www.ijirses.com *ISSN: 2583-1658 (Online) | Volume: 3 Issue: 7 | 2023*

Semantic-Based Natural Language Processing for Classification of Infectious Diseases Based on Ecological Factors

Saviour Inyang[1], Imeh Umoren[2]

[1&2] Department of Computer Science, Akwa Ibom State University, MkpatEnin, Nigeria.

Abstract: *As the world undergoes constant changes, it has become evident that public health measures must adapt to address the evolving needs of different contexts. The transition from an old universal health system to a contemporary public health model has given rise to an emphasis on environmental public health. This approach aims to understand and counter the diverse influences on our health stemming from ecological determinants. Furthermore, in the realm of ecological classification of infectious diseases, Semantic Natural Language Processing plays a crucial role. By analyzing ecological sample text, this technology enables the identification and categorization of diseases based on their ecological characteristics. Firstly, a comprehensive review of ecological factors related to infectious disease transmission was conducted. Then, semantic texts describing these factors were identified and extracted from medical databases, resulting in a final selection of 342 from over 500 journals. The extracted texts formed a corpus, which underwent preprocessing to remove stop words, punctuation, and whitespaces. Next, a document term matrix was created to represent the texts, XGBoots was trained and tested on the document term matrix. The Cross Validation accuracy on the training data was 70% for XGBoots. Additionally, a deep learning model called BERT was integrated with the XGBootsto create an interactive interface for users to input ecological factors sample text and receive infectious disease classifications with confidence intervals.*

Keywords: *Semantic, Ecology, Natural Language Processing, Machine Learning*

1. Introduction

In recent years, it has become more and more evident that parasitical organisms are not only a typical and crucial part of ecosystems, but also affect wild population abundance.Having the power to annihilate their hosts, and are the main driver of evolution [1]. This knowledge of viruses' ubiquitous presence in ecosystems has sparked interest in the field of disease ecology, which is defined as the ecological examination of host-pathogen interactions in relation to their surroundings and evolutionary history. Ecology is the study of how living organisms interact with their environment. More thorough and comprehensive explanations of how biogenetic, psychological, behavioral, social and cultural, and physical environmental factors interact to influence human well-being are provided by ecological research. As a result, pollutants from the environment can cause illnesses like cancer, heart disease, and respiratory problems [2]. In addition to the multifaceted nature of these components and their interrelationships, it is very challenging to track and assess the root causes and consequences of specific diseases. Environment-related variables such as pathogens like fungi and viruses, particles like chemical and biological products wastes, and nutrient and food shortages all have a bearing on the spread of disease in populations and ecosystems [3]. Neglected Tropical Diseases (NTD) have been the subject of enormous effort in recent years, although it is still very difficult to control them, mostly through medicine administration. Given continual destitution, regular exposure to environmental pathogens, and body's acclimatization to medications as a result of the pathogen's persistence in the body, approaches focused purely on treatment by humans or medical care are less effective. However, while seeking to address significant contextual issues in population health, researchers, decision-makers, and public health professionals run into unexpected challenges.

2. Literature Review

Infectious diseases are illnesses caused by microorganisms, usually bacteria, viruses, fungi, or parasites, that are either directly or indirectly passed from one person to another. Infection in humans can also occur as a result of contact with an animal carrying a pathogenic bacterium that can infect people. Simple getting sick, which is different from infectious disease, is the invasion and reproduction of a variety of agents in the body, such as bacteria, viruses, fungi, protozoans, and worms, as well as the tissue's reaction to their presence or to the toxins they produce. In 2013, infectious diseases caused over 9 million fatalities and over 45 million years of lost productivity due to disability [4]. There are times when environmental conditions increase the risk of coming into touch with an infectious pathogen. For instance, environmental disruption following an earthquake may increase the risk of coming into touch with Clostridium tetani and result in host stress, which creates entrance routes for the bacteria. Environmental factors that increase sensitivity may also increase an individual's

vulnerability to infection by altering their physiological state [5].

A. Infectious Disease

A vast variety of pathogens, especially bacteria and viruses, can cause infections [6]. According to WHO, the pathogenic microorganisms that cause infectious diseases, such as bacteria, viruses, parasites, or fungus, can spread the diseases from one person to another either directly or indirectly. These illnesses can be divided into three categories: those that have a high mortality rate, those that lay a large load on disability on populations, and those that, because of their rapid and unanticipated spread, have the potential to have serious worldwide effects.The health sector has little direct control over many of the important factors that affect health and the factors that lead to infectious diseases. Environmental and climatic change, education, agriculture, trade, tourism, transportation, development of industry, and residential are additional sectors that are involved.

B. Disease Ecology

Disease ecology is an area of ecology that focuses on understanding the fundamentals, dynamics, and consequences of host-pathogen interactions, particularly those that lead to infectious diseases[7].There are numerous definitions of disease ecology, but common features are typically emphasized. Kilpatrick and Altizer[8] defined disease ecology as "the ecological study of host-pathogen interactions within the framework of their environment and evolution."They emphasize the discipline's basic objectives of understanding the geographical and biological contexts of parasites, in contrast to parasitology, which has been observed to place more of an emphasis on taxonomy and parasite life cycles. For instance, it looks into the effects and spread of parasites across wildlife populations and communities. [8, 9] Scientists are looking at how diseases spread in the natural environment to learn more about how changes in our environment can impact the migration of viruses and other diseases. Disease ecology, which investigates how interactions between species and abiotic environmental factors influence disease patterns and processes, is a rapidly expanding field of research within ecology.Scientists are looking at how diseases spread in the natural environment to learn more about how changes in our environment can impact the migration of viruses and other diseases. Disease ecology, which looks at how interactions between species and abiotic environmental factors affect disease patterns and processes, is a field of research within ecology that is expanding quickly. Infectious diseases have dominated the study of disease ecology up until now. Thus, the goal of disease ecology is to understand the relationship between ecological interactions and the evolution of diseases. Geographers are nonetheless interested in these effects since disease ecologists have historically focused on how the natural environment affects disease.However, major environmental changes like climate change, urbanization, habitat loss and fragmentation, biodiversity loss, invasions, overexploitation, and pollution, as well as their ripple effects on wildlife and on social and trophic interactions within and between humans and animals, are predicted to have an impact on the spread of disease and host and pathogen reactions.

