the NCBI disease dataset and PMC articles are used
M. Al-Smadi et al. (2020)[16]	Arabic language	8 coarse-grained tags such as Person, Organization, Geopolitical Entity, Location, Facility, Vehicle, Weapon, and Product span over 50 fine-grained	MUSE (Multilingual Universal Sentence Encoder) along with Pooled-GRU based on transfer learning with deep neural network model approach is implemented.	Arabic NER dataset ($WikiFANE_{Gold}$) is used

		classes are classified.		
Pengfei Cao et al. (2018)[17]	Social media and news domain	Person, location, organization, Geo political Entities	BiLSTM+ CRF+ adversarial+ self-attention	WeiboNER and SighanNER Chinese NER datasets and MSR dataset
Nayan Banik and Hasan Hafizur Rahman (2018)[18]	Bangla language NER task	Person, Location, Organization and Date	Gated Recurrent Unit (GRU) is used	Manually annotated Bangla online newspaper dataset
Cetoli A et al. (2018)[19]	General Domain	Location, Miscellaneous, Organization and Person	A neural network architecture with a Bi-directional LSTM model with a Graph Convolution Network, and CRF as the last layer is implemented. Glove embeddings are also utilized.	Experiments are conducted on OntoNotes 5.0 dataset
Chiu J and Nichols E (2016)[19]	General Domain	Location, Miscellaneous, Organization and Person	A hybrid Bidirectional LSTM and CNN model is proposed	CoNLL-2003 dataset and OntoNotes 5.0 dataset are used
Shuwei Wang et. al.(2014) [21]	Financial domain	Name of Financial Organization, Location, Title/Designation, Objects	Conditional random field algorithm is combined with information entropy, mutual information, and word similarity measurement to recognize the abbreviated FNEs.	Chinese financial text dataset with 5500 sentences and MSRA dataset
Doren Singh T (2009)[22]	Manipuri Language	Name of person, Location, Organization and Miscellaneous	SVM is used for token classification.	Data was collected from 'The Sangai Express' , a popular Manipuri news paper is used as dataset.1235 sentences were used for training and 189 for testing.
Chantana Chantrapornchai and Aphisit Tunsakul (2021)[23]	Tourism Domain	Name of Hotel, Location and facility	Pretrained models such as Spacy and BERT	The dataset was generated by crawling websites related to tourism in Thailand.

Transfer learning techniques have a significant impact in the field of Natural NLP by allowing pre-trained models to be fine-tuned for specific tasks, saving computational resources and improving performance. BERT, a transformer-based model, has set new benchmarks for numerous NLP tasks such as text classification, token classification and question answering.

As per the literature, two publicly available data sets for Disease NER are NCBI Disease corpus and BC5CDR Disease corpus [9]. Diseases mentioned in the NCBI Disease Corpus include Cancer, Cardiovascular diseases, and Infectious diseases such as AIDS, tuberculosis, Malaria, etc. Neurological, genetic, endocrine and Mental health disorders are also included, along with Autoimmune diseases and lung diseases such as asthma and chronic obstructive pulmonary disease. Entities mentioned in the BioCreative V CDR corpus contain annotations for associating drugs or chemical entities with disease entities within the text. Both these datasets are unsuitable for studying outbreak detection from a text corpus related to Kerala's health domain. The unavailability of a high-quality benchmark dataset and sufficient resources is a significant hurdle in this research. The most challenging phase of this work is collecting a suitable corpus and cleaning and annotating it for training purposes. The study aims to identify the relationship between disease entities, regions affected by outbreak events and frequency of occurrences.

3. METHODOLOGY

Traditional NER algorithms are feature-based and cannot understand a word based on its context. Transformer-based architecture, such as BERT and its alternatives, can capture contextual information effectively and has demonstrated high efficiency and effectiveness in NLP and Computer Vision tasks. Contextual information is crucial for accurately identifying entities within text. Fine-tuning BERT models can be resource-intensive due to their large size, but variants like DistilBERT offer a smaller yet highly capable alternative. DistilBERT achieves size reduction through knowledge distillation, retaining most of BERT's language understanding abilities while being faster due to architectural simplifications[24]. This work focuses on disease-related entity detection using transfer learning techniques by fine-tuning the DistilBERT model. The steps for developing the NER model are given below:

1. Data gathering and Preprocessing
2. Tokenization and Alignment of labels
3. Loading of pre-trained DistilBERT model and addition of NER Head.
4. Splitting the dataset into training and test sets.
5. Hyperparameter Optimization.
6. Model Training using the Training dataset.
7. Evaluation of performance using Test dataset.

