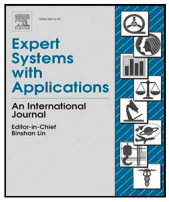
Expert Systems With Applications 250 (2024) 123894

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Explainable epidemiological thematic features for event based disease surveillance

Edmond Menya [a,c,d,*], Roberto Interdonato [b,c], Dickson Owuor [d], Mathieu Roche [b,c]

[a] Université de Montpellier, Montpellier, France
[b] CIRAD, F-34398 Montpellier, France
[c] TETIS - Univ Montpellier - AgroParisTech - CIRAD - CNRS - INRAE, Montpellier, France
[d] SCES - Strathmore University, Nairobi, Kenya

ARTICLE INFO

Keywords:
Epidemiology intelligence
Disease surveillance
Text mining
Corpus classification

ABSTRACT

Event based disease surveillance (EBS) systems are biosurveillance systems that have the ability to detect and alert on (re)-emerging infectious diseases by monitoring acute public or animal health event patterns from sources such as blogs, online news reports and curated expert accounts. These information rich sources, however, are largely unstructured text data requiring novel text mining techniques to achieve EBS goals such as epidemiological text classification. The main objective of this research was to improve epidemiological text classification by proposing a novel technique of enriching thematic features using a weak supervision approach. In our approach, we train and test a mixed domain language model named EpidBioELECTRA to first enrich thematic features which are then used to improve epidemiological text classification. We train EpidBioELECTRA on a large dataset which we create consisting of 70,700 annotated documents that includes 70,400 labeled thematic features. We empirically compare EpidBioELECTRA with both general purpose language models and domain specific language models in the task of epidemiological corpus classification. Our findings shows that epidemiological classification systems work best with language models pre-trained using both epidemiological and biomedical corpora with a continual pre-training strategy. EpidBioELECTRA improves epidemiological document classification by 19.2 F_1 score points as compared to its vanilla implementation BioELECTRA. We observe this by the comparison of BioELECTRA verses EpidBioELECTRA on our most challenging dataset PADI-Web$_{XL}$ where our approach records 92.33 precision score, 94.62 recall score and 93.46 F_1 score. We also experiment the impact of increasing context length of train documents in epidemiological document classification and found out that this improves the classification task by 7.79 F_1 score points as recorded by EpidBioELECTRA's performance. We also compute Almost Stochastic Order (ASO) scores to track EpidBioELECTRA's statistical dominance. In addition, we carry out ablation studies on our proposed thematic feature enrichment approach using explainable AI techniques. We present explanations for the most critical thematic features and how they influence epidemiological classification task We found out that biomedical features (such as mentions of names of diseases and symptoms) are the most influential while spatio-temporal features (such as the mention of date of a given disease outbreak) are the least influential in epidemiological document classification. Our model can easily be extended to fit other domains.

1. Introduction

Epidemic Intelligence systems monitor varying channels for early warning signs to detect and alert on novel and re-emerging infectious diseases. These systems automate the early warning signs detection-and-alerts framework formally defined by the World Health Organization (WHO) under Early Warning and Response (EWAR) standards for acute public or animal health events (WHO, 2014). This is either done through monitoring of *formal sources* such as routine prepared medical reports (e.g., weekly or monthly medical reports prepared by clinicians) or through monitoring *informal sources* such as blogs, hotlines, social media posts and web news articles. These subdivisions leads to two broad classifications of epidemic surveillance systems; Indicator Based Surveillance (IBS) and Event Based Surveillance (EBS) (WHO, 2008).[1]

* Correspondence to: TETIS - Univ Montpellier - AgroParisTech - CIRAD - CNRS - INRAE, 500 rue J.F. Breton, 34093 Montpellier Cedex 5, France.
E-mail addresses: emenya@strathmore.edu (E. Menya), roberto.interdonato@cirad.fr (R. Interdonato), dowuor@strathmore.edu (D. Owuor), mathieu.roche@cirad.fr (M. Roche).

[1] WHO's International Health Regulations IHR-2005 Article 9 on informal sources opened up opportunities for informal epidemic data collection for monitoring purposes and Article 5 on surveillance allows for surveillance using these collected information.

https://doi.org/10.1016/j.eswa.2024.123894
Received 26 May 2023; Received in revised form 18 March 2024; Accepted 1 April 2024
Available online 5 April 2024

Event based surveillance (EBS) monitors informal sources containing real world events and reports on possible public or animal health emergencies. This approach to epidemic surveillance is subtly challenging but far more practical than the classical indicator based approach which monitors formal sources. Informal epidemiological data sources carry more real time information, which when mined, can lead to timely reporting on infectious disease outbreaks way before patients begin flooding health centers for diagnostics. Simply put, EBS does not wait for a gaggle of patients to flock health centers for a health risk alert to be generated. This is especially advantageous for monitoring health events in remote and underdeveloped regions where health care is primarily accessed via informal channels.

Informal source monitoring gives EBS surveillance systems the ability to capture extremely rare but high impact disease outbreaks as well as capturing emerging yet unknown infectious diseases. However, these sources contains unstructured data which proves more challenging to mine in order to produce relevant infectious diseases alerts. In addition, such informal sources are susceptible to rumors, hearsay, fake-news, or even irrelevant reports containing information simply indirectly related to infectious disease such as reports of economic or political consequences that follows the announcing of a disease outbreak. These kinds of (mis)information co-exists alongside genuine epidemiological information that directly reports on disease outbreaks making difficult the task of differentiating between relevant and irrelevant documents.

Contemporary EBS systems follow the five phases of epidemic intelligence namely *detection, triage, verification, risk assessment* and *communication* as captured in Fig. 1. Interests to automated text-based EBS systems mainly covers automation of triage (Arsevska et al., 2018; Brownstein & Freifeld, 2007; Woodall, 2001). Detection is by collection of raw online news articles, social media posts, blogs or emails while triage aims at filtering relevant versus irrelevant articles. A relevant article is one which corresponds to genuine acute public or animal health events.[2] In a manual EBS systems, signaling task is handled by event assessment team of experts who are tasked with the responsibility of assessing each reported public or animal health event after which they can trigger responses (WHO, 2008). In automated EBS systems such as the Platform for Automated extraction of Animal Disease Information from the web (PADI-Web) and HealthMap, this signaling task is automated and handled by an epidemiological corpus classifier engine trained to classify informal corpus sources as either relevant or irrelevant (Brownstein & Freifeld, 2007; Valentin et al., 2021). In other automated EBS systems, such as the Program for Monitoring Emerging Diseases (ProMED), signaling task has traditionally remained a human expert signaling approach to maintain high standards in their alerting system (Woodall, 2001).

Several challenges align with these modern EBS structures. Firstly, there is always a limited number of event assessment human experts to manually classify incoming signals as either relevant or irrelevant such as seen in Woodall (2001) approach to EBS. This challenge lead to a large ratio of the number of reported public or animal health risk events to the size of event assessment unit team of experts, this negatively affects both efficiency and accuracy of event classification. Secondly, classification accuracy for the digital EBS systems is quite challenged by the unstructured nature of source data. These EBS classifiers are trained using informal corpora from varied sources. We observe from Valentin et al. (2021) how poor classification accuracy leads to false alarms on event outbreaks when the system is finally deployed. Thirdly, the approach of training EBS classifiers on informal corpora sourced online leads to bias challenges especially in countries with limited freedom of speech issues (such as restricted or censored communication media) where the amount of data available for training and detection of public

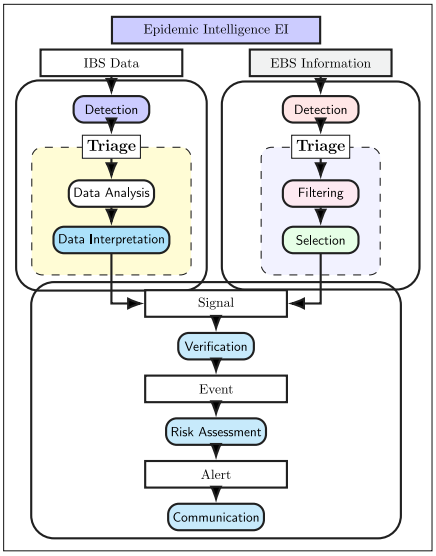


Fig. 1. Epidemic intelligence framework flow.

or animal health events is limited. These luck of quality regional data leads to poor and inaccurate event classifiers (WHO, 2008). Given these challenges, the aim of modern day digital biosurveillance is thus focused on improving efficiency and accuracy of both triage and signaling activities. This aims at avoiding false positives (errors that fails to trigger early disease outbreak alerts) and false negatives (errors that ignores emerging health emergencies).

To achieve these improvements of EBS, various approaches have been suggested. For instance, PADI-Web introduced both keyword based bag of words approach and machine learning document classification techniques (Arsevska et al., 2018). PADI-Web 1.0 used a keyword-based classification approach where epidemiological corpora are classified based on existence of one or more predefined list of disease outbreak-related keywords (Arsevska et al., 2018). PADI-Web 2.0 by Valentin et al. (2020) further enhanced this classifier by incorporating a multilingual module and machine learning techniques based on bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF) approaches (Jones, 1972; Luhn, 1957). Later, Valentin et al. (2021) proposed PADI-Web 3.0 with a fine-grained classification of sentences in order to identify specific classes (e.g. Descriptive epidemiology, Preventive and control measures, Economic and political consequences, etc.).

Further improvement has been introduced using deep learning techniques. EpidBioBERT uses a pre-trained biomedical language model to enrich epidemiological thematic features to enhance its classification approach (Menya et al., 2022). This kind of technique is uniquely attributed to great advances in deep learning text mining approaches based on word embeddings and language models. In this paper, we improve EBS classification to State-Of-The-Art (SOTA) levels by extending the works of Menya et al. (2022) and improving model description. In this extended approach, we propose EpidBioELECTRA, a mixed-domain language model for learning epidemiological thematic features to improve classification in disease surveillance. EpidBioELECTRA is an improved EBS classifier trained on a large dataset one hundred times more compared to EpidBioBERT. EpidBioELECTRA achieves SOTA performance compared to competing language models on the epidemiological classification task. In addition and we carry out in depth study on our thematic-feature approach using explainable AI from our vast dataset. We discover the best pre-training strategy for improving epidemiological classification and the most critical thematic features in epidemic surveillance tasks.

[2] We coin this definition of a relevant article since past digital epidemiological surveillance research have varying definitions depending on the task at hand.

Table 1
Systematic Literature review on early disease surveillance corpus classification systems.

Study	Dataset Origin	Classification algorithms	Main findings
Earley EBS Systems			
ProMED (Woodall, 2001)	OnlineNews, OIE, ProMED curated articles	None (human experts)	**Vast number** of human experts and **longer periods** of time needed for corpus classification.
HealthMap (Brownstein & Freifeld, 2007; Freifeld et al., 2008)	OnlineNews, OIE, ProMED curated articles	N-Gram Parser algorithm with Dictionary matches	Classifier automation **speeds up** EBS early alert systems, however its accuracy needs improvement.
BIOCASTER (Collier et al., 2008)	OnlineNews	SVM	Use of machine learning algorithms **improves detection and classification** of public health rumors in EBS.
MedISys (Jens et al., 2010)	OnlineNews	Dictionary based boolean word combinations	Study showed that MedISys and more generally **media information** from publicly available sources, **can contribute to an integrated monitoring** effort in EBS.

The rest of the paper is organized as follows: Section 2 outlines related works in animal disease surveillance majorly reviewing PADI-Web, EpidBioBERT, HealthMap and ProMED document classifiers; Section 3 presents the formal discussion on thematic features within our model's theoretical framework and pipeline; Section 4 discusses our experiment set up and preparation steps performed on both data and model; Section 5 presents empirical results from our experiments comparing our model with baselines and other competitive models followed by ablation study findings on thematic feature contributions to our improved classifier; Section 6 discusses the limitations of the proposed methods in light of the contributions; Section 7 summarizes the entire paper and recommends future works.

2. Related works on animal disease event based surveillance corpus classification

In the EBS field, several studies on epidemiological document classification exists and we review them in relation to ours. However, EBS processes have been majorly studied under public health domain a lot more than in animal health domain. In this section we thus study state of the art (SOTA) related systems and we contrast them with our proposed contributions in animal health domain. Previously, animal disease surveillance has been studied within a one health context (i.e. in relation to both public health surveillance and environmental surveillance) and in different approaches such as keyword-based (Centre et al., 2011; Chanlekha et al., 2010; Steinberger et al., 2010), machine-learning based (Carter et al., 2020; Collier et al., 2008) and multilingual (Mutuvi et al., 2020; Sahnoun & Lejeune, 2021). We start our literature review by studying classification domain of early EBS systems, followed by contemporary EBS systems and finally EBS systems focused on animal health.

2.1. The corpus classification modules of early EBS systems

Epidemiological corpus classification has been a key module in EBS systems from the very onset. In this section we review two SOTA systems and summarize the rest in Table 1.

2.1.1. ProMED classifier

The program for monitoring emerging diseases (ProMED) is one of the earliest digital biosurveillance system (Woodall, 2001). ProMED curates information from both formal and semi-formal systems inclusive of official reports such as those produced by clinicians, formal websites such as ministry of health or local health department websites, media news reports, social media posts and ground observer reports. These varied information is reviewed, vetted and commented on by a team of expert epidemiologist moderators to create signals in the ProMED network. Infectious disease articles are color-coded as either red, yellow or green ranging from relevant to irrelevant. A key challenge with these approach is the vast number of epidemiologist and long periods of time needed to mine through the vast pool of infectious disease data. These elements are known to affect quality of alerts and signals which are generated to trigger mitigation measures in case of a public or animal health emergencies requiring early response (Yu & Madoff, 2004).