C. Semantic Machine Learning

In machine learning, the task of creating frameworks that approximately represent concepts from a huge collection of documents is known as semantic analysis of a corpus. Usually, prior knowledge of the texts' semantics is not necessary. Human speech can be examined using a metalanguage based on predicate logic [10].Semantic analysis of a text in machine learning involves creating structures that accurately represent concepts from a large collection of documents. Most of the time, prior understanding of the papers' semantics is not necessary. In order to understand natural language (text), semantic analysis is the process of extracting meaningful information from unstructured data, such as context, emotions, and feelings. It makes it possible for computers and systems to understand, examine, and derive meaning from clauses, sentences, reports, registers, files, or any other similarly seeming content. Semantic analysis, a subfield of natural language processing, and machine learning can be used to comprehend the context of any text and the emotions that might be communicated in a sentence [11].

D. Techniques in Semantic Analysis

Depending on the type of information that we would wish to extract from the provided data, there are two different sorts of semantic analysis methodologies. These are semantic extractors and classifiers.

1. Semantic Classification Models

These are the text categorization models that categorize the provided text into any predetermined categories, different semantic classification models are discussed below;

i. Topic Classification

It is a technique for analyzing any text and categorizing it based on its content into many recognized predefined categories. On the other hand, text classification or topic extraction from text requires knowledge of a text's themes prior to analysis because you must tag

data in order to train a topic classifier. Despite the existence of an additional stage, topic classifiers are far more accurate than clustering methods in the long term.

ii. Topic Classification Scope

There are various levels of scope at which topic analysis can be used:

a. Document-level: The topic model pulls the many themes from a full text to create its subjects. For instance, the subject lines of emails or news articles.

b. Sentence-level: The topic model is able to identify a sentence's topic. For instance, the headline of a news article.

c. sub-sentence level: The topic model derives the subject of sub-expressions from within a phrase at the sub-sentence level. For instance, various subjects could be included in a single sentence of a product evaluation.

d. Sentiment analysis

It is a technique for evaluating whether a text has hidden sentiment that is positive, negative, or neural. This approach aids in comprehending any statement's urgency. Sentiment analysis, often known as opinion mining, is a natural language processing (NLP) method for identifying the positivity, negativity, or neutrality of data. Textual data is frequently subjected to sentiment analysis in order to assist organizations in monitoring brand and product sentiment in consumer feedback and comprehending client wants.

e. Intent classification: With the help of machine learning and natural language processing, intent recognition also known as intent classification associates text data and expression with a certain intent. To put it another way, intent recognition uses a query as an input and connects it with the intended class. It is a technique for differentiating any text based on the customers' purpose.

2. Semantic Extraction Models

Semantic extraction model uses keywords or entity in finding identified entities in text, such as names of individuals, organizations, and locations we enumerate the different types of semantic extraction model below;

i. Keyword Extraction

It is a technique for identifying the precise insights in any text by extracting the pertinent words and expressions. It frequently coexists with the many classification models. It is used to determine which words are "positive" and which words are "negative" after analyzing various keywords within a corpus of text. The most often used subjects or terms may reveal information about the text's purpose.

ii. Entity Extraction

Entity extraction, also known as entity name extraction or named entity recognition (NER), is a method of information extraction that recognizes important textual components before categorizing them. Due to this, unstructured data can now be machine-readable (or organized) and used for common natural language processing (NLP) tasks including information retrieval, fact extraction, and question answering.

3. RESEARCH METHOD

Methodology refers to a way for analyzing a study issue. As a result, it is interested in looking into numerous steps a researcher might take to conduct research as well as the mechanisms that support those steps. The procedures or strategies used to find, select, process, and analyze information about a topic are referred to as research methodology. In this research, the methods we adopt are as follows;

3.1 Data Gathering

Methods for gathering data for research purposes include approaches and procedures that can use either quantitative or qualitative data collection techniques. In this research, we gather epidemiological data from medical literatures in different databases.

3.2 Conceptual Context for Data Gathering

This the data gathering on the ecological factors that affects infectious disease is aim at using ecological sample text from a given environment to classify infectious disease The goal of this dissertation project is to present a qualitative and quantitative analysis in details, in an exhaustive way that describe and identify how ecological factors can affect infectious disease.