3.1 Data Gathering and Preprocessing:

In this step, the dataset for the experiments is prepared by collecting unstructured text data from various Online news websites covering Kerala News, Open access resources such as Medline Plus(https://medlineplus.gov/), Wikipedia, Journal articles available in PubMed, NIH Magazine (https: //magazine. medlineplus.gov/), etc. An annotated corpus with all relevant entities appropriately marked is indispensable for training the NER model. News articles related to disease outbreaks can be extracted from HTML and XML documents using web scraping. Preprocessing is carried out to remove noise from text data collected from digital news websites. MedlinePlus provides curated health information by gathering information from NLM(the National Library of Medicine), NIH(the National Institutes of Health), other U.S. government agencies, and health-related organisations. It provides reliable health information in English and Spanish that is layperson-accessible to nonprofessional users. Articles published by WHO and NIH MedlinePlus Magazine are also utilized. Then, the collected data is manually annotated using Label Studio and UBIAI tools to identify named entities such as names of diseases, pathogens, places affected by the outbreak, date, events such as death, cases reported, suspected/diagnosed cases, etc. After data collection, extensive pre-processing removes special characters, punctuations and irrelevant information. After tokenizing the sentences, part-of-speech tagging is done using the NLTK library to get syntactic and semantic information about text. NER tagging is done manually. The collected corpus is then converted into CoNLL-2003 format, a standard data format for representing NER tasks in natural language processing. Entities are labelled using IOB tagging.

3.2 Tokenization and Alignment of Labels:

Input is fed in the form of sentences and labels associated with each word in the sentence. The input text is tokenized using the tokenizer. Tokenization is breaking down a sentence into words, sub-words, or characters. It facilitates the conversion of raw text data into a

format conducive to efficient machine analysis. WordPiece tokenization - the algorithm used by DistilBERT allows the model to handle languages with complex word formations and morphological variations. The WordPiece tokenizer starts with subword tokens, such as individual characters or common word parts, then iteratively adds the most frequent out-of-vocabulary (OOV) subword units. The vocabulary of the proposed model is built from these tokens. Thus, the words not explicitly seen during training can be represented efficiently, enabling DistilBERT to work effectively with diverse text data. To achieve custom tokenization patterns, regular expressions are used to process input data before passing it through the tokenizer so that the tokenizer will treat them as a single unit. Two unique tokens, [CLS], classification token, and [SEP] token, enable DistilBERT to know the start and end of the sequence of tokens given as input to the pre-trained DistilBERT model. The maximum size of tokens fed into the proposed DistilBERT model is 137. Sequences with less than 137 tokens are padded with [PAD] tokens to ensure uniform token size fed as input. The model then outputs a dense vector of size 768 for individual tokens. These dense vector representations of words in a continuous vector space can capture contextual information and semantic similarities between words in the embedding space. Identifying entities in the health domain is more challenging than solving tasks in other domains [10]. In disease names such as Covid-19, SARS-CoV-2, JN.1, Hepatitis B, Hepatitis A, Hepatitis C, Hepatitis D and pathogen names like BA.2.10, BA.2.86, BA.2.75, etc., the appearance of special symbols like '-' or '.' or space are very common, which do not occur in person names or location names. This resulted in new sub-words and discrepancies in the sequence length and corresponding NER labels. Proper alignment of labels is carried out by assigning the same label among all the sub-words that belong to the same token. 50 unique tokens are appended to the DistilBERT tokenizer by a process called special tokenization to avoid splitting disease and pathogen names with special symbols. Proper tokenization and handling of out-of-vocabulary words influence the performance results of a token classification model. The proposed uncased DistilBERT model has a vocabulary of size 30576 tokens.