2.1.2. HealthMap classifier

HealthMap is a semi-automatic biosurveillance event based system (Brownstein & Freifeld, 2007). HealthMap collects epidemiological corpora from diverse sources namely; online news articles, World Organization for Animal Health (OIE) data, national health authorities and ProMED curated articles. The early signaling capabilities of HealthMap classifier was built around human moderators rating incoming documents on a scale of 1 to 5 ranging from less relevant to most relevant disease outbreaks of international concern (Brownstein et al., 2008). However, the need to introduce automation led to algorithm driven classification engine which demonstrated significant usefulness in managing large volume of information processed by HealthMap (Freifeld et al., 2008). This classification engine used parser algorithms with dictionary databases to find keywords of interest that are used to determine the relevance of an epidemiological article. Furthermore, fine grained classes were introduced corresponding to the 1 to 5 scale are *breaking news, warning, old news, context* and *not disease related* as captured in Fig. 2. However, even though this early introduction of epidemiological classification using keywords was a major breakthrough towards EBS automation, the key challenge was achieving competitive accuracy scores with solely using keywords and the limitation of dictionary vocabularies. Also, handling the ever increasing fast number of online article sources was a major challenge (Brownstein et al., 2008). In our approach, we propose a model that uses deep contextualized models which uses entire document sentences as (opposed to simple keywords) to learn patterns in an epidemiological document which significantly improve classification accuracy.

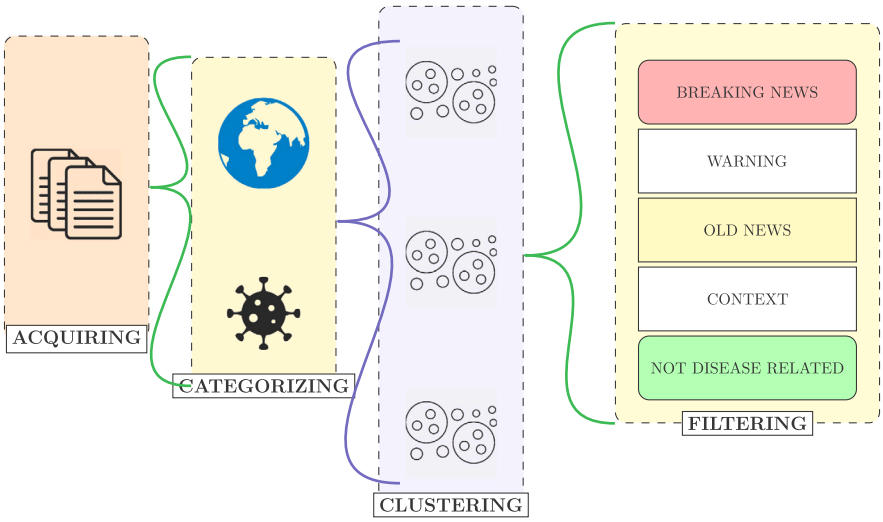


Fig. 2. HealthMap classifier pipeline.

Table 2
Systematic Literature review on current disease surveillance corpus classification systems.

Study	Dataset Origin	Classification Algorithms	Statistical test	Main findings
Current EBS Systems				
GRITS (Huff et al., 2016)	Social media, online news outlets, ProMED-mail reports, blogs	TF-IDF	None	Monitoring digital disease signals for infectious disease threats means that EBS capacity can be extended to areas where public health infrastructure is inadequate. The use of binary relevance algorithm combined with enriched features mined using NLP models improves ensembled logistic regression classifier.
VGCN+BERT (Mutuvi et al., 2020)	ProMed News Articles	BERT, VGCN	None	Models based on both fine-tuned language models and graph convolution networks achieve very good performance on the classification of multilingual epidemiological texts.
DANIEL (Mutuvi et al., 2020b; Sahnoun & Lejeune, 2021)	ProMed News Articles	Language-agnostic text-level features extraction using multinomial naive Bayes, random forest, neural network classifiers and BERT	None	Use of both Open Information Extraction (OIE) and extendend Ontologies.enhances epidemiological text classification and related tasks such as named entity extraction.

2.2. The corpus classification modules of modern EBS systems

In order to improve EBS classification task beyond classical machine learning approaches, new classification systems focusing on using language processing techniques were later introduced. We review them in this section and provide a summary in Table 2.

2.2.1. GRITS classifier

Global Rapid Identification of Threats System for Infectious Diseases in Textual Data Sources (GRITS) Huff et al. (2016) is a biosurveillance system that uses binary relevance method to predict the disease referred to by a body of text. GRITS uses ensemble learning with logistic regression classifiers and having each classifier estimates the probability that a text passage is associated with a given disease. GRITS' classification engine also extracts vector features from textual documents using NLP algorithms and uses these to enrich their features obtained through binary relevance in order to improve classification. A key challenge with GRIT classification approach is that use of NLP extracted vector features combined with ensemble-learnt features does not enrich thematic features enough to improve classification. Our proposed approach advances this aspect by using pre-trained language models to enrich features beyond machine learning techniques such as logistic regression.

2.2.2. DANIEL classifier

The Data Analysis for Information Extraction in any Language (DANIEL) Sahnoun and Lejeune (2021) system is multilingual news surveillance system that mines for journalistic writing style patterns in order to classify an epidemiological document (Mutuvi et al., 2020b). DANIEL system performs classification by tracking key questions about

Table 3
Systematic Literature review on current disease surveillance corpus classification systems focused on Animal Health.

Study	Dataset Origin	Classification Algorithms	Statistical test	Main findings
Animal Health Current EBS Systems				
PADI-Web (Valentin et al., 2021)	Google News	SVM	None	Use of Machine learning algorithms specifically SVM, improves epidemiological document classification as compared to Logistic regression.
APHA (Arguello-Casteleiro et al., 2019)	Google News	Unified Medical Language System (UMLS) Metathesaurus	None	Relevant veterinary medical terms can be automatically identified in the free-text summaries entered by Veterinary Investigation Officers (VIOs) in clinical reports using NLP techniques to mine Metathesaurus.
EpidBioBERT (Menya et al., 2022)	Google News	BERT, BioBERT	None	Use of attention based pre-trained networks, specifically **domain specific language model**, significantly improves epidemiological document classification.

how news is reported and linking this with patterns in news to improve classification (e.g. how the beginning and ending of an epidemiological news article are written). In our proposed approach, we learn such patterns in epidemiological news documents using pre-trained language models. We compare how different language models enrich features and improve classification by setting up experiment involving language models with differing pre-trained strategies and from different domains.

2.3. Contemporary EBS focused on animal health and their corpus classification modules

Finally in this section, we review EBS systems focused on animal health and provide a summary in Table 3.

2.3.1. PADI-Web classifier

PADI-Web is an event based biosurveillance system that monitors the emergence and spread of infectious animal diseases by monitoring online news sources in order to detect and alert on (re)-emerging epizootics (Valentin et al., 2020, 2021). The system has collected over 500,000 news articles since 2016 and has evolved over three versions starting with PADI-Web 1.0 (Arsevska et al., 2018). This early version of PADI-Web was a keyword-based classification-approach system that used a predefined list of disease outbreak keywords and classified input epidemiological corpora based on existence of one or more of these preset keywords as found in the document. This technique combined both rule-based and data-mining approaches to mine for epidemiological keywords over 352 English news articles collected from google news. This version of PADI-Web introduced corpora collection via Really Simple Syndication (RSS) feeds. Google News RSS feeds are mined based on disease names search and terms describing clinical signs of hosts of a given disease. These raw corpora are then cleaned and classified as either relevant or irrelevant. After this process, epidemiological information extraction of disease names, event location, event date, and disease host names are carried out over relevant articles as tagged by the system. A Support Vector Machine (SVM-RBF) engine trained over a set of curated rules is used to achieve the corpora classification process. We capture this pipeline in Fig. 3.

PADI-Web 2.0 extends surveillance from four animal diseases in its initial version to monitoring nine infectious animal disease outbreaks and eight syndromes in five animal hosts. Its classification module uses supervised machine learning techniques to identify relevant corpora. As of its main model, PADI-Web 2.0 converts news corpora to bag-of-words and Term Frequency-Inverse Document Frequency (TF-IDF) features sets (Jones, 1972; Luhn, 1957) and relevant news corpora are further classified into five fine-grained categories namely; *confirmed outbreak, suspected or unknown outbreak, preparedness* and *impact* (Valentin et al., 2020). PADI-Web 3.0 improves the Fig. 4 classification process by introducing sentence classification with bag-of-words sentences representations. This presents a more fine grained approach to improve the overall document classification by highlighting fine-grained epidemiological information such as *risk events, preventive* and *control measures*. A dedicated annotated corpus is built for training purposes (Valentin et al., 2019). A domain-specific biomedical model is built for epidemiological information extraction to replace generic approaches used in prior PADI-Web versions. For further terminology extractions and annotations, both Brat and BioTex (Lossio-Ventura et al., 2016) are incorporated.

2.3.2. EpidBioBERT classifier

This work majorly extends PADI-Web 1.0 (Arsevska et al., 2018), PADI-Web 2.0 (Valentin et al., 2020) and PADI-Web 3.0 (Valentin et al., 2021) introducing enrichment of deep thematic embeddings to improve over PADI-Web infectious animal disease news article classification. EpidBioBERT adopts BioBERT(+PubMed) by Lee et al. (2019) as it is pre-trained biomedical language model and fine tunes it for the task of epidemiological document classification (Menya et al., 2022). EpidBioBERT classifier takes inputs in the form of [CLS]$ThemTok_1^1, \ldots, ThemTok_M^N$[EOS] representing N thematic feature tokens from M sentences in the annotated train corpus. The deep thematic based model then learns a probability distribution over document classes *relevant* and *irrelevant* represented as $\{c_1, c_2\}$ as shown in Fig. 5.

This thematic feature enrichment concept towards improving epidemiological corpora classification focuses on *disease, host, location* and *date* thematic features as found in epidemiological corpora and enriches them using a pre-trained word embeddings approach. Menya et al. (2022) compares builds such embeddings using two approaches; Bag-of-Words and Term-Frequency Inverse Document Frequency (TF-IDF) and trains them over classical classification approaches such as Support Vector Machines (SVM), Long-Short Memory Networks (LSTM) and Bidirectional LSTM (Bi-LSTM) and current BERT model. These classifiers are then benchmarked against EpidBioBERT model. The deep thematic embeddings approach of EpidBioBERT concludes that using a pre-trained biomedical language models such as BioBERT enrich features further thus improving epidemiology document classification.

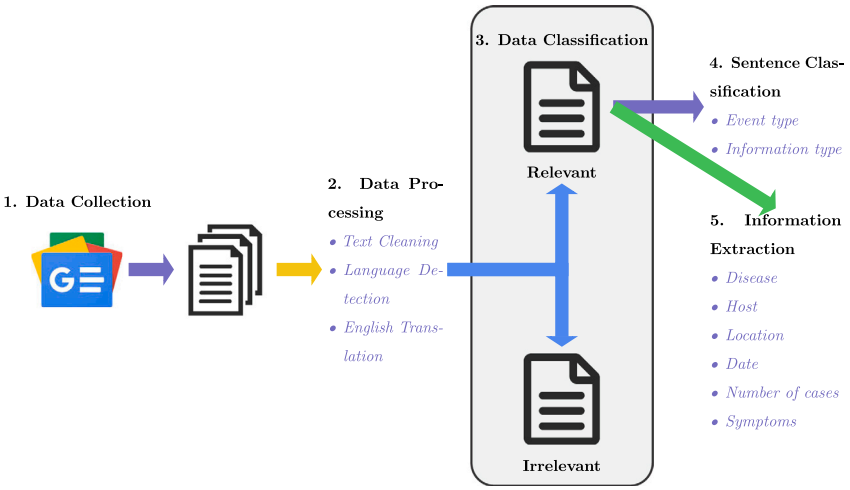


Fig. 3. PADI-Web epidemiology document classifier.

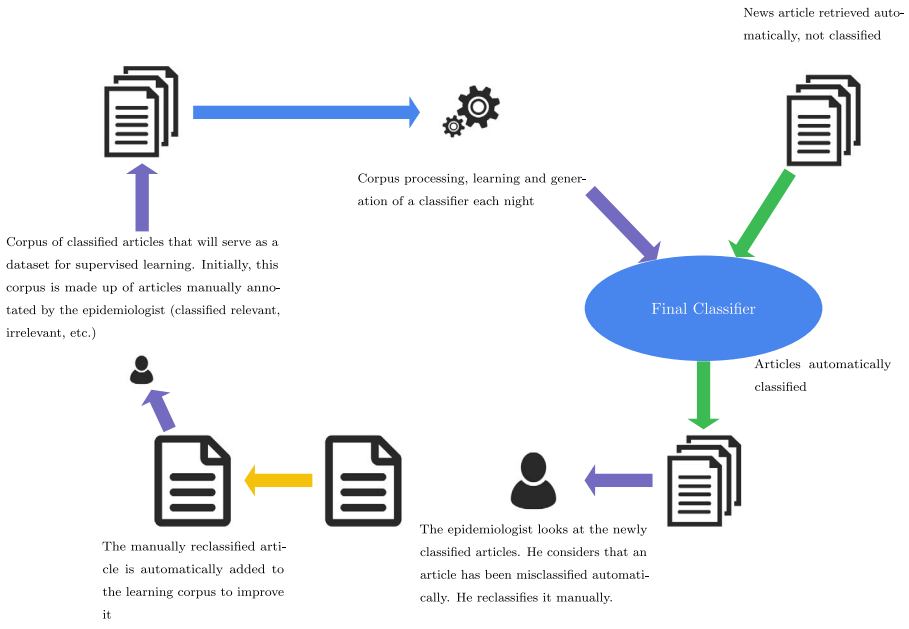


Fig. 4. PADI-Web pipeline flow.

In this paper, we further improve this deep thematic feature approach to epidemiological classification. Our technique experiments enriching of thematic features over SOTA biomedical language models listed in the BLURB leader-board (Gu et al., 2022).