3.3 Search Terms

A search of journals, books, related reports and policy statements was conducted in other to provide answers the to the research theme on ecological factors that affect infectious disease based on the selected infectious disease chosen.The search was done in other to discover studies (clinical reports, articles on ecological impacts of infectious disease). The search terms were built around the sample of interest (Ecological factors that affect infectious diseases), the phenomenon of interest (Infectious Disease), evaluation of the studies (qualitative research design.), and the type of research (qualitative and mixed research types). Boolean operators, ‘OR’, ‘AND’ and

wildcard operator '*' are used for combining and stringing search terms in the database. Trial searches was first conducted to ascertain the number of publications before determination of the final search terms. One major reason for conducting trial searches was to ensure that the search terms was able to identified a journal which was independently selected. This process ensures the accuracy of the search terms. The search terms were also selected to sufficiently produce accurate and specific results without necessary large and excessive numbers of publications. Four categories of search strings were used; the first and second search terms focused on the population and the subject of interest; while the third and fourth specify the methods and the type of studies and guaranteed that the studies captured were related to evaluation of the intervention using qualitative research mechanism. The following search strings were used:

("Ecology Factors" OR Ecological Factors* OR Ecological Factors in Nigeria*)

AND

("Infectious Disease" OR "Infectious Disease in Nigeria" or "Ecological Factors in Infection Disease Transmission")

AND

(Effect* OR Transmission* OR Studies* OR "experiments* OR assess* OR Evaluat* OR Efficien*)

AND

("Qualitative" OR Mixed method*)

3.4 Search Strategy

Search strategy is a plan used to identified studies that are relevant to the research. The plan is targeted at discovering literatures which includes clinical samples and reports that will address and identified ecological factors that affects infectious diseases. We choose the following databases for the searching of the ecological samples. The following databases were searched: PubMed and Medline.

i. PubMed: This is an internet-based repository offered by the US National Library of Medicine (NLM), providing a wide range of complimentary materials for searching and accessing journals, reports, books, and other literature pertaining to biomedical, life sciences, and related fields

   Why use PubMed?

   With an impressive collection exceeding 33 million abstracts and bibliographies of biomedical literature, this platform stands out for its extensive resources. Recognized as a trusted and authoritative source in the field of health and medicine, it offers unparalleled reliability. PubMed offers users two search methods: Basic Search and Advanced Search. Basic Search allows for the input of search terms without the necessity of formatting, utilizing logical operators (such as OR, AND, or NOT) to refine the results.

ii. Medline: Medline is the primary module of PubMed; it contains more than 28 million citations including journal papers in biomedical and life science (NLM, 2021). Journals selected for Medline and reviewed by the Literature Section Technical Review Committee; it has a span of 5200 journals in the world over 40 languages.

3.5 Inclusion and Exclusion Criteria

The development of the search terms, inclusion and exclusion criteria were based on the SPIDER search tool which stand for 'Sample' (S), ideally for small samples size usually found in qualitative related researches, 'Phenomenon of Interest' (PI), describing the 'why' and 'how' of certain behaviour or reaction, 'Design' (D), describes the qualitative research framework, 'Evaluation' (E), describes measures of outcomes in a research and 'Research type' (R), supports three research types: qualitative, quantitative or mixed of qualitative and quantitative.

The full detail of the inclusion and exclusion criteria is presented in Table 1 below

Table 1 Inclusion and exclusion criteria

Criteria	Inclusion	Exclusion
Sample	Studies containing data of ecological factors that contribute to the transmission of infectious diseases.	Studies containing data of ecological factors that do not contribute to the transmission of infectious diseases.
Phenomenon of Interest	Infectious disease transmission based in ecological factors.	Other infectious disease transmission that is not based on ecological factors.
Design	Studies that uses observatory, case studies, focus groups, surveys, methods.	Studies that used only surveys.

Evaluation	Evaluation of outcomes include improved infectious disease management, quality of environment, environmental pollution reduction, commitment to efficient environmental care.	Studies that do not report results related improved infectious disease transmission.
Research Type	Qualitative and Quantitative studies with observation, survey, case studies and focus group designs.	Quantitative studies with experimental and other design methods.

3.6 Study Selection

The process of searching for studies can be inefficient at times, as online databases often generate a significant number of irrelevant results. To ensure that the studies chosen for this research meet the selection criteria, the titles and abstracts of the studies were initially reviewed to determine their relevance before obtaining and thoroughly screening the full papers. Additional selection criteria encompassed studies published in English, within the past ten years (2012-2022).Nevertheless, we present a PRISMA flow diagram that was used for the assessment of our ecological Samples results. The search process and study selection are described in the PRISMA chart in figure 1 below.

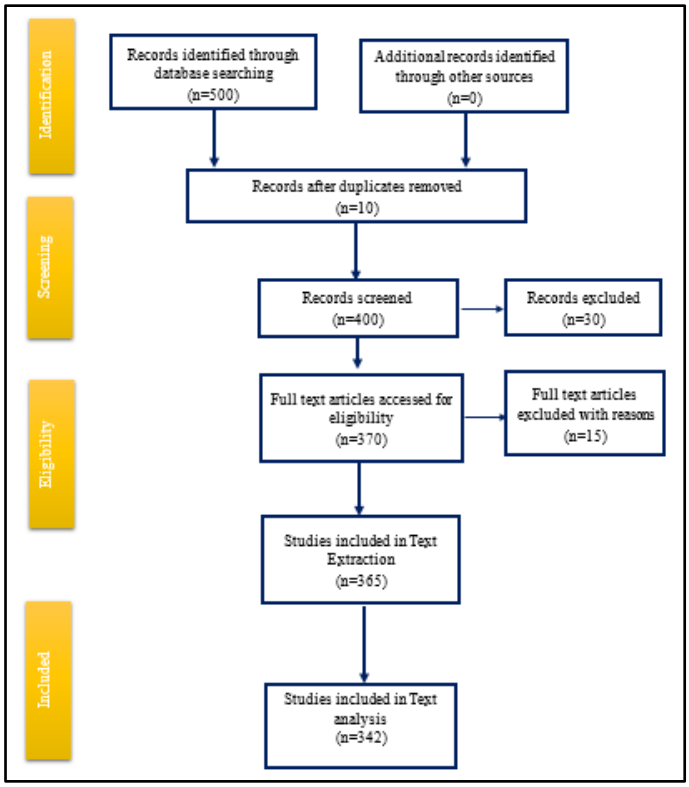

Fig1: PRISMA Flow Diagram

Furthermore, we present the cross section of the Ecological Datasets gathered on infectious disease which we present in table 2.