3.3. Load the pre-trained DistilBERT model and add the NER Head.

Large Language Models (LLMs) have demonstrated cutting-edge performance across numerous natural language processing (NLP) applications[24]. Fine-tuning and in-context learning are two strategies to tailor a pre-trained LLM to specific tasks within a given domain. The fine-tuning method enhances the performance of a pre-trained Large Language Model (LLM) using a custom dataset tailored to a specific downstream task, while in-context learning supplies task-specific context or examples during inference to direct the model's response generation[25]. Our proposed Named Entity Recognition (NER) model for disease outbreak detection utilizes DistilBERT- an LLM model designed for next-sentence prediction- fine-tuned on our dataset for outbreak detection purposes. The attention mechanism helps transformer-based models to focus on various parts of the input sequence while processing each token [26]. DistilBERT is based on transformer architecture that implements actual bi-directional attention. Understanding the relationship between words and their context is essential for accurate entity recognition. Transformers can deal better with long-term dependencies. The transformers package developed by Hugging Face Company is used to conduct experiments in this work.

NER is a token classification task where a token or word present in a sentence is assigned an entity type e ε **E** in a given sentence S={w1, w2, w3……,wn} where E denotes the set of entity labels and n denotes the number of tokens in a particular sentence. We initialized the DistilBERT model with these pre-trained weights, and on top of it, a custom NER classification layer is added for token classification. DistilBERT also incorporates positional embeddings to encode the position or order of tokens within the input sequence so that the sequential structure of the input text can be captured. DistilBERT's word embeddings provide contextualized word representations that capture the semantics and relationships of words within the context of the input text. Positional embeddings give the model information about the positions of words in a sentence. Input embeddings pass through the attention layer, residual connections and layer normalization modules. The model learns by combining the token and positional embeddings to comprehend a document's context and relationships

between tokens. Attention mechanisms capture long-term dependencies. Most of the parameters of the model are contained in the feed-forward layer. Multiple self-attention and feedforward neural network layers are categorized into the Transformer Encoder Layers and the Output Layer. DistilBERT model architecture contains 6 layers of Transformer Encoder Layers, which are the core components of DistilBERT. The last layer of the DistilBERT model is typically adapted for a specific downstream task. A linear output layer maps the contextual representations from the Transformer Encoder Layers to the desired output classes or entity labels. A SoftMax activation is applied to generate probabilities for each possible output class. Using the training data, fine-tuning is conducted to predict named entities accurately and efficiently. During the training loop, we iterate over the data loader, move the inputs and labels to the appropriate device (CPU or GPU), pass the inputs to the model, compute the loss, backpropagate the gradients, and update the model's parameters using the Adam optimizer. Finally, we saved the fine-tuned model and tokenizer for future use.

3.4. Training and Test sets

The dataset is available in CoNLL-2003 format, where each line represents a word in a sentence along with its associated linguistic annotations, such as POS tag and entity label. Entity label denotes the type of entity. It is a standard data format for representing named entity recognition (NER) tasks in natural language processing. The number of words belonging to each of the 12 Entity classes with IOB entities in combined format is shown in Table 2, given below.

Table 2: Details of Entity Classes

Entity Type	Count
Disease	6803
District	1826
GPE	2910
Event	3404
State	1055
Date	1404
Cardinal	3058
Person	1015
Pathogen	842
Percent	328
Nation	362
O	116094

Entities are labelled using IOB tagging where B refers to the Beginning of an Entity, I mean inside of an entity, and O is Other or non-entity.

3.5 Hyper Parameter Optimization

The efficacy and performance of ML and DL models are significantly influenced by parameters and hyperparameters. Parameters are dynamic variables optimized throughout training to minimize the disparity between predicted and actual output. On the other hand, hyperparameters dictate the model's generalization capacity, are set before the training process and remain constant throughout the training process [27][28]. Optuna, an automated Hyper Parameter Optimization (HPO) framework, is employed to streamline hyperparameter tuning, particularly for optimizing essential hyperparameters in the DistilBERT model. These include learning rate, epochs, batch size, optimizer, warm-up steps, sequence length, fine-tuning strategy, etc. A systematic exploration of various hyperparameters is conducted through 50 trials to enhance the precision and accuracy of the proposed model. The table 3 below outlines the specific values tested during the experiment:

Table 3: Hyper Parameter Tuning Details

Hyper Parameters	Type	List/ Range of Values	Optimum Value selected
Batch Size	Integer	[5, 8, 10, 16, 20,50,75]	8
Optimizer	Categorical	[AdamW, Adam,SGD, Ada]	AdamW
Warmup steps	Integer	[0,10]	4
Sequence length	Integer	[122, 512]	137
Learning rate	Real	[1e-5, 5e-5]	3.45e-05
No.of Epochs	Integer	[1,10]	8

Within gradient descent optimization, the learning rate governs step size, while epochs dictate the number of times the entire dataset is traversed. Batch size determines the number of samples processed per iteration, with larger sizes potentially leading to faster convergence, albeit with increased

memory demands. Warm-up steps stabilize training and mitigate abrupt fluctuations in model weights, whereas the optimizer orchestrates weight updates. With a fixed maximum sequence length of 137, adequate for capturing crucial information from training samples, only the top layers are fine-tuned, reducing the requisite training data.