2.4. Challenges in modern EBS classification models

Most epidemiological classification models rely on conventional supervised learning technique which requires human-expertly labeled datasets which are time consuming and expensive to achieve limiting the amount of data used in training (Delon et al., 2024). For instance, (Menya et al., 2022; Mutuvi et al., 2020; Valentin et al., 2019) classifiers are trained using supervised learning. Unsupervised learning is another technique that can be applied as a potential area to improve the problem of domain specific text classification and entity recognition since it requires unlabeled datasets which is always largely available. For example, G. et al. (2023), uses unsupervised learning approach to identify thematic named entities in agricultural domain. However, this is still a challenging area due to the unstructured nature of textual data. These challenges calls for other current learning techniques such as weak supervised learning to be explored.

On the other hand, deep learning based approaches have been shown to improve text classification (Bai et al., 2018; Kowsari et al., 2018; Li et al., 2021). These approaches work by applying a pre-trained large language model that is fine tuned to a specific text classification task. This approach has been applied with success in general English articles but remains challenging in domain specific tasks. Jiang et al. (2023) for example, explores novel fine tuning techniques for text classification in plant health domain finding that conventional fine tuning does not always work in domain specific text classification. Other related tasks have also employed deep learning with fine tuning (Lample et al., 2016; Mutuvi et al., 2020), however most of these approaches focus on enriching general named entities and events. In addition, for most deep learning approaches, even though they improve classification, model explanation is not usually provided making it difficult to understand how the classifier improves internally.

In this paper, we propose a deep learning based model using pre-trained language modeling. We experiment our proposed model on a large dataset using weak supervision approach and study the underlying thematic feature behaviors in how they improve epidemiological classification using explainable AI techniques. Our large train set enables our model to generalize well on both precision and recall metrics

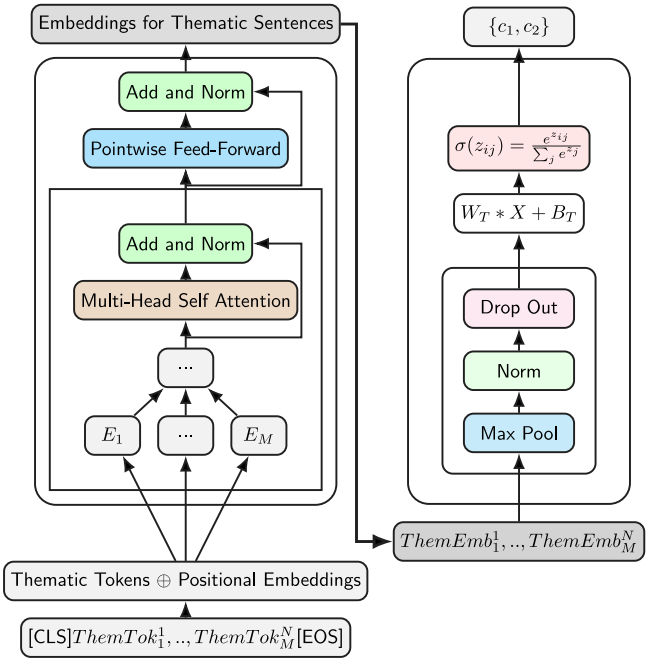


Fig. 5. EpidBioBERT epidemiology document classifier.

and uncovers novel strategies for epidemiological surveillance shown through in-depth study of thematic feature behavior that we present as part of our model explainability.

3. Epidemiological thematic features

3.1. Definition of thematic features

In our approach, epidemiological corpora classification is improved with enrichment of epidemiological thematic features. These features include; spatio-temporal, biomedical, and named entities as captured in Fig. 6. **Spatio-temporal** entities consists of location, date and time of a reported acute public or animal health event. In this context, location is the geographical region where the case is originating such as a country, state or city, where as date and time are the period of reported disease outbreaks such as days of the week and seasons of the year e.g. *last Tuesday* or *early summer*.

Biomedical entities are names of outbreak diseases, names of disease hosts and disease symptoms or syndrome. Disease name examples include *Avian influenza* and *COVID-19* while disease hosts include names of agents that are infected by a disease such as *pigs* or *chicken*. Finally, symptoms include names of conditions caused by a disease such as *fever, vomiting* or *death*.

Lastly, **Named** entities are made up of conventional named entities. Categories of named entities includes names or mentions of key persons, organizations, and number of reported cases. For instance, names and mentions of entities such as *WHO director general* or *the cabinet secretary of the ministry of health*, and organizations e.g. *the World Food Program* are of interest in this study.

3.2. Classical language processing methods for thematic feature representation

In this subsection we look at the classical language processing approach to thematic feature representation. We start by a formal definition to thematic features followed by various techniques for representing thematic features namely; bag of words and TF-IDF approaches.

3.2.1. Formal definition of thematic features

Given an epidemiological corpus set of $size = N$ news articles denoted as $D = \{d_1, \dots, d_n\}$ and news article $d_j \in D$ containing I epidemiological thematic features in set T denoted as $T = \{t_1, t_2, \dots, t_i\}$ we define a *term-document* matrix $X_{n,i}$

$$X = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \dots & t_i \end{array} \\ \begin{bmatrix} f(t_1,d_1) & f(t_2,d_1) & \dots & f(t_i,d_1) \\ f(t_1,d_2) & f(t_2,d_2) & \dots & f(t_i,d_2) \\ \vdots & \ddots & \vdots & \\ f(t_1,d_n) & f(t_2,d_n) & \dots & f(t_i,d_n) \end{bmatrix} \end{array}$$

From X^T we get $t_i = [f(t_i,d_1), f(t_i,d_2), \dots, f(t_i,d_n)]$ which we represent as thematic vector $\vec{w}_i$ of thematic feature t_i represented as $\vec{w}_i \in \mathbb{R}^{1*N}$ where N is the number of documents in our corpus set. Function $f(t_i,d_n)$ can be defined in several ways discussed below in Sections 3.2.2, 3.2.3 and 3.3.1.

Epidemiological corpus set D is divided into train, test and validation sets Γ, τ, ν respectively. Given the set of K features where $K \subset T$ and $[t_1, t_2, \dots, t_k] \in d_j$ having $d_j \in \Gamma$, and epidemiological target classes $C = \{c_1, c_2\}$ where $c_1 = relevant, c_2 = irrelevant$, we learn the conditional probability distribution $p(c_i|d_j)$ where $c_i \in C$ and $d_j \in \Gamma$ to obtain vector $\vec{y} = [p(c_i|d_1), p(c_i|d_2), \dots, p(c_i|d_n)]$ where $\vec{y}^T$ becomes the target vector.

3.2.2. Bag of words thematic features

In the bag of words (BOW) approach, every thematic feature is taken in isolation ignoring its relation with other features and their respective positions in an epidemiological corpus. In other words, BOW assumes that every token feature is not related to other tokens in the epidemiological domain. Following this definition, function $f(t_i,d_n) \in X$ is defined as

$$f(t_i,d_n) = \begin{cases} 1, & \text{for } t_i \in d_n \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Eq. (1) follows that thematic feature $t_i \in X^T$ is represented with a one hot vector $\vec{w}_i \in \{\mathbb{W} := \{0,1\}\}^{1*N}$. The feature matrix X^T is thus a high dimensional sparse matrix since most terms t_i will occur in fewer documents. This sparse matrix can be decomposed into a dense matrix that is more efficient and richer using singular vector decomposition (SVD) or principal component analysis (PCA) techniques. Menya et al. (2022) experiments over SVD and PCA epidemiological thematic features converting t_i from $\{\vec{w}_i \in \mathbb{W} := \{0,1\}\}^{1*N}$ to $\vec{w}_i \in \mathbb{R}^{1*N}$ dense representations. Though useful, these types of epidemiological thematic features are not rich enough to improve epidemic intelligence corpora classification mainly due to the BOW assumption aforementioned which leads to poor classifier performance.

3.2.3. TF-IDF thematic features

Term-Frequency Inverse Document Frequency (TF-IDF) approach introduced by Jones (1972), Luhn (1957), counters the BOW challenge introduced in Section 3.2.2. Instead of generating a one hot representation of thematic feature t_i, TF-IDF computes function $f(t_i,d_n) \in X$ as:

$$\begin{aligned} &f(t_i,d_n) = tf_{t_i,d_n} * idf_{t_i}, \\ &\text{Where } tf_{t_i,d_n} = \text{frequency of term } t_i \text{ in document } d_n, \\ &idf_{t_i} = \ln(\tfrac{N}{df(t_i)}), \\ &\text{Where N = number of documents, } df(t_i) = \text{count documents} \\ &\text{out of N containing } t_i \end{aligned} \tag{2}$$

Both Valentin et al. (2020) and Menya et al. (2022) experimented over TF-IDF thematic features. Menya et al. (2022) builds smoothed L2-norm TF-IDF feature sets which avoids the zero word vectors in the

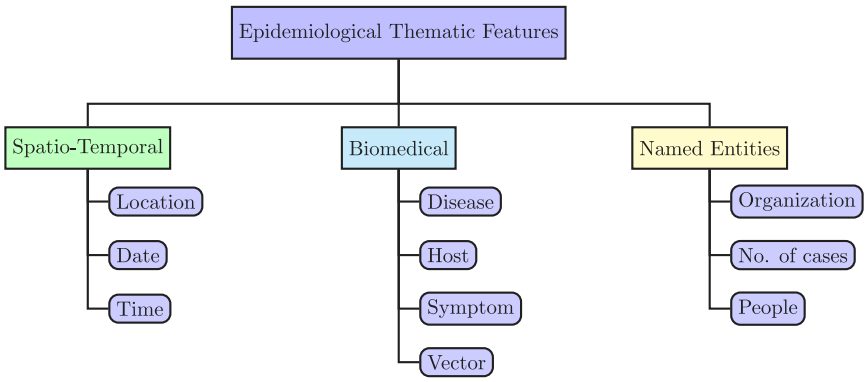


Fig. 6. Thematic feature classifications.

case of $t_i \in \forall n, n \in D$ where the term occurs in all documents causing $idf_{t_i} = 0$. In smoothed TF-IDF version, the idf formula is updated to:

$$idf_{t_i} = \ln(\frac{N+1}{df(t_i)+1}) + 1 \tag{3}$$

while L2-norm is used to normalize the $\vec{w}_i$ of t_i using the euclidean formula $L_2(\vec{w}) = \|w\|_2$. The resulting thematic feature is a normalized dense vector in the $\{\mathbb{R} := \{0,1\}\}^{1*N}$ space. TF-IDF epidemiological thematic features are richer than BOW thematic features since they relate every token to a given document's context which gives much weight to rare tokens thus improving discrimination between relevant and irrelevant classes.

3.3. Deep learning methods to thematic feature representation

In this subsection we look at the deep learning language processing approach to improving thematic feature representation beyond classical approaches presented in the previous sections. We start by presenting contextualized word embeddings as a model for representing enriched thematic features, followed by a presentation on pre-trained language modeling approach to learning thematic feature embeddings.

3.3.1. Deep contextualized thematic features

In the deep learning approach to representing thematic features, instead of *term-document matrix* $X_{n,i}$ as seen in Section 3.2.1 we shift to **term-term matrix** $M_{i,v}$ for I terms and V context-set which is the vocabulary set defined as the set of all unique words in all N documents in our epidemiological corpus.

$$M = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_i \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \dots & t_v \end{array} \\ \begin{bmatrix} f(t_1,t_1) & f(t_2,t_1) & \dots & f(t_i,t_1) \\ f(t_1,t_2) & f(t_2,t_2) & \dots & f(t_i,t_2) \\ \vdots & \ddots & \vdots & \\ f(t_i,t_1) & f(t_i,t_2) & \dots & f(t_i,t_v) \end{bmatrix} \end{array}$$

Function $f(t_i, t_v)$ is computed as:

$$f(t_i, t_v) = p(t_i|t_v), \tag{4}$$

Where t_v is the context word

Terming the vocabulary set V as context, we learn the conditional probability of term t_i given context t_v. This property leads to similar words having similar vectors since similar words are known to co-occur in similar contexts (Firth, 1957; Harris, 1954; Joos, 1950). Thematic feature t_i is now represented as $t_i = [p(t_i,t_1), f(t_i,t_2), \dots, f(t_i,t_v)]$ which is a vector that encodes information on how t_i relates to all other terms in T. Various techniques have been introduced of learning this complicated distribution. For instance, Mikolov et al. (2013) learns:

$$p(t_v|t_i) = \frac{e^{(t_v \cdot t_i)}}{\sum_{s \in |V|} e^{(t_s \cdot t_i)}} \tag{5}$$

However this approach requires computing the dot product with every other word in the vocabulary set making it computationally expensive to compute the denominator. In addition, computing $p(t_i|t_v)$ presents a challenge similar to BOW assumption since language is a sequence that unfolds in time thus any particular term should be related to every prior term that occurs before it in an epidemiological sentence. Thus considering document $d_j \in D$ which contains P epidemiological sentences represented as set $S = \{s_1, s_2, \dots, s_p\}$ with $s_p = \{t_1, t_2, \dots, t_i\}$ we learn $p(t_i|t_{i-1}, t_{i-2}, \dots, t_1)$ (also $p(t_i|t_{i+1}, t_{i+2}, \dots, t_v)$ in the Bi-directional approach) where t_i is our epidemiological thematic feature and set $\{t_{i-1}, t_{i-2}, \dots, t_1\}$ is the context of t_i. Considering this complex contextualization in learning word vectors for our thematic features leads to a language modeling approach. Thematic features learnt this way are far richer in context as opposed to thematic feature learnt through the non-contextualized approaches such as BOW and TF-IDF.