Table 2: Ecological Datasets

Disease	Sentence	Ecological Factor	Journal Title	Journal Source	Year of Journal
Malaria	Deforestation disrupts natural ha	Deforestation	"The Impacts of Deforestation on Malaria in Nigeria"	Environmental Science and Pollution Research	2018
Malaria	Standing water in discarded tires	Discarded Tires	"Ecological Factors Affecting Malaria Transmission in	Journal of Vector Borne Diseases	2016
Malaria	Poor waste management leads to	Waste Management	"Effect of Solid Waste Management on Malaria Transm	Nigerian Journal of Parasitology	2017
Malaria	Urbanization results in increased	Urbanization	"The Role of Urbanization in Malaria Transmission in N	Environmental Health Insights	2015
Malaria	Agricultural practices promote ma	Agricultural Practices	"Impact of Agricultural Activities on Malaria Transmis	Nigerian Journal of Medical Research	2019
Malaria	Mining activities disrupt natural e	Mining Activities	"Malaria Transmission and Mining Activities in Nigeria	Journal of Environmental and Public Health	2018
Malaria	Dams and reservoirs provide bree	Dams and Reservoirs	"Role of Dams in Malaria Transmission in Nigeria"	Journal of Water and Health	2017
Malaria	Climate change alters mosquito d	Climate Change	"Impacts of Climate Change on Malaria Transmission i	Environmental Research Letters	2016
Malaria	Oil exploration and drilling disru	Oil Exploration and Drillin	"Ecological Consequences of Oil Exploration on Malari	African Journal of Environmental Science and Te	2018
Malaria	Increased deforestation for agricu	Agricultural Expansion	"Deforestation, Agriculture, and Malaria Transmission	Malaria Journal	2015
Malaria	Water pollution from human activ	Water Pollution	"Role of Water Pollution in Malaria Transmission in Ni	Journal of Environmental Science and Health, Pa	2017
Malaria	Urban expansion leads to decreas	Urban Expansion	"Effects of Urbanization on Malaria Transmission in Ni	African Journal of Biotechnology	2016
Malaria	Construction activities disrupt na	Construction Activities	"Ecological Impacts of Construction Activities on Mala	Journal of Environmental Planning and Manage	2019
Malaria	Pesticide use in agriculture and v	Pesticide Use	"Pesticide Resistance and Malaria Transmission in Nig	International Journal of Environmental Health R	2017
Malaria	Erosion from deforested areas co	Erosion	"Deforestation and Erosion Effects on Malaria Transmi	Nigerian Journal of Clinical and Biomedical Res	2018
Malaria	Livestock farming leads to increas	Livestock Farming	"Role of Livestock Farming in Malaria Transmission in	Veterinary World	2016
Malaria	Standing water provides breeding	Water Stagnation	Journal of Vector Ecology	Wiley Online Library	2021
Malaria	Increased rainfall leads to a rise	Rainfall	Malaria Journal	BMC	2019
Malaria	Deforestation disrupts natural ec	Deforestation	Acta Tropica	Elsevier	2018
Malaria	Polluted water sources attract mo	Water Pollution	Journal of Environmental Health	NEHA	2020
Malaria	Urbanization and population grov	Urbanization	Journal of Infectious Diseases	Oxford Academic	2022
Malaria	Agricultural practices such as irri	Irrigation	Parasites & Vectors	BMC	2020
Malaria	Temperature fluctuations can affe	Temperature	Parasitology Research	Springer	2017
Malaria	Land use changes alter mosquito l	Land Use	Environmental Research	Elsevier	2021
Malaria	Seasonal changes affect mosquito	Seasonal Changes	BMC Infectious Diseases	BMC	2019
Malaria	Deficient waste management lead	Waste Management	Journal of Environmental Health	NEHA	2022

3.7 MODEL DESIGN

Today, AI enterprises are being implemented in a variety of businesses for a variety of uses. Predictive modeling, pattern recognition systems, automated vehicles, interactive systems, high energy activities, and goal-driven systems are examples of these applications. Each of these ventures has one thing in common: they're all based on an understanding of the business challenge and the need to apply data and machine learning algorithms to it, leading inside a machine learning model that meets the design specifications. A machine Learning model is a computer program which has been trained to recognize specific patterns[12]. We train the model on a set of data and give it an algorithm to use to reason about and learn from that data. Once the model has been trained, you can use it to reason over data it hasn't seen before and make predictions about it[13]. Hence, the following steps aided in our model design in this research work;

i. Define our problem clearly (goal, expected outcomes, etc.).

ii. Obtain information (i.e., data).

iii. Select a metric for success.

iv. Determine the framework and the various procedures that are available.

v. Prepare the information (dealing with missing values, with categorical values).

vi. Correctly spill the info.

vii. Explain the differences between overfitting and underfitting, including what they are and how to avoid them.

viii. A brief description of how a model learns.

ix. What is regularization, and when should it be used?

x. Create a benchmarking model.

xi. . Select an appropriate model and fine-tune it to achieve the best potential results.

3.8 Problem Definition / Model Formulation

Classifying infectious diseases based on ecological factors poses a significant challenge in the field of epidemiology. The transmission dynamics of diseases are influenced by various ecological factors, including climate, vegetation, and animal populations. However, there is a pressing need to develop an effective and precise classification system that utilizes these ecological factors to enhance disease surveillance, prevention, and control strategies.