4. RESULTS AND DISCUSSION

The NER dataset is unbalanced, with a significantly higher number of 'O' (Outside entity or non-entity) labels compared to actual entity labels, making dataset balancing challenging. Therefore, the most appropriate metrics for evaluating NER models are recall, precision, and F1 score[28]. CRF, LSTM with FastText Embeddings (LSTM FE), BERT and DistilBERT are the algorithms implemented in the experiments. BERT and DistilBERT gave relatively better results even if the time and resources required for training the models were high. CRF has 14199 trainable parameters, whereas LSTM with fast text embeddings has 561521 parameters for training. DistilBERT and BERT have 65248533 and 107773461 trainable parameters, respectively. The size of the trainable parameters is the reason for the increased training time of the BERT and DistilBERT models. Training time for CRF, LSTM FE, BERT and DistilBERT algorithms are 29s, 45s, 356s and 188s, respectively. Precision, recall and accuracy metrics of transfer learning-based methods are high because of their ability to understand contextual information.

The performance of baseline methods such as CRF, LST and Finetuned BERT and proposed DistilBERT models are compared in terms of Entity level Precision(P), Recall (R)and F1-score (F1) and high scores obtained for various entities are highlighted and shown in Table 4 given below:

Table 4: Comparison of Various NER Algorithms

Algorithms	CRF			LSTM FE			BERT			DistilBERT		
Entities	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
B-CARDINAL	0.85	0.79	0.82	0.62	0.52	0.57	0.93	**0.93**	0.93	**0.93**	0.92	**0.93**
B-DATE	**0.92**	0.50	0.65	0.45	0.26	0.33	0.91	**0.64**	**0.75**	**0.91**	**0.62**	**0.73**
B-DISEASE	0.93	0.90	0.91	0.77	0.63	0.69	0.93	0.97	0.95	**0.94**	**0.97**	**0.95**
B-DISTRICT	0.99	0.99	0.99	0.45	0.19	0.26	0.99	1.00	0.99	**0.99**	**1.00**	**0.99**
B-EVENT	0.84	0.81	0.82	**0.91**	0.76	0.83	0.81	0.81	0.81	0.83	0.81	0.82
B-GPE	0.81	0.86	0.83	0.39	0.46	0.42	0.98	0.93	0.93	**0.98**	**0.94**	**0.96**
B-NATION	**0.99**	**0.94**	**0.96**	0	0	0	0.85	0.83	0.84	0.87	0.83	0.85
B-PATHOGEN	0.45	0.40	0.42	0.27	0.28	0.27	0.51	0.45	0.48	**0.56**	**0.50**	**0.53**
B-PERCENT	0.91	0.86	0.89	0	0	0	0.95	1.00	0.97	**0.97**	**1.00**	**0.98**
B-PERSON	0.81	0.56	0.66	**1.00**	0.01	0.02	0.91	**0.90**	**0.90**	0.90	**0.87**	0.89
B-STATE	**0.99**	0.89	0.94	0.23	0.19	0.21	0.94	**0.98**	0.96	0.97	0.97	**0.97**
I-CARDINAL	0.44	0.17	0.24	0	0	0	0.78	**0.22**	**0.35**	**0.87**	0.21	0.33
I-DATE	0.91	0.62	0.74	0.87	0.07	0.13	0.95	**0.78**	**0.86**	**0.96**	0.77	0.85
I-DISEASE	0.79	0.61	0.69	0.67	0.40	0.50	**0.78**	0.64	0.70	0.77	**0.66**	**0.71**
I-EVENT	0.55	**0.53**	0.54	**0.76**	0.51	**0.61**	0.60	0.42	0.49	0.65	0.51	0.57
I-GPE	0.50	0.67	0.57	0	0	0	**1.00**	0.82	0.90	0.89	**0.94**	**0.91**
I-PATHOGEN	**0.68**	0.36	0.47	0	0	0	0.59	0.40	0.47	**0.62**	**0.47**	**0.53**
I-PERCENT	**0.97**	0.82	0.89	0.98	1.00	**0.99**	0.93	1.00	0.97	0.93	**1.0**	**0.97**
I-PERSON	0.70	0.75	0.7	0	0	0	**0.79**	0.91	**0.84**	**0.75**	**0.91**	0.82
O	0.97	0.99	0.98	0.93	.94	0.93	0.98	0.99	0.98	**0.98**	**0.99**	**0.98**