3.3.2. The language modeling approach to thematic features

The language modeling approach of learning contextualized epidemiological thematic features leads to context-rich features. However, these complicated contexts are hard to manage and learn from. To this regard, several context management approaches have been introduced to learning word embeddings; for instance the use of *neural networks* (Bojanowski et al., 2016; Pennington et al., 2014), *recurrent neural networks* (RNN) (Peters et al., 2018), *Long-Short Term Memory Networks* (LSTMs) and *Attention based Encoder–Decoder Networks* (Vaswani et al., 2017). In our approach, we learn epidemiological thematic embeddings using the Attention Networks approach in which we train an Encoder–Decoder model to learn the most significant dimensions of the context $\{t_{i-1}, t_{i-2}, \dots, t_1\}$ that the model pays attention to with respect to how much these terms contributes to a rich representation of feature t_i.

$$t_i = \sum_{m \le i} \alpha_{im} \cdot t_m, \forall m \le i$$

Where α_{im} represents the weight of terms t_m in representing term t_i for example a cosine similarity function between t_m and t_i (6)

The attention technique learns parameter α using a set of trainable weights namely *query*, *key* and *value* represented as W^Q, W^K, W^V (Vaswani et al., 2017). Thus Eq. (6) becomes:

$$t_i = (W^Q t_i \cdot W^K t_i \cdot W^V t_i) + (W^Q t_i \cdot W^K t_{i-1} \cdot W^V t_{i-1}) + \dots + (W^Q t_i \cdot W^K t_1 \cdot W^V t_1) \tag{7}$$

Where $\alpha_{im} = W^Q t_i \cdot W^K t_m$

The term α is then normalized using softmax and computations done in matrix form, thus we set up an encoder–decoder attention network

to learn epidemiological thematic feature representations of M as:

$$AttentionScore(W^Q M, W^K M, W^V M) = \sigma(\frac{W^Q M, W^K M}{\sqrt{K_{dim}}})W^V M$$

Where σ is the softmax equation $\sigma(x) = \frac{e^x}{\sum_{\forall i} e^{x_i}}$

and we normalize with K_{dim} to avoid exploding/vanishing gradients

(8)

In these approaches, we end up with a rich and dense vector representations of our thematic features. However, two more challenges exists in this attention based approach of enriching thematic features. First, the approach is computationally expensive to train from scratch thus we use pre-trained language model with a transfer learning approach. Secondly, traditional transfer learning approaches are largely pre-trained on general English corpora thus using this approach will mainly only enrich terms mostly used in English (such as named entities). Such an approach does not significantly enrich thematic terms such as biomedical and spatio-temporal thematic features since they fail to exist in pre-training vocabulary. We thus introduce approaches to enrich epidemiological thematic features using novel techniques that counter these challenges. Our approach proposes a mixed-domain language model for enriching epidemiological thematic features.

3.3.3. Deep contextualized epidemiological thematic features

In order to enrich our epidemiological thematic features, our approach adopts a pre-trained biomedical language model with epidemiological fine tuning. This choice is inspired by the close link between biomedical domain and epidemiological domain in the study of infectious diseases and how they spreads. As an improvement to Menya et al. (2022) who introduced the (Lee et al., 2019) BioBERT architecture for epidemiological corpora classification task, we achieve our architecture using BioELECTRA (Kanakarajan et al., 2021).

BioELECTRA's underlying architecture improves over BERT based language models by introducing *replaced token prediction* technique that counters BERT's masked language model technique allowing BioELECTRA to learn from all tokens in a train set. In addition, BioBERT's Weakness of word-piece tokenization hampers generalization in downstream biomedical tasks making it perform lower than BioELECTRA (Gu et al., 2022). BioELECTRA ranks higher than BioBERT in the BLURB leaderboard's biomedical tasks including named entity recognition which is strongly related to our task goal. In Section 5 we compare our approach with other SOTA language models namely BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Tinn et al., 2021) and EpidBioBERT (Menya et al., 2022) over the epidemiological corpora classification task.

4. Experimental setup

In this section we discuss our corpus set and its contents, how it was generated and prepared. We also specifically discuss our thematic feature corpus set and its statistical distribution. Finally we outline our model configurations and hyperparameter settings.

4.1. PADI-Web corpus set

Our experimental data is made up of three sets of corpora. First we prepare PADI-Web$_{gold}$ which consists of 800 human-expert labeled corpora which is balanced between relevant and irrelevant articles (Rabatel et al., 2017). This first set of corpus is used for strong supervision testing of our model. Secondly, we derive our main corpora from PADI-Web dataset by Menya et al. (2023) for a weak supervised learning approach. Weak supervised learning approach provides large corpus for improved training and testing beyond gold labeled dataset (Mutuvi, 2022). Our PADI-Web dataset contains automatically annotated news articles collected from google news via RSS feeds. In addition, this collected corpus set contains relevant (consists of animal epizootics news articles) and irrelevant (articles that mentions other general topics or those simply related to epidemiology) article labels. These labels are automatically assigned by PADI-Web system (Valentin et al., 2020). To minimize labeling noise from our machine generated data, we use a *human annotator intervention* approach to manually walk-through the data. This intervention leads to text cleaning approaches including correctly re-classifying the small portion of miss-classified corpus and the elimination of outlier articles. We thus exclude articles that are either too short or too long within a given threshold. Outliers are more prone to causing miss-classification errors (Valentin et al., 2020).

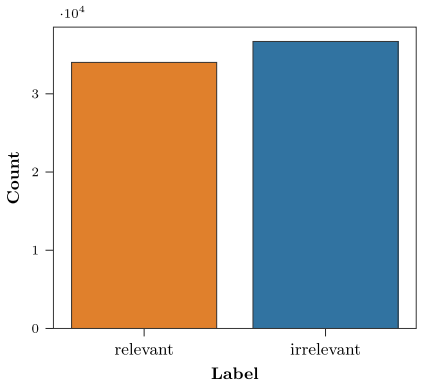


Fig. 7. Distribution of relevant vs. irrelevant epidemiological news articles in our collected PADI-Web dataset.

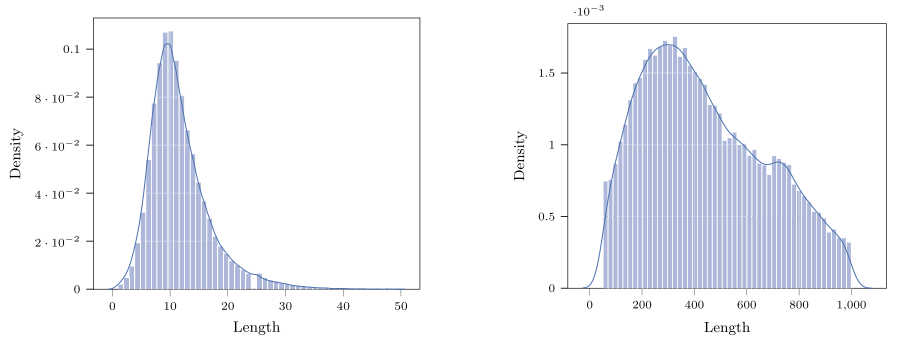
In the next step of data preparation on our main corpus set, we apply a two step procedure. We first clean the raw PADI-Web articles (removing hyperlinks, lowercasing) and then perform tokenization using pre-trained BioELECTRA tokenizer. This main corpus set consists of 70,707 news articles with a fair relevant to irrelevant distribution ratio of $48:52$ translating to 34,015 relevant and 36,692 irrelevant as shown in Fig. 7. The aim of having a corpus set of this size is to improve our model training and testing beyond the 800 class-imbalanced news articles used to train and test EpidBioBERT (Menya et al., 2022). EpidBioELECTRA is thus trained on almost 100 times the data of EpidBioBERT. Our prepared corpus set is then subdivided into train (60%), test (20%) and validation (20%) sets for experimental purposes as summarized in Table 4.

In addition to the above steps, from our main corpus set we further prepare a new data segment named PADI-Web$_{titles}$ consisting of PADI-Web article titles with their associated article labels. This new data consists of shorter corpora length since the average length of news titles measured from these set is approximately 11 words Fig. 8(a) compared to the average length of article content Fig. 8(b). We experiment over PADI-Web$_{titles}$ in Section 5.1 and find that it is more efficient to train from this set though trained models though the resulting classifier suffers in performance compared to training with longer articles. For further experimentation's based on corpus length, we prepare PADI-Web$_{short}$ with maximum length of 128 words per article, PADI-Web$_{long}$ of length 256 and PADI-Web$_{XL}$ with 512 words per article as informed by the average size of most news articles in our original PADI-Web corpus set Fig. 8(b).

The last portion of our data is termed PADI-Web$_{annotated}$ which is collected from the annotated PADI-Web corpus containing labeling information and meta-data about thematic features found in PADI-Web (Menya et al., 2023). We use this second data portion to study our thematic features with their labels, distributions and influence that they have in our classifier using explainable AI techniques in Section 5.5. We also use PADI-Web$_{annotated}$ to compute thematic features frequency distribution in Section 4.2 below. Our model, EpidBioELECTRA uses the two prepared portions of data as shown in Fig. 9.

Table 4
Overview of PADI-Web dataset showing frequency, train and test percentages (Train : Test), maximum set corpus length (Length) and percentages of relevant versus irrelevant articles (Rel : Irrel).

Dataset	*Frequency*	*Train : Test*	*Length*	*Rel : Irrel*
PADI-Web$_{gold}$	800	80 : 20	–	49 : 51
Titles				
PADI-Web$_{titles}$	70,707	60 : 40	16	48 : 52
Length				
PADI-Web$_{short}$	70,707	60 : 40	128	48 : 52
PADI-Web$_{long}$	70,707	60 : 40	256	48 : 52
PADI-Web$_{XL}$	70,707	60 : 40	512	48 : 52
Thematic Feature				
PADI-Web$_{annotated}$	70,400	–	–	100 : 0



(a) Average title length in PADI-Web$_{titles}$ (b) Average corpus length in PADI-Web

Fig. 8. On the left the average number of tokens per title in epidemiological news article in our collected PADI-Web dataset. On the right the average number of tokens per document in the same dataset.

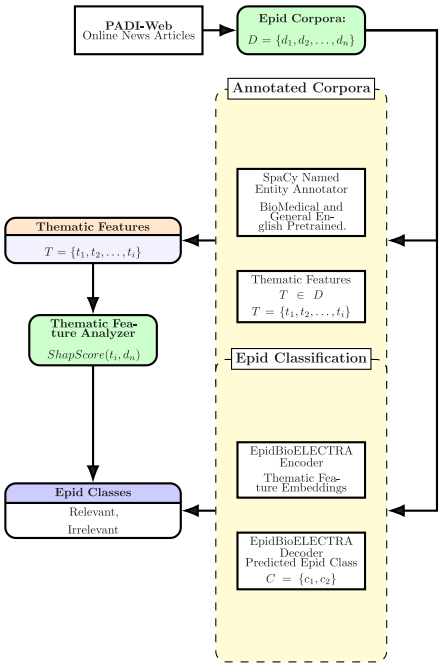


Fig. 9. EpidBioELECTRA architecture flow.

4.2. Thematic feature distribution in PADI-Web

Our second portion of data, PADI-Web$_{annotated}$, is richly annotated with epidemiological information as contained in PADI-Web corpus with information on document id, sentences and positions where a given thematic feature appears. PADI-Web$_{annotated}$ contains 70,400 thematic features and their related information, out of the 70,400 we count 9073 uniquely mentioned thematic features. From PADI-Web$_{annotated}$ we compute some useful statistics about PADI-Web dataset. We generate both bar graph and sankey plots to visualize thematic feature frequencies from both relevant and irrelevant classes of PADI-Web$_{annotated}$. From Fig. 10, we observe that PADI-Web corpus set uniquely mentions 202 diseases, 129 hosts, 2763 named locations and 35 symptoms.

In Fig. 11 we generate a sankey plot of three of the top thematic features of each type (disease, host, keyword, location, symptom, date and time) showing how they rank by frequency. We note that PADI-Web corpus contains *african swine fever* as the top mentioned disease in most of its epidemiological news articles, while *pig*, *cases* and *China* rank as the top host, keyword and location thematic features respectively. Likewise, *fever* and *2021* are the most mentioned symptom and date respectively.

4.3. Model configurations and hyperparameter settings

We perform fine tuning on our model maintaining the hyperparameters experimented by Menya et al. (2022) and selected through grid search, we thus have hidden embedding size of 768, 12 Attention Heads and 12 Transformer blocks for our model and baselines to maintain a fair competition among models. However, different from Menya et al. (2022), we increase our batch size and sequence length depending on the type of data under experimentation and after performing grid search on our held out corpus. For PADI-Web$_{titles}$ we set batch size to 64 and sequence length to 32. For the rest of corpora PADI-Web$_{short}$, PADI-Web$_{long}$ and PADI-Web$_{XL}$ we set batch size to 32 and sequence length corresponding to limits set in 4.1. We then experiment using cross-validated models with each running for 50 epochs and we obtain the averaged metrics. For our decoder model we experiment with dropout rates of 0.1, 0.2 and 0.3 to control the model's overfitting tendency. For

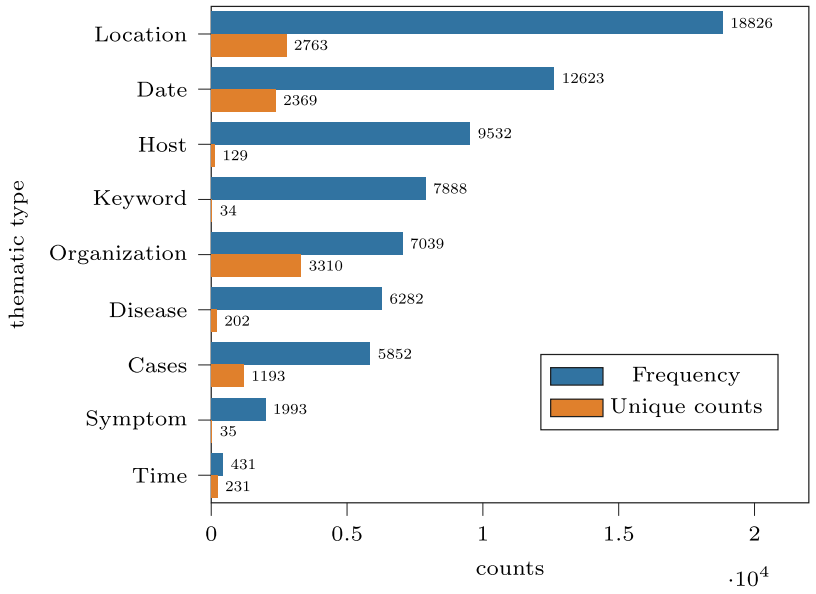


Fig. 10. Frequencies of mentioned thematic features in PADI-Web. Unique counts are recorded as we count every individual feature once without repetition.