The current issue revolves around the limited understanding of the intricate interactions between ecological factors and infectious diseases. Existing classification systems primarily focus on specific pathogens or geographic regions, failing to consider the broader ecological context. Consequently, the ability to predict disease outbreaks, identify vulnerable populations, and allocate resources efficiently is hindered.

To overcome this challenge, it is imperative to create a comprehensive infectious disease classification framework that integrates ecological factors as crucial determinants. This framework should employ advanced computational techniques like machine learning and data mining to analyze extensive ecological and epidemiological datasets. By accurately categorizing diseases based on ecological factors, public health agencies and policymakers can make informed decisions to effectively mitigate the impact of infectious diseases on human and animal populations. In this research, we employ Semantic-Based NLP system for the classification of infectious disease. Natural Language Processing (NLP) is an interdisciplinary field that combines linguistics, computer science, and artificial intelligence to enable computers to comprehend and process human language. One common application of NLP is topic classification, where the objective is to categorize text documents into predefined topics or categories. The process that is embarked on the semantic-based NLP topic classification in our model involves several steps.

i. **Data Preparation**: In this step, the raw text data is prepared by eliminating noise like punctuation marks and stopwords, and then tokenizing the text into individual words or phrases. The mathematical representation of this step can be defined as follows: Given a raw text document say "D", the preprocessed document D' is obtained by removing noise and tokenizing the text: D' = preprocess(D)

ii. **Feature Extraction**: for us to presents the documents of our data sets numerically, relevant features must be extracted. One common method is the application of a bag-of-words model, where each document is represented as a vector of word frequencies. Therefore, we can compute the term frequency-inverse document frequency (TF-IDF) score for

each word in the document to give more importance to words that are rare in the overall corpus. The mathematical representation of this step presented as;

For each document D' in the corpus C, compute the TF-IDF score for each word say "w"

$$TF - IDF(D^{'}, w) = TF(D^{'}, w) * IDF(w) \#1$$

where TF(D', w) is the term frequency of word w in document D', and IDF(w) is the inverse document frequency of word w across the entire corpus.

iii. **Model Formulation**: nevertheless, in order to classify the documents into topics, we need to train a classification model. One popular model for text classification is the XGBoost classifier, which assumes that the features (word frequencies) are given the class label. The mathematical formulation of the XGBoost classifier can be defined as: Given a document D' represented by feature vector X, we predict the class label y:y = argmax(P(y) * ∏ P(x_i|y)) mwhere P(y) is the prior probability of class y, P(x_i|y) is the probability of feature x_i given class y, and the product is taken over all features in X.

iv. Iv. Training and Evaluation of Classification Model: In this step, we split the labeled dataset into a training set and a test set. We use the training set to train the topic classification model, and then evaluate its performance on the test set using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score. The mathematical formulation of the evaluation metrics can be defined as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

where TP (True Positive) is the number of correctly classified positive instances, TN (True Negative) is the number of correctly classified negative instances, FP (False Positive) is the number of wrongly classified positive instances, and FN (False Negative) is the number of wrongly classified negative instances. By following these steps and using appropriate mathematical formulations, we can build a natural language processing system for topic classification.

v. Model Deployment with DistilBERT and ML Algorithm. Hence, this step followed for model deployment for Application programing interface

4. RESULTS

We present the results in this research as follows;

4.1 Label Encoding:

Label encoding is another form of data preprocessing technique which involves the conversion of categorical labels into format acceptable in the machine learning process. From figure above, we observed that the class of diseases labelled 'Disease' in the dataset is in categorical data type, therefore, a one hot label encoding technique was used to convert the categorical data into numbers. All the nine classes of disease considered in this research was converted into numbers starting from 0 to 8. Therefore, malaria class of diseases was replaced by 0 while tuberculosis was replaced by 1 as shown in the Figure 2 below.

	Disease	Sentence	label_encode
0	Malaria	Deforestation disrupts natural habitats, incre...	0
1	Malaria	Standing water in discarded tires creates idea...	0
2	Malaria	Poor waste management leads to stagnant water,...	0
3	Malaria	Urbanization results in increased population d...	0
4	Malaria	Agricultural practices promote mosquito breedi...	0

Fig 2: One-hot label encoding

4.2 Model Training and Testing Split

After the preprocessing of the dataset, the processed data will be split into training and testing sets using the popular sklearn python library known as the train_test_split as illustrated the code snippet shown below. The data will be split into 75% for training while the remaining 25% was reserved for testing which is shown in figure 3.

Fig 3: Train-Test Split

From the use of Xgboots Algorithm in training the data, we present the result of the training in figure 4.

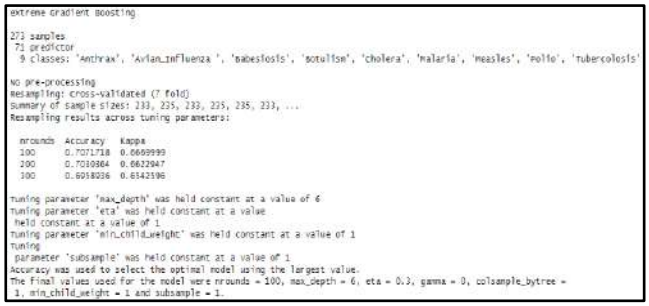

Fig 4XGboost Training Results

Also, we present the accuracy of XGboost using nrounds of 100, 200 and 300 the Accuracy and Kappa results shows that the final values used for the model were nrounds = 100, max_depth = 6, eta = 0.3, gamma = 0, colsample_bytree = 1, min_child_weight = 1 and subsample = 1, which gives 70% on accuracy and 66% Kappa.