Upon careful analysis, we can understand that Entity-wise metrics are high for pre-trained models. LSTM with FastText embedding gave poor results for some entities because it failed to handle out-of-vocabulary words. CRF gave better results, even though it lacks contextual information. To illustrate the impact of the proposed NER model on real-world Applications, two sample texts and the output obtained from the model are shown below:

Sample Text 1: The Alappuzha district registered 2168 dengue cases on Saturday. The Health Department has not released the test positivity rate for Covid-19. Of the fresh cases, 2087 people contracted Covid-19 through local transmission.

```
Entities: ['alappuzha', '2168', 'dengue', 'cases', 'COVID-19', 'cases', '2087', 'COVID-19']
tags : ['DISTRICT', 'CARDINAL', 'DISEASE', 'EVENT', 'DISEASE', 'EVENT', 'CARDINAL', 'DISEASE']
```

Sample Text 2: 3 cases of Covid-19 reported at Ernakulam. Measles and Dengue reported at Kannur. Monkeypox at Thrissur. Angamaly and Ernakulam on Nipah Alert.

```
Entities: ['3', 'cases', 'COVID-19', 'ernakulam', 'measles', 'dengue', 'kannur', 'monkeypox', 'thrissur', 'angamaly', 'ernakulam', 'nipah']
tags : ['CARDINAL', 'EVENT', 'DISEASE', 'DISTRICT', 'DISEASE', 'DISEASE', 'DISTRICT', 'DISEASE', 'DISTRICT', 'GPE', 'DISTRICT', 'DISEASE']
```

Table 5 summarises the performance of our proposed NER model compared with selected existing works for disease outbreak detection in terms of precision, recall, and F1-score.

Table 5 : Comparison with Previous Works

Reference	Algorithm/Methodology used	Performance
Antonella Dellanzo et al.(2020)[29]	A rule-based algorithm is used along with Freeling to perform NER. Regular Expressions and Gazetteer are utilized.	Precision: 48% Recall: 46% F1 score: 47%
Minh-Tien Nguyen and Tri-Thanh Nguyen(2013)[30]	Rules and Dictionaries are used	Precision: 89.47% Recall: 94.44% F1 score: 91.89%
Lejeune G, et al. (2015)[31]	The proposed Daniel System can process multiple languages for epidemic detection using a lexicon with disease names and locations.	Precision: 82% Recall: 82% F1 score: 82%
Proposed model	Optimized and Fine-tuned DistilBERT model	**Precision: 96%** **Recall: 96%** **F1 score: 96%**

In view of recent advancements in NLP, particularly in the domain of healthcare and epidemiology, our research stands as a progressive step forward by harnessing the potential of transformer-based models, specifically DistilBERT, for disease entity recognition. Our research aligns with the current trend of utilizing cutting-edge techniques and extends the boundaries of NLP applications in the healthcare sector. While some prior research has concentrated on theoretical aspects or limited experimental evaluations, our study prioritizes real-world application and practical implications. By showcasing the efficacy of DistilBERT in identifying disease-related entities, our work facilitates a crucial advancement in bio-surveillance capabilities. Timely and accurate detection of disease outbreaks from textual data is imperative for public health agencies and epidemiologists. We provide a comprehensive analysis of the performance of the DistilBERT model compared to existing methods, highlighting its superiority in terms of accuracy and efficiency. This differentiation elucidates the unique contribution of our work to the existing body of literature on disease outbreak detection.

5. EXPERIMENTAL SETUP

Experiments were conducted in Colab Pro using a V100 GPU processor utilizing 54.8GB of High RAM. We have trained the model with a learning rate of 3.45e-05, batch sizes 8. AdamW is the optimization algorithm used, and the number of epochs is 8. Increasing epochs beyond 10 did not give improved results. 80 % of the samples are used for training and 20% for testing.