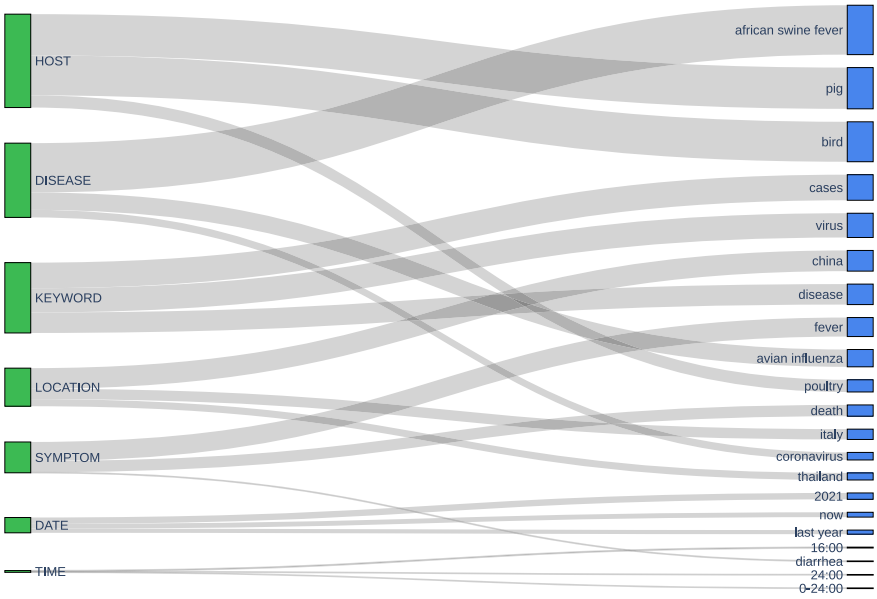


Fig. 11. Sankey Plot of top thematic features (on the right) by frequency (captured by arc thickness) and their general types (on the left) computed from PADI-Web annotations. We observe Disease, Host, Keyword and Location as popular thematic features while Symptom and Date as least popular.

the optimizer, we maintain the AdamW optimizer, which has been reported as the best-performing hyperparameter in previous studies, with decoupled weight decay setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$ (Loshchilov & Hutter, 2019). We also set ϵ = 1e-8 and weight decay = 0.01 for the optimizer. In addition, we also set small initial learning rates of 1e-5 and 2e-5 balancing this with our cross-validation epochs to favor our fine tuning approach following (Ruder, 2021).

5. Results and ablations

In this section, we present experimentation results of EpidBioELECTRA compared with other pre-trained baselines experimented on PADI-Web$_{titles}$, PADI-Web$_{short}$, PADI-Web$_{long}$ and PADI-Web$_{XL}$ corpus sets. We first discuss overall results for all competitive models and we gold test the most competitive models on PADI-Web$_{gold}$ dataset. Later, we present and discuss ablation findings on understanding the behaviors of these classifiers and how thematic features influence classification improvement.

5.1. Experimentation metrics

In our first empirical step, we apply an extrinsic evaluation approach where we focus on measuring and comparing model classification skills as captured by how they balance between recall and precision. Recall metric is also known as a models' sensitivity while precision is known as a models' positive predictive value (PPV). In the context of epidemiology surveillance, Sensitivity is defined as the ability of an EBS system to detect health risks while PPV is defined

Algorithm 1: EpidBioELECTRA classifier algorithm

Data: Epidemiological news documents $D = \{d_1, ..., d_n\}$ with class labels $D = \{l_1, ..., l_n\}$
Result: Determination of whether d_n is relevant or irrelevant

$T \leftarrow \{disease, host, location, date, time\}$;
$D_{annotated} \leftarrow \{\}$;
$D_{general} \leftarrow \{\}$;
for $d_n \in D$ **do**
  for *each word* $w_i \in d_n$ **do**
    if $w_i \in T$ **then**
      $w_i \leftarrow thematic$;
      $D_{annotated} \leftarrow w_i$; /* This is a thematic feature */
    else
      $D_{general} \leftarrow w_i$; /* This is a general English word */
    end
  end
end
for $d_a \in D_{annotated}$ **do**
  for $t_i \in d_a$ **do**
    $W_{t_i} \leftarrow LM$;
    $d_e m \leftarrow W_{t_i}$;
    $y \leftarrow wx + b$; /* Compute class prediction */
    $L \leftarrow -(cln(y) + (1-y)ln(1-y))$; /* Compute model loss */
    $W \leftarrow w - \alpha * \frac{\partial L}{\partial W}$; /* Update weights for better prediction */
  end
end

as the probability of a raw signal detected by an EBS to correspond to a genuine health risk (WHO, 2014, chap. 6). We perform evaluation by comparing precision–recall curve (PR) as well as receiver operating characteristic curve (ROC) of EpidBioELECTRA against competitive classifiers in their skill of performing epidemiological document classification by discriminating between relevant and irrelevant signals.

Precision and Recall performance measures are inspired by the expected practicality of EBS systems. An effective EBS system must report epidemiological events before official sources detect such cases. By taking this approach, we are measuring the added value of an EBS system over conventional methods of epidemic surveillance. A clear balance has to be struck between timeliness and accuracy of digital epidemic surveillance classifiers. WHO (2014, chap. 4) advise is to prioritize sensitivity (recall) above PPV (precision) when detecting for emerging and novel epidemics while evoking a vice-versa prioritization when detecting for common re-emerging diseases. In-line with this, we monitor F_1 scores of EpidBioELECTRA and competitive classifiers in their ability to balance sensitivity and PPV as captured by F_1 score. These monitored F_1 scores inform our experimentation on how skilled a model is in discriminating between relevant and irrelevant epidemiological corpora.

We compute Almost Stochastic Order (ASO) ϵ_{min} scores (Del Barrio et al., 2018; Dror et al., 2019) of EpidBioELECTRA versus all the competitive models using the implementation of Ulmer et al. (2022) and we present these results in Section 5.4. These ASO ϵ_{min} scores quantifies the level of confidence in the difference among all competitive models. ASO computes a specific metric that informs how far a given algorithm is from being significantly better than another. When ASO $\epsilon_{min} = 0$ then it means the algorithm is stochastically dominant over the comparative algorithm, while ASO $\epsilon_{min} = 0.5$ stochastic is undefined. On the other hand, when ASO $\epsilon_{min} = 1$ then the algorithm in question is considered not stochastically dominant to the competitive algorithm.

5.2. Experimentation results

We present our first experimental findings in Table 5. In this table, we note that EpidBioELECTRA performs better than competitor classifiers on F_1 score in both PADI-Web$_{titles}$ and PADI-Web$_{short}$ datasets also outperforming competitor models in recall score in PADI-Web$_{short}$. Performance improves almost ten-fold in PADI-Web$_{short}$ which has longer context compared to PADI-Web$_{titles}$. This improvement is recorded by all competitive models except the baseline models BioELECTRA and BERT though we note that BioELECTRA makes a slight improvement above BERT on F_1 score. We note that increasing the number of words (more sentences) in the train set benefits the performance of all models. This performance improvement can be attributed to the increases in context on which the classifiers make their decision. This context is much longer in long documents compared to short ones. However, though the improvement is true on all models, it is EpidBioELECTRA that records the highest improvement of +7.79% points on F_1 score. This EpidBioELECTRA's improvement is +0.76% higher than EpidBioBERT's which comes second in PADI-Web$_{short}$ dataset.

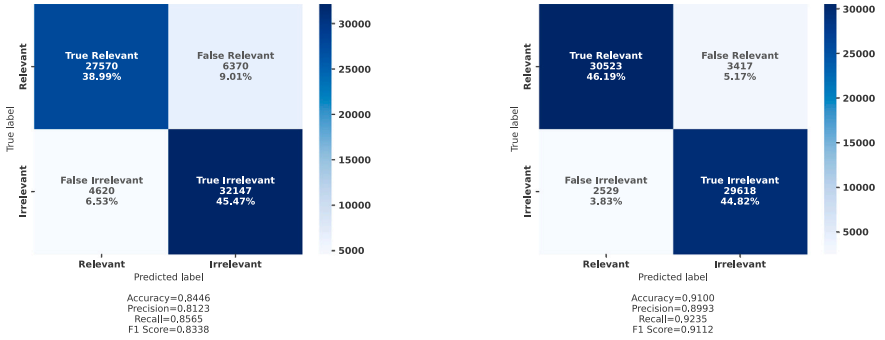
However, still in Table 5, we can see that BioELECTRA, a model that is not fine tuned on epidemiological classification task, records the highest recall values in both datasets with the none fine tuned BERT coming in second. This means that these two none fine tuned models, perform best at avoiding false negatives thus they do well on sensitivity. On the contrary, the two baselines perform the poorest at committing false positives errors thus have the lowest precision values (PPV) meaning they are more prone to sounding false alarms on acute health events. This underlies the key goal of fine tuning EpidBioELECTRA which is to balance the model's sensitivity and PPV as normalized by having a better F_1 score thus avoiding both missing out on epidemiological emergencies and triggering of false alarms (see Figs. 12–14). We also note that PubMedBERT, a biomedical language model, is also competitive in this task as it only falls short by −0.81% of EpidBioELECTRA on F_1 score.

Following the results recorded in Table 5, we make a critical observation that; in enriching features for effective epidemiology surveillance three strategies stand out. First the data used in pre-training the language model, secondly, the pre-training strategy and finally the fine-tuning technique used contribute a lot to a model's performance. We track the pre-trained data used in all our competitive models in Table 6. From this table we note that, building disease surveillance system (such as epidemiology text classifier in our case) works best with language models trained using biomedical corpora such as BioELECTRA and PubMedBERT followed by fine tuning them with epidemiological data. Models pre-trained on other domain corpora, such as SciBERT and ClinicalBERT, tend to suffer in epidemiological task performance. These pre-training strategies tend to enable the base model to learn morphological, syntactical and semantic information from the pre-trained corpus and transfers these knowledge to downstream tasked fine-tuned models. Our fine-tuned model thus learns domain specific thematic features such as names of diseases, hosts and symptoms from the biomedical corpora which is a major contributor to enriching our classification. We study these aspects further in Section 5.5.

Consequent to our observations (in EpidBioELECTRA's context) it follows that it was a benefiting empirical step to use BioELECTRA base model since epidemiological news sources such as PADI-Web corpus have a lot of biomedical information vis-à-vis normal English information. The data used in pre-training BioELECTRA i.e. PubMed-Abstracts, and PMC Fulltexts, benefits our downstream classification process by reducing cases of out-of-vocabulary (OOV) instances. OOV is where many words in the train set are not found in the underlying model's vocabulary thus they are assigned a shared none-discriminative embedding. Also we note that deep Event Based Surveillance systems

Table 5
Results Table comparing EpidBioELECTRA performance against the baseline models in PADI-Web$_{titles}$ and PADI-Web$_{short}$ datasets with best scores in **Bold**. Prec%, Rec% and F_1% refer to precision percentage, recall percentage and F_1 score percentage respectively.

	PADI-Web$_{titles}$			PADI-Web$_{short}$		
Model	**Prec%**	**Rec%**	F_1%	**Prec%**	**Rec%**	F_1%
PubMedBERT	**82.21**	84.12	83.15	89.55	91.12	90.31
BioELECTRA	48.58	**99.64**	65.32	55.83	**99.92**	71.64
BERT	50.53	98.67	66.83	55.85	97.86	71.12
SciBERT	81.7	84.6	83.0	89.71	90.2	89.95
EpidBioBERT	81.04	85.32	83.08	89.71	91.11	90.36
EpidBioELECTRA	81.23	85.65	**83.33**	**89.93**	92.35	**91.12**
ClinicalBERT	79.94	84.85	82.28	88.2	89.0	88.59



(a) Confusion matrix calculated over PADI-Web$_{title}$ (b) Confusion matrix calculated over PADI-Web$_{short}$

Fig. 12. EpidBioELECTRA's confusion matrices.

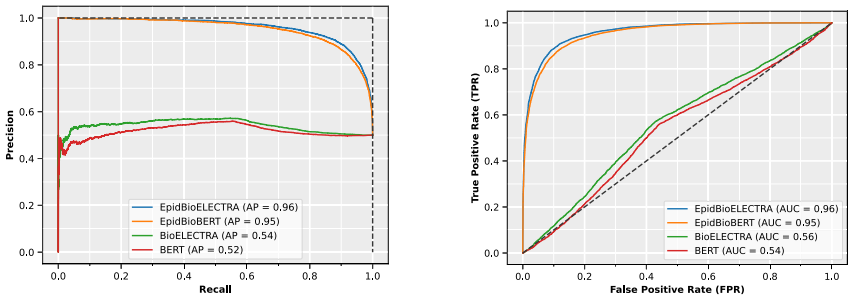
Table 6
Corpora used for pre-training strategies of competitive models. GenEnglish, BioMed, Epid, Clinical, CompSci refer to general Eglish domain, biomedical domain, epidemiological domain, clinical domain and computer science domain respectively.

Model	Pre-trained Corpus Category ($GenEnglish$[1] $BioMed$[2], $Epid$[3], $Clinical$[4], $CompSci$[5])	Corpora
Domain Specific		
PubMedBERT (Abstracts)	2	PubMedAbstracts
PubMedBERT (Abstracts+PMC)	2	PubMedAbstracts, PMC Fulltexts
BioELECTRA	2	PubMedAbstracts, PMC Fulltexts
General Purpose		
BERT	1	English Wikipedia, BookCorpus
Mixed Domain		
SciBERT	2, 5	SciERC, JNLPBA, GENIA, SciCite, BC5CDR, NCBI-disease, EBM-NLP, ChemProt, Paper Field, ACL-ARC
Continual Pre-trained		
EpidBioBERT	1, 2, 3	Wikipedia, BookCorpus, PubMed Central, PADI-Web
EpidBioELECTRA	1, 2, 3	Wikipedia, BookCorpus, PubMed Central, PADI-Web
ClinicalBERT	1, 4	Wikipedia, BookCorpus, MIMIC III

tend to work best with domain specific language models as opposed to general purpose language models such as BERT. However we observe that this domain specific language model has to be fine tuned further with a continual pre-training approach. Such an approach means that we initiate the fine-tuning step to pick up from where the pre-training step left of and we train further to make our model skilled in a related downstream task. This is the approach we take with EpidBioELECTRA.