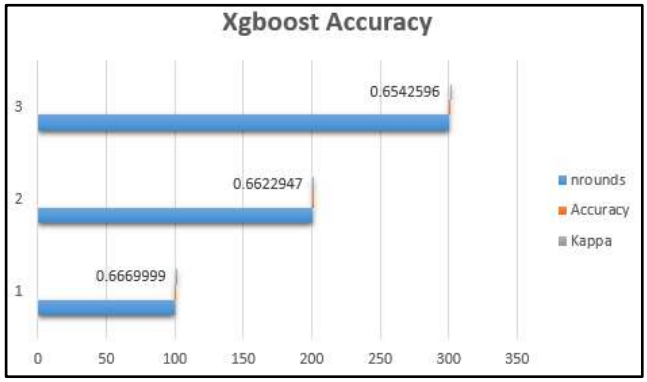

Fig 5: XGBoost Accuracy

Again, figure 6 XGBoots depicts different bosting iterations for cross-validations.

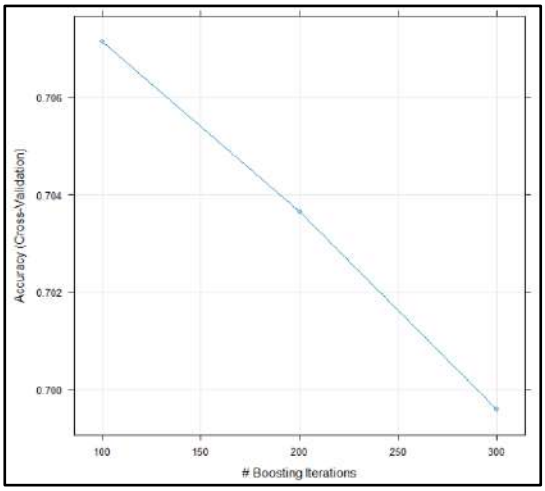

Fig :6 XGBoots Accuracy

4.2 Importing DistilBERT Pretrained Model and Tokenizer

At this point, we will import The DistilBERT pretrained model to be trained along side our classification Algorithm which here is XGBoots which gave 70% based on cross validation accuracy measure .

4.3 Tokenization

After the importation of the pretrained model and its tokenizer, the dataset through a lambda function will be tokenized. The tokenization will be done in batches and the output is presented in Figure 7 below.

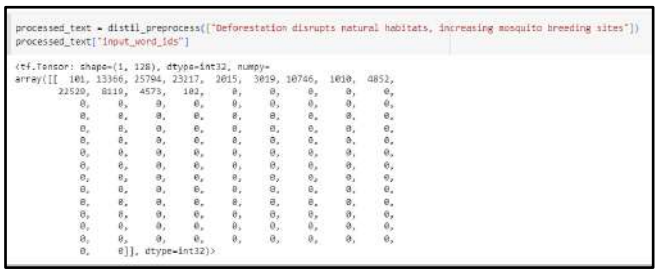

Fig 7: Output of tokenizer function

The output of the DistilBERT tokenizer is a tuple of shape (1, 128) for one sequence. As its observed from Figure 4 above, the sentence "Deforestation disrupts natural habitats, increasing mosquito breeding sites" does not contain up to 128 words, therefore, 0 is used to complete the array in order for all the dataset to have the same fixed length. This process of adding 0s to the array is known as padding. Figure 8 below illustrates the processing of the tokens by the DistilBERT model.

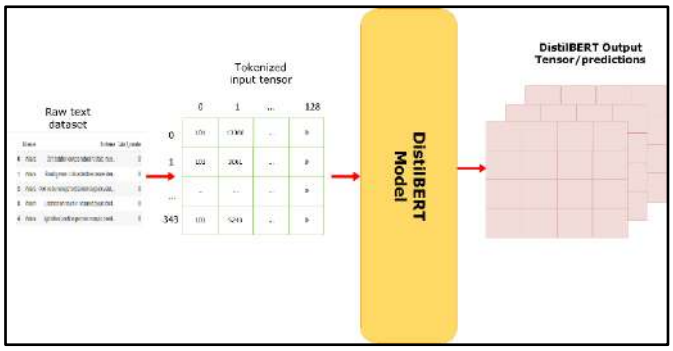

Fig8: Tokens processing

The full process that involves the transformation of each sentence from the dataset is presented in the Figure9 below.

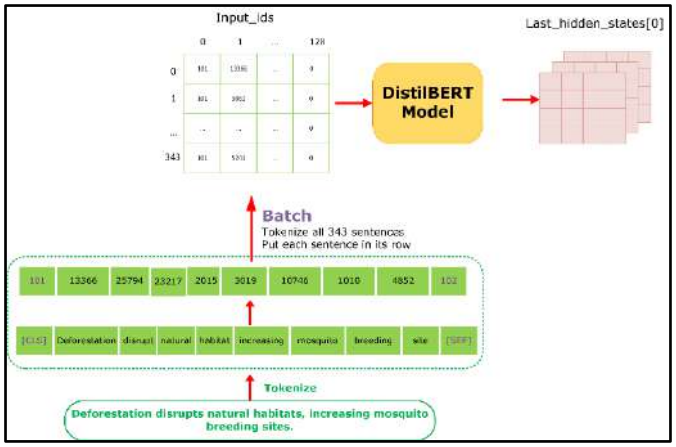

Figure 9: dataset transformation

In sentence classification, only the BERT output with (Classification Token) are considered. In this case, a slice of the output cube is selected and everything else is discarded as illustrated in figure 10 below.