6.CONCLUSION

The key objective of the study was to enhance disease outbreak detection through the accurate recognition of disease-related entities from textual data. We have experimented with different NER algorithms – both transfer learning- and non-transfer learning approaches – and evaluated the performance of the algorithms in terms of precision, recall, and F1 score of most relevant entities in the health domain and processing time of these algorithms. Through the implementation of a fine-tuned DistilBERT model for Named Entity Recognition (NER), we could bring significant improvements in entity detection accuracy, as demonstrated by the provided examples. By comparing our proposed model's performance with other baseline NER models, we aim to provide a comprehensive assessment of the capabilities of transfer learning techniques for token classification tasks. The transfer learning technique is highly effective even with a relatively low dataset size. By exploiting the transfer learning technique, our fine-tuned DistilBERT model could achieve 96% accuracy in the NER task with higher precision, recall and F1 score values for individual entities, effectively fulfilling our research's core objective. However, it is important to acknowledge the limitations and potential threats to validity inherent in this study. An intrinsic limitation is the dependence on a tailored dataset for refining the DistilBERT model. Although this enabled the model to adapt specifically to domain-specific entities, the generalizability of the model to unseen data or diverse contexts may be impacted. Additionally, the performance of the NER model could be influenced by factors such as data quality, field specificity, and language variations, which may introduce biases or inaccuracies in entity recognition. Addressing the challenges posed by data imbalance and bias in annotated datasets used for training NER models remains an open issue, and solving these issues could enhance the model's performance. Incorporating temporal dynamics into NER models to capture the evolving nature of disease outbreaks is an area requiring further exploration. Exploring the integration of multimodal data sources, such as textual, spatial, and temporal data, could provide a richer context for disease surveillance. Examining techniques for seamlessly integrating multiple modalities into NER models will be investigated further.

ACKNOWLEDGEMENT

During the preparation of this work, the authors used Grammarly to paraphrase and check grammar. After using the tool/service, the author reviewed and edited the content as needed and took full responsibility for the publication's content.

REFERENCES:

[1] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition", Dec. 2018, doi: 10.1109/TKDE.2020.2981314.

[2] N. Collier *et al.*, "BioCaster: Detecting public health rumours with a Web-based text mining system", *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, Dec. 2008, doi: 10.1093/bioinformatics/btn534.

[3] S. Sreedharan, "Analysing the Covid-19 Cases in Kerala: a Visual Exploratory Data Analysis Approach", doi: 10.1007/s42399-020-00451-5/Published.

[4] J. P. Jayan, R. R. R, and E. Sherly, "A Hybrid Statistical Approach for Named Entity Recognition for Malayalam Language",2013.

[5] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named Entity Recognition and Relation Detection for Biomedical Information Extraction", *Frontiers in Cell and Developmental Biology*, vol. 8. Frontiers Media S.A., Aug. 28, 2020. doi: 10.3389/fcell.2020.00673.

[6] W. Khan, A. Daud, K. Shahzad, T. Amjad, A. Banjar, and H. Fasihuddin, "Urdu Named Entity Recognition Using Conditional Random Fields", *Applied Sciences (Switzerland)*, vol. 12, no. 13, Jul. 2022, doi: 10.3390/app12136391.

[7] V. Gangadharan and D. Gupta, "Recognizing Named Entities in Agriculture Documents using LDA based Topic Modelling Techniques", in *Procedia Computer Science*, Elsevier B.V., 2020, pp.

1337–1345. doi: 10.1016/j.procs.2020.04.143.

[8] L. Kühnel and J. Fluck, "We are not ready yet: limitations of state-of-the-art disease named entity recognizers", *J Biomed Semantics*, vol. 13, no. 1, Dec. 2022, doi: 10.1186/s13326-022-00280-6.

[9] P. Hiremath and S. B. R, "Approaches to Named Entity Recognition in Indian Languages: A Study", 2014.

[10] A. Shaker, A. Aldarf, and I. A. Bessmertny, "Using LSTM and GRU With a New Dataset for Named Entity Recognition in the Arabic Language." [Online], Available: https://github.com/Alaa-Shaker/

[11] S. Azizi, D. B. Hier, and D. C. Wunsch, "Enhanced neurologic concept recognition using a named entity recognition model based on transformers", *Front Digit Health*, vol. 4, Dec. 2022, doi: 10.3389/fdgth.2022.1065581.