This observation on choosing a suitable base model is quite interesting since it seems to underscore that though related, epidemiology surveillance is different from biomedical surveillance given the nature of train datasets. Those of epidemiology surveillance blend a mixture of general English and biomedical terms for example epidemiological news articles contains reporting of disease outbreaks in a news reading context. Models based on none biomedical training corpora such as ClinicalBERT perform below the rest on epidemiological classification task.

To understand how longer context improves epidemiological classification, we train and test all models on two more datasets; PADI-Web$_{long}$ and PADI-Web$_{XL}$, and we present results in Table 7. We observe from this table that EpidBioELECTRA benefits the most from this long corpus training as recorded in its +1.69% (in PADI-Web$_{long}$) and +2.34% (PADI-Web$_{XL}$) improvement in F_1 score from PADI-Web$_{short}$. EpidBioELECTRA also improves its recall from 93.49% in PADI-Web$_{long}$ to 94.62% in PADI-Web$_{XL}$ which is +0.33% ahead of PubMedBERT's improvement on the same metric. With longer contexts corpus in Table 7 experiment results, we still make observations on importance of biomedical domain in epidemiology surveillance. Longer context tend to simply amplify model benefits recorded in Table 5. This explains the simultaneous improvements of the top three models (EpidBioELECTRA, EpidBioBERT and PubMedBERT) on precision as captured by their shared metric score of +92.33%. ClinicalBERT suffers the most in performance compared to the other models falling short −3.08% below EpidBioELECTRA in F_1 score (see Table 8).



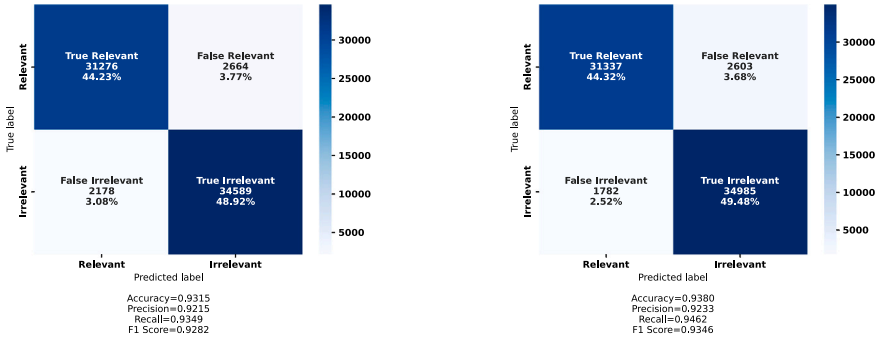
(a) Precision-Recall Curve comparing sensitivity versus PPV

(b) Receiver Operating Curve with area under a curve (AUC)

Fig. 13. Plots of EpidBioELECTRA skill on balancing sensitivity and PPV compared to EpidBioBERT and baseline models over PADI-Web$_{long}$.

Table 7
Results Table comparing EpidBioELECTRA performance against the baseline models in PADI-Web$_{long}$ and PADI-Web$_{XL}$ datasets with best scores in Bold. Prec%, Rec% and F_1% refer to precision percentage, recall percentage and F_1 score percentage respectively.

	PADI-Web$_{long}$			PADI-Web$_{XL}$		
Model	**Prec%**	**Rec%**	F_1%	**Prec%**	**Rec%**	F_1%
PubMedBERT	91.36	92.72	92.03	**92.33**	93.52	92.92
BioELECTRA	56.42	**99.95**	72.14	60.33	96.57	74.26
BERT	56.36	98.27	71.63	56.64	**99.91**	72.32
SciBERT	90.35	91.65	90.99	91.46	91.82	91.63
EpidBioBERT	90.95	92.95	91.93	**92.33**	93.28	92.8
EpidBioELECTRA	**92.15**	93.49	**92.81**	**92.33**	94.62	**93.46**
ClinicalBERT	89.09	91.65	90.99	89.96	90.82	90.38



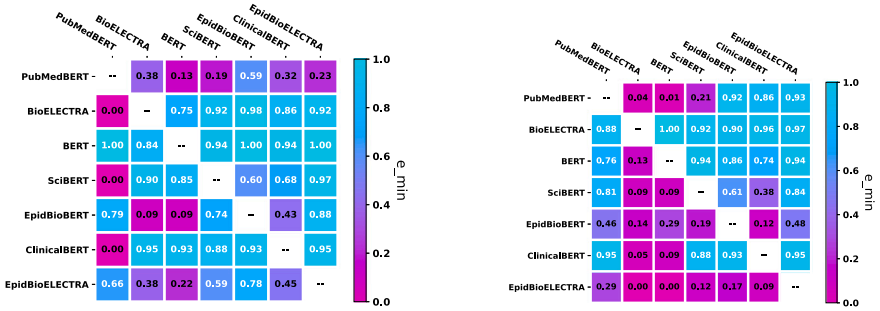
(a) Confusion matrix calculated over PADI-Web$_{long}$

(b) Confusion matrix calculated over PADI-Web$_{XL}$

Fig. 14. EpidBioELECTRA's confusion matrices.

Table 8
Fine tuning memory and time requirements of all models in all our datasets.

Model	PADI-Web$_{titles}$		PADI-Web$_{short}$		PADI-Web$_{long}$		PADI-Web$_{XL}$	
	GPU Memory Usage (GB)	Runtime (s)	GPU Memory Usage (GB)	Runtime (s)	GPU Memory Usage (GB)	Runtime (s)	GPU Memory Usage (GB)	Runtime (s)
PubMedBERT	5.5	10 405.6	13.1	170 389.9	21.5	304 893.7	27.15	617 733.6
BioELECTRA	2.8	6439.6	10.3	126 758.1	18.3	250 125.6	25.3	399516.0
BERT	1.5	5593.8	11.5	108 742.6	18.8	269 003.1	20.2	412350.3
SciBERT	3.8	9483.5	12.4	112 678.3	28.3	283 706.0	24.6	415002.8
EpidBioBERT	3.5	9732.4	14.5	130 212.8	20.5	315 668.2	27.4	609533.3
EpidBioELECTRA	3.6	9936.3	16.4	144 863.0	23.5	335 634.4	28.3	614746.1
ClinicalBERT	3.2	8769.4	15.7	126 488.3	19.7	295 124.6	26.8	528368.9



(a) ASO calculated over PADI-Web$_{short}$ (b) ASO calculated over PADI-Web$_{long}$

Fig. 15. Almost Stochastic Order (ASO) scores expressed in ϵ_{min} with a confidence interval of $\alpha = 0.5$ adjusted as per the Bonferroni correction (Bonferroni, 1936). Bold shades shows stochastic dominance showing that one algorithm (row) is better than the other (column). E.g. in (a) EpidBioELECTRA (row) is stochastically dominant over BioELECTRA (column) $\epsilon_{min} = 0.38$.

Table 9
Performance of EpidBioELECTRA classifier on PADI-Web$_{gold}$ compared to two competitive models. Best scores are in **bold**.

Model	F_1 *Score*	*Precision*	*Recall*	*Accuracy*
Competitive Models				
PubMedBERT	95.01	94.31	95.73	95.84
EpidBioBERT	95.61	95.33	95.89	96.13
Ours				
EpidBioELECTRA	**97.6**	**97.29**	**97.92**	**96.91**

5.3. Strong supervision performance testing for EpidBioELECTRA

In this subsection we provide test results of testing EpidBioELECTRA on our small gold labeled dataset PADI-Web$_{gold}$ which is made up of 800 humanly labeled articles. We take this approach in order to ascertain that the weak learning approach benefited our model. Gold labeling a dataset is a slow and expensive process that hampers strong supervision training, as a results weak supervision helps models generalize well as the amount of data is larger. For this testing approach, we track accuracy of competitive models (as trained using PADI-Web$_{XL}$) in that the best model should achieve an accuracy closer to 100% meaning it has well captured the human-expert's labeling decisions in PADI-Web$_{gold}$. We also set PADI-Web$_{gold}$ length to 512 to match PADI-Web$_{XL}$ used for training the competitive models. We present results in Table 9 from where we note that EpidBioELECTRA achieves both the highest classification accuracy and F_1 score compared to the closest competitive models. The accuracy performance of EpidBioELECTRA is respectively +1.11 and +2.91 above those recorded by EpidBioBERT (Menya et al., 2022) and PADI-Web classifier system (Valentin et al., 2019) on PADI-Web$_{gold}$.

5.4. Confidence levels for competitive models

In order to verify that EpidBioELECTRA's performance does not simply benefit from statistical chance, we compute Almost Stochastic Order (ASO) ϵ_{min} scores. We present confidence level values as measured by statistical dominance in Figs. 15(a) and 15(b) where we compares our models' classification performance confidence against competitive models in PADI-Web$_{short}$ and PADI-Web$_{long}$ datasets respectively. We observe statistical dominance of EpidBioELECTRA over competitive models recorded in both instances interpreted as 0.00 => ϵ_{min} <= 0.1 along the row as compared to competitive models along the column. Fig. 15(a) shows that EpidBioELECTRA (row) is stochastically dominant with significance over BioELECTRA, BERT and ClinicalBERT (column). We also note from the same table that EpidBioELECTRA is dominant compared to SciBERT, PubMedBERT and EpidBioBERT with a margin. Fig. 15(b), we observe the same pattern of EpidBioELECTRA's statistical dominance over competitive models with significance uniformity. This improvement can be attributed to the longer context length of the PADI-Web$_{long}$ test dataset.

5.5. Thematic feature influence

As an improvement to Menya et al. (2022) approach of measuring thematic feature importance in epidemiological classification, we employ an explainable AI approach to understand how much impact each thematic feature has in our epidemiology corpus classifier model. To this regard, we compute SHapley Additive exPlanations (SHAP) values on our test set corpus. SHAP values introduced by Lundberg and Lee (2017) are SOTA explainable AI computations for explaining the individual impact of every feature in a black box model. For example Fig. 16 shows a document in the test set where some key high value SHAP points are highlighted in color.

From the SHAP values we also learn feature directionality i.e. whether a feature point towards relevant or irrelevant directions in a classified document. For example in Fig. 17, we plot the feature directionality of thematic features in Fig. 16 based on their SHAP values (SHAP values have boht magnitude and direction). We further experiment on SHAP values of complete epidemiological phrases beyond single unigrams. For example, Fig. 18 plots most popular phrases of the corpus from Fig. 16 and the directionality of the phrases.

5.6. EpidBioELECTRA SHAP computations

As outlined in Section 4.1, our test set i.e. PADI-Web$_{test}$ covers (20%) of our corpus set. These consists of 14,400 documents that we use not only to test all models in Section 5.2, but also in computing our shap explanations. PADI-Web$_{test}$ contains 10,900 thematic features out of these we count 1300 unique features which we study in this section and present recorded results in Table 10.

We modify the approach of Lundberg and Lee (2017) from which we compute probabilistic shap values. These computed values answer the question of how much probability mass does one thematic feature contribute in influencing classification decision of a given corpus. We compute them based on Eq. (9). Given a corpus d_j in PADI-Web$_{test}$ denoted as τ and containing n thematic features $F = \{f_1, f_2, \ldots, f_n\}$ influencing classification decision $p(c_i|d_j)$ where $c_i \in C$ and $d_j \in \tau$ with $C = \{relevant, irrelevant\}$ we compute $SHAP_{f_n}(p(c_i|d_j))$ the shap value of feature f_n in influencing $p(c_i|d_j)$ for document $d_j \in \tau$ as:

$$SHAP_{f_n}(p(c_i|d_j)) = \sum_{N+1}^{n=1}[n\binom{N}{n}]^{-1}\psi_{f_n}(p(c_i|d_j))$$

Where $\psi_{f_n}(p(c_i|d_j))$ computes the marginal contribution of f_n in classification decision $p(c_i|d_j)$ and $[n\binom{N}{n}]^{-1}$ computes its weight (9)

...the sardinia region is trying to get out of the nightmare represented by african swine fever, just as the rest of italy is on alert. in all on the island there are only three outbreaks in breeding and since 2019 there have been no more cases in pigs illegally kept in the wild and while among wild animals the virus has no longer been found, the island tries to reappear on the pork market beyond regional borders. the recent action of councillor murgia the regional councillor of agriculture, gabriella murgia, with the other colleagues of the agricultural policies commission of the state regions conference, met on 10 march...

Fig. 16. An example of an epidemiological document of class relevant from PADI-Web$_{test}$ with thematic features tagged by EpidBioELECTRA. The shades of green represents shap values and their magnitude (captured by shade opacity) as used by the model to classify this document.

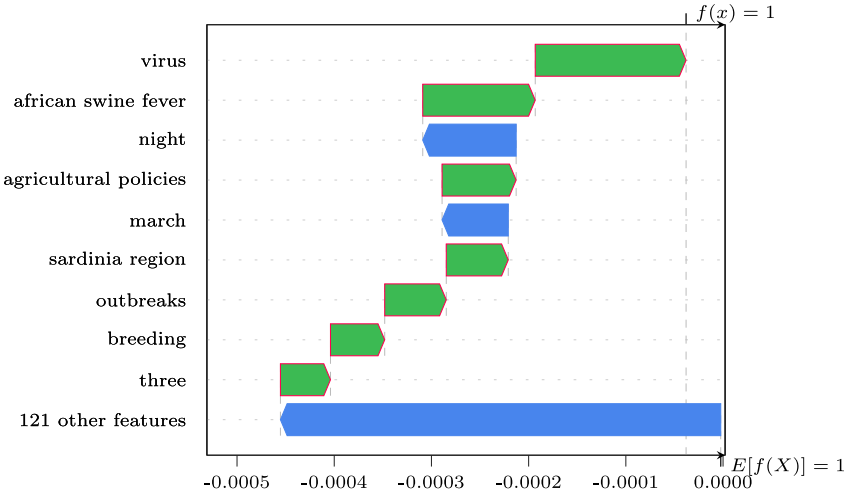


Fig. 17. Plot showing how thematic features shap values are impacting both relevant and irrelevant decisions by EpidBioELECTRA model on document shown in Fig. 16. Green shades represents relevant class and blue shade irrelevant class. The arrows point on the directional of shap pull with the classification value settled at the demarcated y-axis line.