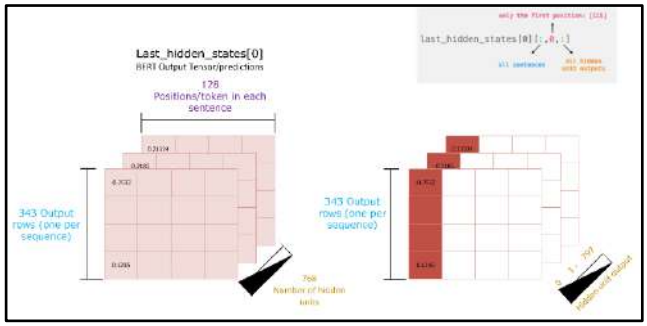

Fig10: Slicing of the last hidden state of the BERT outputs

From figure9, the 3-dimensional tensor is sliced resulting in a 2-dimensional tensor appropriate for the training of our Classification Model (XGBoots). The feature set containing the 2-dimensional array of the sentence embeddings for all the sentences in the project dataset is presented in Figure 11 below.

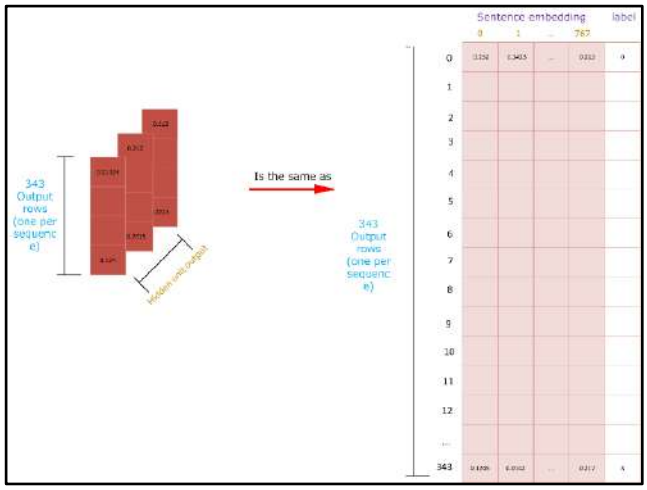

Fig11: Feature set generated from the sliced output of the BERT model

After extracting the output of the BERT model, we have successfully created a new dataset for the XGboots model. The 769 columns from Figure 7 above represent the features and the label column is the encoded label from the initial dataset. Hence, we present the final features which we will use Bert and XGbost algorithm for the final classification, which is presented in the Figure 12 below.

Fig: 12 features and labels for XGBoost

Nevertheless, we present the final Application Programming Interface for the ecological infection disease classification where a user can describe the ecological factors within his or her environment and proceed to click the classifier button which then classifies the disease with a confidence interval among all the classes of the disease. Hence, figure 13 depicts the ecological classification interface.

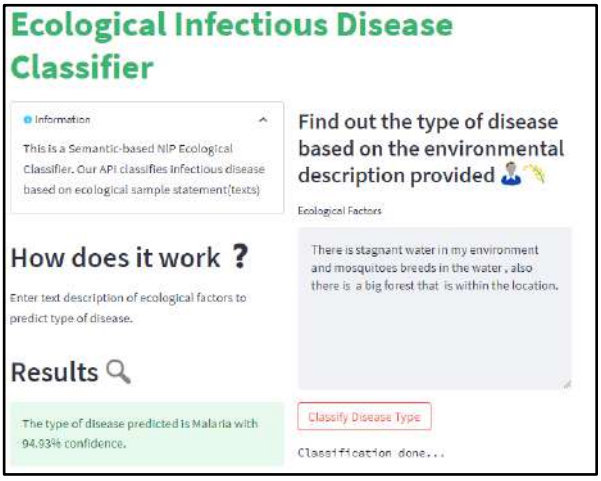

Fig 13: Ecological Classification Interface

Furthermore, we present the model interpretability results of the classification of infectious disease how the decision was made and the criteria it uses in the decision making which is depicted in figure 14.

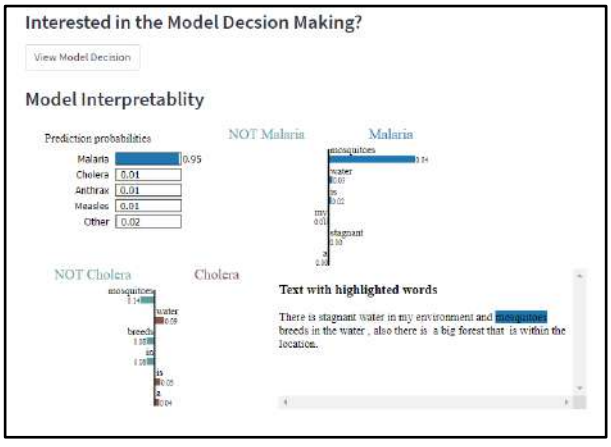

Fig 14: Model Decision Making

5. CONCLUSIONS

In the contemporary world, the role of ecological factors in the transmission of infectious diseases has gained prominence. The intricate connection between ecosystems, human activities, and pathogens has created an environment conducive to the emergence and spread of diseases. Several ecological factors contribute to this phenomenon, including the destruction of habitats, climate fluctuations, urbanization, and the global movement of people and goods. Human-induced habitat destruction and deforestation, particularly driven by activities like agriculture and urban expansion, disrupt natural ecosystems and bring humans into closer contact with wildlife. This proximity heightens the risk of zoonotic diseases, which are illnesses transmitted from animals to humans. Another significant ecological factor influencing disease transmission is climate change. Alterations in temperature and rainfall patterns impact the distribution of disease vectors such as mosquitoes and ticks, thereby facilitating the spread of diseases like malaria, dengue fever, and Lyme disease to new regions. Urbanization also plays a pivotal role. Rapid and poorly planned urban growth leads to overcrowded living conditions, inadequate sanitation, and improper waste management. These conditions create ideal breeding grounds for disease-carrying vectors and facilitate the transmission of infectious diseases.