[12] R. Islamaj *et al.*, "NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature", *Sci Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1038/s41597-021-00875-1.

[13] M. Abadeer, "Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and Medical Concepts", 2020. [Online]. Available: https://github.com/huggingface/

[14] J. Y. Lee, F. Dernoncourt, and P. Szolovits, "Transfer Learning for Named-Entity Recognition with Neural Networks", [Online]. Available: https://github.com/Franck-Dernoncourt/NeuroNER

[15] M. Al-Smadi, S. Al-Zboon, Y. Jararweh, and P. Juola, "Transfer Learning for Arabic Named Entity Recognition with Deep Neural Networks", *IEEE Access*, vol. 8, pp. 37736–37745, 2020, doi: 10.1109/ACCESS.2020.2973319.

[16] P. Cao, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism", [Online]. Available: https://github.com/CPF-NLPR/AT4ChineseNER

[17] N. Banik, M. Hasan, and H. Rahman, "GRU based Named Entity Recognition System for Bangla Online Newspapers", 2018. [Online]. Available: https://www.prothomalo.com

[18] A. Cetoli, S. Bragaglia, A. D. O'harney, and M. Sloan, "Graph Convolutional Networks for Named Entity Recognition", [Online]. Available: https://github.com/contextscout/gcn_ner.

[19] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs",[Online], Available: http://nlp.stanford.edu/projects/glove/

[20] S. Wang, R. Xu, B. Liu, L. Gui, and Y. Zhou, "Financial named entity recognition based on conditional random fields and information entropy", in *Proceedings - International Conference on Machine Learning and Cybernetics*, IEEE Computer Society, Jan. 2014, pp. 838–843. doi: 10.1109/ICMLC.2014.7009718.

[21] T. Doren Singh, K. Nongmeikapam, A. Ekbal, and S. Bandyopadhyay, "Named Entity Recognition for Manipuri Using Support Vector Machine 1", 2009. [Online]. Available: http://www.thesangaiexpress.com/

[22] C. Chantrapornchai and A. Tunsakul, "Information extraction on tourism domain using SpaCy and BERT", *ECTI Transactions on Computer and Information Technology*, vol. 15, no. 1, pp. 108–122, Apr. 2021, doi: 10.37936/ecti-cit.2021151.228621.

[23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[24] Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., & Wang, G., "GPT-NER: Named Entity Recognition via Large Language Models", *ArXiv. /abs/2304.10428*, 2023.

[25] Shekhar, S., Bansode, A., & Salim, A.," A Comparative Study of Hyper-Parameter Optimization Tools, *ArXiv*. /abs/2201.06433, 2022.

[26] U. Zaratiana, P. Holat, N. Tomeh, and T. Charnois, "Hierarchical Transformer Model for Scientific Named Entity Recognition", March 2022, [Online]. Available: http://arxiv.org/abs/2203.14710

[27] Lee, Myungho & Jafar, Abbas, " Comparative Performance Evaluation of State-of-the-Art Hyperparameter Optimization Frameworks" , *Transactions of the Korean Institute of Electrical*

Engineers, 2023, vol.72,pp. 607-619. 10.5370/KIEE.2023.72.5.607.

[28] M. Y. Landolsi, L. Hlaoua, and L. Ben Romdhane, "Information extraction from electronic medical documents: state of the art and future research directions", *Knowledge and Information Systems*, vol. 65, no. 2. Springer Science and Business Media Deutschland GmbH, pp. 463–516, Feb. 01, 2023. doi: 10.1007/s10115-022-01779-1.

[29] Antonella Dellanzo, Viviana Cotik, and Jose Ochoa-Luna, "A Corpus for Outbreak Detection of Diseases Prevalent in Latin America", in *Proceedings of the 24th Conference on Computational Natural Language Learning*, 2020, pp. 543–551.

[30] M. T. Nguyen and T. T. Nguyen, "Extraction of disease events for a real-time monitoring system," in *ACM International Conference Proceeding Series*, 2013, pp. 139–147. doi: 10.1145/2542050.2542084.

[31] S. Sahnoun and G. Lejeune, "Multilingual Epidemic Event Extraction: From simple Classification methods to Open Information Extraction (OIE) and Ontology," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2021, pp. 1227–1233. doi: 10.26615/978-954-452-072-4_138.

[1]