Table 10
Impact of each Thematic Feature on influencing our classifier results as captured during EpidBioELECTRA training. Features with the most information contribute the most to both relevant and irrelevant classes, we show their scores in bold.

Thematic feature	Freq	Cumulative Impact		Impact Per Mention	
		rel	irrel	rel	irrel
Keyword	26	21.3157	0.1053	**0.8198**	0.004
Disease	33	23.9848	0.0207	0.7268	0.0006
Host	46	9.2895	0.2456	0.2019	0.0053
Symptom	20	3.5717	0.1873	0.1785	0.0093
Organization	337	**29.394**	**1.769**	0.087	0.005
Cases	246	15.7584	0.3192	0.064	0.0012
Location	264	16.5104	1.117	0.062	0.0042
Date	**387**	15.9813	1.4973	0.0412	0.0038
Time	28	0.2989	0.3304	0.0106	**0.0118**

In this our approach we can track the computations of Eq. (9) using Table 11. We notice that to compute the probabilistic shap value of n thematic features we have to construct 2^n test models to cover all possible combinations of feature-sets in order to study how each combination influences the model's classification decision. The first constructed model is made up of no features (an empty feature set ϕ) and is also known as the *base model*. On the other extreme, the complete model is made up of a full feature set F which is equivalent to the complete model versions tested in Section 5.2. In between these two model extremes, we form all possible feature combinations and compute individual probabilistic shap values.

As an example, if we are computing the probabilistic shap value of a disease thematic feature namely **african swine fever**, we can track the computations as captured by Fig. 19. Following this approach, we get the probability shap value for token **african** i.e. $SHAP_{african}(p(relevant|d_j))$ using Eq. (11). The term $\psi_{f_n}(p(c_i|d_j))$ is known as the **marginal**

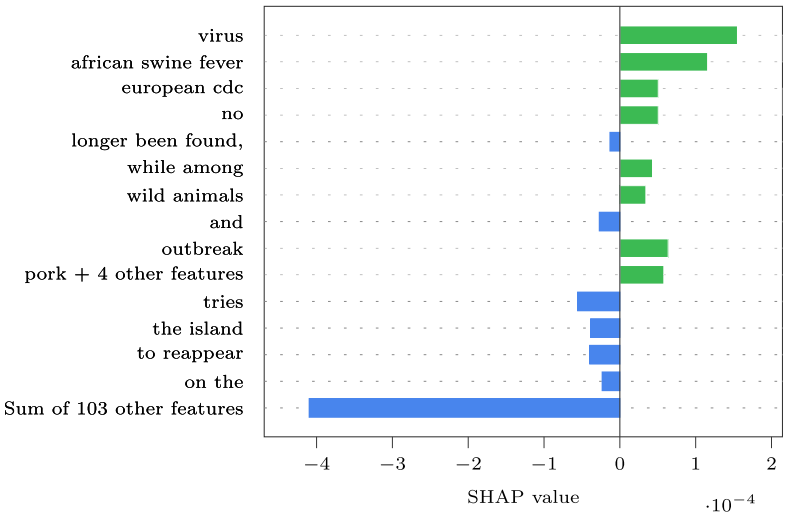


Fig. 18. Most popular phrases by shap value and their class directions in the entire document presented on Fig. 16.

Table 11
Tracking shap computations in 2^N EpidBioELECTRA test models for thematic features in document d_j. The first row represents a test model constructed with an empty feature set while the last row represents the complete model with a full feature set.

Iteration	Thematic set	Model prediction
0	ϕ	$p(c_l\|d_j)_1$
1	${}^nC_1$	$p(c_l\|d_j)_1, \ldots, p(c_l\|d_j)_3$
2	${}^nC_2$	$p(c_l\|d_j)_1, \ldots, p(c_l\|d_j)_3$
⋮	⋮	⋮
N	${}^nC_n$	$p(c_l\|d_j)_1$

contribution of feature f_n in the classification's probability mass $(p(c_i|d_j))$.

$$\psi_{f_n}(p(c_i|d_j)_{x,y}) = \begin{cases} [p(c_i|d_j)_{\{x,y\}} - \\ p(c_i|d_j)_{\{x-1,y\}}], & \text{for } f_n \notin p(c_i|d_j)_{\{x-1,y\}} \forall y \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Where x, y are co-ordinates of current cell by row and column and $x-1, y$ are the cells in preceding row

In Fig. 19, we track marginal contribution of token *african* along the edges leading to nodes where *african* is involved in local calculation. We calculate the overall contribution of token *african* by summing up its weighted marginal contributions in all models where the token appears in the model's thematic set. This visualization is achieved by considering the edge connecting parent to child node with the condition that the parent node does not contain feature *african* while the child node contains this feature in order to measure the effect of adding token *african* to a thematic feature set. We compute marginal contribution of a feature using Eq. (10) and finally these values are weighted and summed. Lundberg and Lee (2017) proposed weightings that follows two key rules; all weights must sum to one and weights at any given row must be equal to each other, thus we adopt $[n\binom{N}{n}]^{-1}$ weight index.

$$\begin{aligned} SHAP_{african}(p(relevant|d_j)) = {} & [1\tbinom{3}{1}]^{-1}\psi_{african}(p(relevant|d_j))+ \\ & [2\tbinom{3}{2}]^{-1}\psi_{african}(p(relevant|d_j))+ \\ & [2\tbinom{3}{2}]^{-1}\psi_{african}(p(relevant|d_j))+ \\ & [3\tbinom{3}{3}]^{-1}\psi_{african}(p(relevant|d_j)) \end{aligned} \tag{11}$$

Beyond shap values of an individual thematic feature, we study the overall impact of each broad class of thematic features over both relevant and irrelevant classes in the 1300 features in PADI-Web$_{test}$. To understand which features contribute the most to an epidemiological document being classified as either relevant or irrelevant, we sum up all computed probabilistic shap values for both relevant and irrelevant classes in a metric which we term **cumulative impact** computed as in Eq. (12). Since its trivial that a high frequency feature will have a higher cumulative impact mass, we compute a secondary metric termed **impact per mention** (IPM) which is as the norm value of cumulative impact of a feature divided by its frequency computed as in Eq. (13). We present these results in Table 10 and we rank features as per their IPM value over the relevant class.

$$CumImp_{f_n}(p(relevant|d_j)) = \sum_{\forall f_i \in F} SHAP_{f_i}(p(relevant|d_j)) \tag{12}$$

$$IPM_{f_n}(p(relevant|d_j)) = \frac{CumImp_{f_n}}{Count_{f_i}} \tag{13}$$

From Table 10, we note the ranking of thematic features *Keyword, Disease, Host* and *Symptom* as the most influential respectively. These features significantly contribute to the relevance of an epidemiological corpus as concluded from our test set. We interestingly note that the least useful feature, in this case *Time* (a spatio-temporal feature), is the most contributing to the irrelevance of an epidemiological document. From this our study, biomedical features are shown to be more influential while spatio-temporal features are the least influential in epidemiological classification task. We further study this interesting observation by plotting sankey diagrams of *Disease* Fig. 20 compared to that of *Time* Fig. 21 over their shapely values.

We note that presence of keyword feature followed by disease feature in a document contributes more to its relevance while presence of time feature splits between relevant and irrelevant decisions meaning the model losses its discriminative skill. We finally plot the overall thematic classes in Fig. 22 and we note that class relevant has the highest impact mass from all thematic features, this is key since it signifies that enriching thematic features improves epidemiological classification task.

6. Discussion

In this section we discuss the constraints and assumptions inherited in our approach in light of key improvements of our approach to the EBS architecture.

Our approach to epidemiological document classification focused on improving accuracy in discriminating between relevant and irrelevant

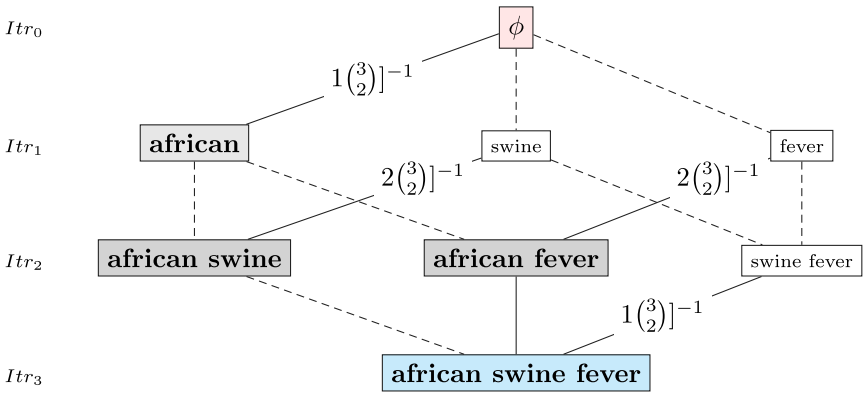


Fig. 19. Demonstrating the calculation of shap value for disease thematic feature **african swine fever**. The bold edges contain weights used while the bold-colored vertices show the nodes active when computing the shap value for token **african**. Each vertex represents an individual model and its feature set.

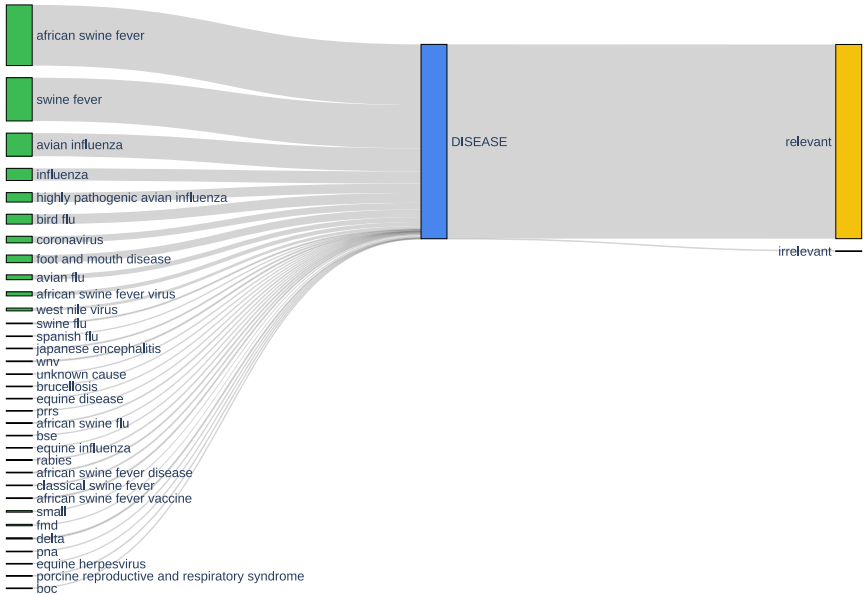


Fig. 20. Sankey Plot of Disease thematic feature shap scores over PADI-Web$_{test}$.

labels in such documents. In order to achieve better accuracy, we have proposed the use of mixed-domain pre-trained language models which are used to enrich thematic features that are fed into a neural network classifier architecture. To train our model using a large dataset, we created a machine labeled dataset out of a previously humanly labeled dataset in a semi-supervised learning fashion. Our approach thus achieves two key goals towards improving EBS; first our approach improves the quality of classification, secondly it improves the explanability of classification in EBS in order to understand how key thematic features influence this task. Two key concerns are plausible in our approach to improve EBS classification task; first using other pre-trained language models could impact the classifier in different ways. Secondly, even though using a semi-supervised approach with machine labeled dataset increases accuracy of our classifier, it does introduces label bias in the train and test sets. Future works could investigate ways to further de-bias our dataset.

Using explainable AI, we investigate EpidBioELECTRA's internal workings by the computation of probabilistic shap scores. We show that EpidBioELECTRA classifies epidemiological documents by focusing on relevant thematic features such as disease and symptom names while giving less focus to irrelevant features such as date and time of disease outbreaks. This approach towards explainability opens ground towards further understanding of epidemiological thematic features and how they affect text-based epidemiological surveillance tasks.

7. Conclusion and future work

This paper presents epidemiological thematic feature enrichment as a technique for improving epidemiological document classification. Our approach is based on the continual fine-tuning of a mixed domain language model using curated epidemiological datasets that is partly hand labeled and partly automatically labeled using weak supervision. Our model (EpidBioELECTRA) acquires significant thematic features

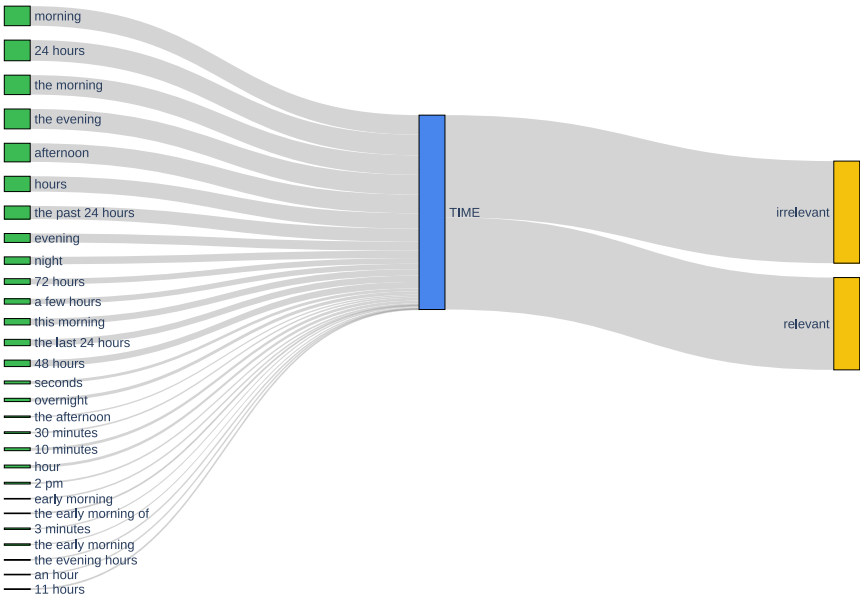


Fig. 21. Sankey Plot of Time thematic feature shap scores over PADI-Web$_{test}$.