Nevertheless, effectively controlling and mitigating the impact of these ecological factors on disease transmission necessitates comprehensive measures addressing both the root causes and immediate consequences. This involves implementing sustainable land-use practices, preserving natural habitats, promoting responsible urban planning, and enhancing surveillance systems and early warning mechanisms.

Consequently, the emergence of natural language processing (NLP) has significantly impacted the healthcare sector. NLP, a branch of artificial intelligence focused on computer-human language interaction, has revolutionized various aspects of healthcare, including medical diagnosis, clinical decision-making, and patient monitoring.

In the realm of infectious diseases and ecology, NLP has the potential to contribute significantly by aiding in the classification and analysis of disease-related data. By processing extensive textual information from sources such as research articles, clinical reports, and surveillance data, NLP algorithms can extract valuable insights into the ecological factors influencing disease transmission. This knowledge can inform the development of effective prevention strategies and targeted interventions.

Moreover, NLP facilitates the integration of data from diverse sources, including ecological data, clinical records, and genetic information. This holistic approach enhances our understanding of the complex dynamics between ecological factors and infectious diseases, empowering researchers and policymakers to make informed decisions and implement proactive measures to mitigate the impact of ecological factors on disease transmission.

REFERENCES

[1] Hudson, P. J., Rizzoli, A., et al. The Ecology of Wildlife Diseases. Oxford, UK, Oxford Press (2002).

[2] Brusseau, M.L., Ramirez-Andreotta, I.L., & Maximillain, J. (2019). Environmental Impacts on Human Health and Well-Being. Environmental and Pollution Science, 3, 477-499. https://doi.org/10.1016/B978-0-12-814719-1.00026-4.

[3] McMichael AJ. (1993). Planetary Overload: Global Environmental Change and the Health of the Human Species. Cambridge (UK): Cambridge University Press.

[4] Naghavi M., Wang H., Lozano R. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015;385:117–171.

[5] van Seventer JM, Hochberg NS. Principles of Infectious Diseases: Transmission, Diagnosis, Prevention, and Control. International Encyclopedia of Public Health. 2017:22–39. doi: 10.1016/B978-0-12-803678-5.00516-6. Epub 2016 Oct 24. PMCID: PMC7150340

[6] Sehgal, M., Ladd, H. J., & Totapally, B. (2020). Trends in epidemiology and microbiology of severe sepsis and septic shock in children. *Hospital Pediatrics*, 10(12), 1021-1030.

[7] Ostfeld, Richard S. (2015), "Disease Ecology", Ecology, Oxford University Press, doi:10.1093/obo/9780199830060-0128, ISBN 978-0-19-983006-0.

[8] Kilpatrick, A. M., AND S. Altizer. (2010). Disease ecology. Nature Education Knowledge 1: 13.

[9] Hawley, Dana M.; Altizer, Sonia M. (2011). "Disease ecology meets ecological immunology: understanding the links between organismal immunity and infection dynamics in

natural populations". *Functional Ecology*. 25 (1): 48–60.

[10] Nitin Indurkhya; Fred J. Damerau (2010). Handbook of Natural Language Processing. CRC Press. ISBN 978-1-4200-8593-8.

[11] Umoren I., Abe O., Ansa G., Inyang S., and Umoh I. (2023) A New Index for Intelligent Classification of Early Syndromic of Cardiovascular (CVD) Diseases Based on Electrocardiogram (ECG), *European Journal of Computer Science and Information Technology*, 11 (4), 1-21

[12] Ekong, A., Silas, A., & Inyang, S. (2022). A Machine Learning Approach for Prediction of Students' Admissibility for Post-Secondary Education using Artificial Neural Network. *Int. J. Comput. Appl*, 184, 44-49.

[13] Umoren, I. J., & Inyang, S. J. (2021). Methodical Performance Modelling of Mobile Broadband Networks with Soft Computing Model. *International Journal of Computer Applications, 174*(25), 7-21.

Authors' Biographies

Saviour J. Inyang holds the National Diploma in computer Science from AkwaIbom state Polytechnic, BSc degree in Computer Science from Akwa Ibom State University, Nigeria and he is currently undergoing MSc. in Computer Science, Akwa Ibom State university, Nigeria.His research interests include Natural language Processing, Machine Learning Soft Computing, Computational intelligence, Artificial intelligence, and Modeling of Communication Networks. He has co-authored several articles in reputable local and international journals.

Imeh J. Umoren holds B.Sc., MSc. and PhD Degrees in Computer Science respectively. In the early stage of His career, he was trained in University of Calabar, Nigeria on Unix/Linux Technologies and Data Processing from 1986-1988. He held a Total E&P Graduate Scholarship from 2010-2012. Currently, he is an academic staff in the Department of Computer Science, Akwa Ibom State University (AKSU), Mkpat Enin, Nigeria. Dr. Umoren, has more than 12 years' experience in the Academic, Research and Global ICT industry. He is a strategic thinker, Analyst and a Multi-Tasking Development Specialist focusing on people, process and technology. He is passionate about the use of technology (Computational Intelligence) to transform lives, create wealth and accelerate development in Africa (Nigeria) and globally. Imeh Umoren is a member of several high-profile professional bodies including - IEEE (an elite group within the global engineering community), a recognized member of Polish Neural Networks Society, a member of Nigeria Computer Society (NCS), Internet Society etc. Dr. Imeh Umoren has over 60 published articles and conference papers. His research interests focus on Mobile Computing, Complexity Theory, Algorithms, Computational Intelligence, Machine Learning (ML) and Natural Language Processing (NLP).