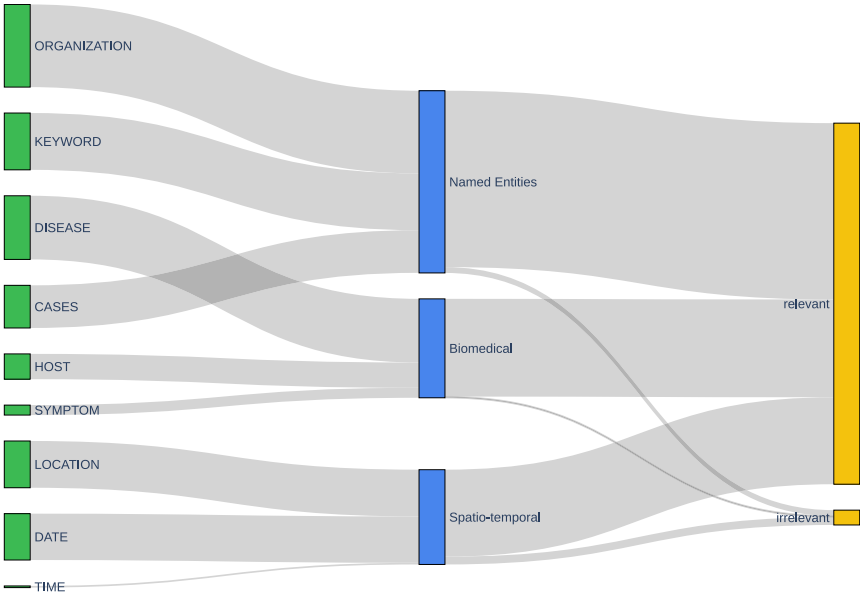


Fig. 22. Sankey Plot of thematic features and their types showing how they are contributing to both relevant and irrelevant classifications over the test set.

knowledge from our fine-tuning steps improving the downstream target task of epidemiological document classification. EpidBioELECTRA records a 19.2 F_1 score points improvement on our most challenging dataset PADI-Web$_{XL}$ compared to BioELECTRA. Also in this paper, we improve experimentation beyond EpidBioBERT model by testing EpidBioELECTRA on a hundred times the data used in EpidBioBERT. We experiment the ability of our model improving on epidemiological document classification by training using long context documents. We found out that this improves the classification task by 7.79 points compared to training using shorter context documents.

In future works, we propose investigating other robust language models with varied pre-training strategies, such as generative pre-trained approaches (GPT) in addition to investigating varied cross-domain training data sets to further improve text-based epidemiological surveillance tasks. Effects of epidemiological events duplication in corpora and text-quality effects can also be investigated in an attempt to understand thematic features further.

Ethics statement

The authors confirm compliance with the ethical policies of the journal, as noted on the journal's author guidelines page. No ethical approval was required because this study did not involve any experimental protocol on humans or animals, and only open source online data were used.

Code availability

The code developed and used in training and testing EpidBioELECTRA model is available at: https://github.com/menya-edmond/EpidBioELECTRA.

CRediT authorship contribution statement

Edmond Menya: Resources, Investigation, Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing,

Visualization. **Roberto Interdonato:** Methodology, Writing – review & editing, Supervision. **Dickson Owuor:** Methodology, Writing – review & editing, Supervision. **Mathieu Roche:** Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors make available the data used in experimentation and model evaluations published at Menya et al. (2023) and can be accessed at request.

Acknowledgments

This study was partially funded by "Ambassade de France - Nairobi", French General Directorate for Food (DGAL) and by EU grant 874850 MOOD and is catalogued as MOOD075. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop* (pp. 72–78). Minneapolis, Minnesota, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-1909, URL: https://aclanthology.org/W19-1909.

Arguello-Casteleiro, M., Jones, P. H., Robertson, S., Irvine, R. M., Twomey, F., & Nenadic, G. (2019). Exploring the automatisation of animal health surveillance through natural language processing. In M. Bramer, & M. Petridis (Eds.), *Artificial intelligence XXXVI* (pp. 213–226). Cham: Springer International Publishing.

Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., & Roche, M. (2018). Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PLOS ONE*, *13*, 1–25. http://dx.doi.org/10.1371/journal.pone.0199960.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, CoRR abs/1803.01271. arXiv:1803.01271.

Beltagy, I., Cohan, A., & Lo, K. (2019). SciBERT: Pretrained contextualized embeddings for scientific text, CoRR abs/1903.10676. arXiv:1903.10676.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information, CoRR abs/1607.04606. arXiv:1607.04606.

Bonferroni, C. E. (1936). Pubblicazioni del r instituto superiore di scienze economiche e commerciali di firenze. *Teoria statistica delle classi ecalcolo delle probabilità*, *8*, 3–62.

Brownstein, J. S., & Freifeld, C. (2007). HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Weekly Releases (1997–2007)*, *12*(48), 3322.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLOS Medicine*, *5*(7), 1–6. http://dx.doi.org/10.1371/journal.pmed.0050151.

Carter, D., Stojanovic, M., Hachey, P., Fournier, K., Rodier, S., Wang, Y., & de Bruijn, B. (2020). Global public health surveillance using media reports: Redesigning GPHIN, CoRR abs/2004.04596. arXiv:2004.04596.

Centre, J. R., for the Protection, I., of the Citizen, S., Linge, J., Belyaeva, J., & Mantero, J. (2011). *How to maximise event-based surveillance web-systems : The example of ECDC/JRC collaboration to improve the performance of MedISys*. Publications Office, http://dx.doi.org/10.2788/69804.

Chanlekha, H., Kawazoe, A., & Collier, N. (2010). A framework for enhancing spatial and temporal granularity in report-based health surveillance systems. *BMC Medical Informatics and Decision Making*, *10*, 1.

Collier, N., Doan, S., Kawazoe, A., Matsuda Goodwin, R., Conway, M., Tateno, Y., Ngo, H., Dien, D., Takeuchi, K., Shigematsu, M., & Taniguchi, K. (2008). BioCaster: Detecting public health rumors with a web-based text mining system. *Bioinformatics (Oxford, England)*, *24*, 2940–2941. http://dx.doi.org/10.1093/bioinformatics/btn534.

Del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2018). An optimal transportation approach for assessing almost stochastic order. In *The mathematics of the uncertain* (pp. 33–44). Springer.

Delon, F., Bédubourg, G., Bouscarrat, L., Meynard, J.-B., Valois, A., Queyriaux, B., Ramisch, C., & Tanti, M. (2024). Infectious risk events and their novelty in event-based surveillance: new definitions and annotated corpus. *Language Resources and Evaluation*, 1–19.

Dror, R., Shlomov, S., & Reichart, R. (2019). Deep dominance - how to properly compare deep neural models. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the Association for Computational Linguistics, ACL 2019, florence, Italy, July 28- August 2, 2019, volume 1: Long papers* (pp. 2773–2785). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1266.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, *15*(2), 150–157. http://dx.doi.org/10.1197/jamia.M2544, arXiv:https://academic.oup.com/jamia/article-pdf/15/2/150/2086063/15-2-150.pdf.

G., V., Kanjirangat, V., & Gupta, D. (2023). AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems with Applications*, Article 120440. http://dx.doi.org/10.1016/j.eswa.2023.120440, URL: https://www.sciencedirect.com/science/article/pii/S0957417423009429.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. http://dx.doi.org/10.1145/3458754, URL: http://dx.doi.org/10.1145/3458754.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Huff, A. G., Breit, N., Allen, T., Whiting, K., & Kiley, C. (2016). Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. *Interdisciplinary Perspectives on Infectious Diseases*, *2016*.

Jens, L., Ralf, S., Flavio, F., Stefano, B., Monica, G., Jenya, B., Delilah, A. K., Roman, Y., & Erik, V. D. G. (2010). MedISys - Medical information system. URL: https://api.semanticscholar.org/CorpusID:166930212.

Jiang, S., Cormier, S., Angarita, R., & Rousseaux, F. (2023). Improving text mining in plant health domain with GAN and/or pre-trained language model. *Frontiers Artificial Intelligence*, *6*, http://dx.doi.org/10.3389/frai.2023.1072329.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.

Joos, M. (1950). Description of language design. *The Journal of the Acoustical Society of America*, *22*(6), 701–707.

Kanakarajan, K. r., Kundumani, B., & Sankarasubbu, M. (2021). BioELECTRA:Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th workshop on biomedical language processing* (pp. 143–154). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.bionlp-1.16, URL: https://aclanthology.org/2021.bionlp-1.16.

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., & Barnes, L. E. (2018). RMDL. In *Proceedings of the 2nd international conference on information system and data mining*. ACM, http://dx.doi.org/10.1145/3206098.3206111.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 260–270). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-1030, URL: https://aclanthology.org/N16-1030.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. http://dx.doi.org/10.1093/bioinformatics/btz682.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2021). A survey on text classification: From shallow to deep learning. arXiv:2008.00364.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. arXiv:1711.05101.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, *19*(1), 59–99.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, *1*(4), 309–317. http://dx.doi.org/10.1147/rd.14.0309.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. http://dx.doi.org/10.48550/ARXIV.1705.07874, URL: https://arxiv.org/abs/1705.07874.

Menya, E., Interdonato, R., Owuor, D., & Roche, M. (2023). PADI-web corpus used for the EpidBioELECTRA approach. http://dx.doi.org/10.18167/DVN1/WD1UC2.

Menya, E., Roche, M., Interdonato, R., & Owuor, D. (2022). Enriching epidemiological thematic features for disease surveillance corpora classification. In *Proceedings of the language resources and evaluation conference* (pp. 3741–3750). Marseille, France: European Language Resources Association.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mutuvi, S. (2022). *Epidemic event extraction in multilingual and low-resource settings* (Ph.D. thesis), La Rochelle Université, URL: https://theses.hal.science/tel-03978917.

Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., & Odeo, M. (2020). Multilingual epidemiological text classification: A comparative study. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6172–6183). Barcelona, Spain (Online): International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.543, URL: https://aclanthology.org/2020.coling-main.543.

Mutuvi, S., Doucet, A., Lejeune, G., & Odeo, M. (2020). A dataset for multi-lingual epidemiological event extraction. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4139–4144). Marseille, France: European Language Resources Association, URL: https://aclanthology.org/2020.lrec-1.509.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1162, URL: https://www.aclweb.org/anthology/D14-1162.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (Long papers)*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n18-1202.

Rabatel, J., Arsevska, E., de Goër de Hervé, J., Falala, S., Lancelot, R., & Roche, M. (2017). *PADI-web corpus: news manually labeled*. CIRAD Dataverse, http://dx.doi.org/10.18167/DVN1/KMTIFG.

Ruder, S. (2021). Recent Advances in Language Model Fine-tuning. http://ruder.io/recent-advances-lm-fine-tuning.

Sahnoun, S., & Lejeune, G. (2021). Multilingual epidemic event extraction : From simple classification methods to open information extraction (OIE) and ontology. In *Proceedings of the international conference on recent advances in natural language processing* (pp. 1227–1233). Held Online: INCOMA Ltd., URL: https://aclanthology.org/2021.ranlp-1.138.

Steinberger, R., Fuart, F., van der Goot, E., Best, C., ETTER, P., & Yangarber, R. (2010). Text mining from the web for medical intelligence. *JRC*.

Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Fine-tuning large neural language models for biomedical natural language processing, CoRR. arXiv:2112.07869.

Ulmer, D., Hardmeier, C., & Frellsen, J. (2022). Deep-significance-easy and meaningful statistical significance testing in the age of neural networks. arXiv preprint arXiv:2204.06815.

Valentin, S., Arsevska, E., Falala, S., de Goër, J., Lancelot, R., Mercier, A., Rabatel, J., & Roche, M. (2020). PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, *169*, Article 105163. http://dx.doi.org/10.1016/j.compag.2019.105163, URL: https://www.sciencedirect.com/science/article/pii/S0168169919310646.

Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., & Roche, M. (2020). *PADI-web: An event-based surveillance system for detecting, classifying and processing online news* (pp. 87–101). http://dx.doi.org/10.1007/978-3-030-66527-2_7.

Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., & Roche, M. (2020). PADI-web: An event-based surveillance system for detecting, classifying and processing online news. In Z. Vetulani, P. Paroubek, & M. Kubis (Eds.), *Human language technology. Challenges for computer science and linguistics* (pp. 87–101). Cham: Springer International Publishing.

Valentin, S., Arsevska, E., Rabatel, J., Falala, S., Mercier, A., Lancelot, R., & Roche, M. (2021). PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance. *One Health*, *13*, Article 100357. http://dx.doi.org/10.1016/j.onehlt.2021.100357, URL: https://www.sciencedirect.com/science/article/pii/S2352771421001476.

Valentin, S., Valérie, D. W., Aline, V., Elena, A., Renaud, L., & Mathieu, R. (2019). Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus. http://dx.doi.org/10.18167/DVN1/YGAKNB.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

WHO (2008). *A guide to establishing event-based surveillance*. Manila: WHO Regional Office for the Western Pacific.

WHO (2014). *Early detection, assessment and response to acute public health events: Implementation of early warning and response with a focus on event-based surveillance: Interim version* (p. v, 59 p.). World Health Organization.

Woodall, J. P. (2001). Global surveillance of emerging diseases: the ProMED-mail perspective. *Cadernos de Saude Publica*, *17*, S147–S154.

Yu, V. L., & Madoff, L. C. (2004). ProMED-mail: an early warning system for emerging diseases. *Clinical Infectious Diseases*, *39*(2), 227–232.