*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3741–3750
Marseille, 20-25 June 2022


# Enriching Epidemiological Thematic Features For Disease Surveillance Corpora Classification

**Edmond Menya[1], Mathieu Roche[2,3], Roberto Interdonato[2,3], Dickson Owuor[1]**
[1] SCES Strathmore University, Nairobi, Kenya
[2] CIRAD, F-34398 Montpellier, France
[3] TETIS - Univ Montpellier - AgroParisTech - CIRAD - CNRS - INRAE, Montpellier, France
{emenya, dowuor}@strathmore.edu
{mathieu.roche,roberto.interdonato}@cirad.fr

**Abstract**

We present **EpidBioBERT**, a biosurveillance epidemiological document tagger for disease surveillance over PADI-Web system. Our model is trained on PADI-Web corpus which contains news articles on Animal Diseases Outbreak extracted from the web. We train a classifier to discriminate between relevant and irrelevant documents based on their epidemiological thematic feature content in preparation for further epidemiology information extraction. Our approach proposes a new way to perform epidemiological document classification by enriching epidemiological thematic features namely disease, host, location and date, which are used as inputs to our epidemiological document classifier. We adopt a pre-trained biomedical language model with a novel fine tuning approach that enriches these epidemiological thematic features. We find these thematic features rich enough to improve epidemiological document classification over a smaller data set than initially used in PADI-Web classifier. This improves the classifiers ability to avoid false positive alerts on disease surveillance systems. To further understand information encoded in EpidBioBERT, we experiment the impact of each epidemiology thematic feature on the classifier under ablation studies. We compare our biomedical pre-trained approach with a general language model based model finding that thematic feature embeddings pre-trained on general English documents are not rich enough for epidemiology classification task. Our model achieves an F1-score of $95.5\%$ over an unseen test set, with an improvement of $+5.5$ points on F1-Score on the PADI-Web classifier with nearly half the training data set.

**Keywords:** Epidemiology Intelligence, Thematic Epidemiology Features, Disease Surveillance, Text Mining, Corpus Classification

## 1. Introduction

In recent years, with the rise of infectious disease outbreaks, more focus has been put on epidemic intelligence and biosurveillance intelligent systems. These epidemiological surveillance systems have gained track within the natural language processing community. Such systems are able to monitor official intergovernmental digital sources as well as unofficial sources such as unstructured web news articles for early detection and reporting of existing, reemerging and novel disease outbreaks (Woodall, 2001; Arsevska et al., 2018; Valentin et al., 2021).

Early disease surveillance systems generally used an indicator based approach that uses formal rule systems to monitor relevant official sources (Paquet et al., 2006). Current surveillance systems such as `ProMED`, `HealthMap` and the Platform for Automated extraction of Animal Disease Information from the web `PADI-Web`, use an event based approach with multiple different corpora and language sources (Woodall, 2001; Brownstein and Freifeld, 2007; Arsevska et al., 2018).

PADI-Web is an event based biosurveillance system developed for the French Epidemic Intelligence System (FEIS) focused on monitoring online news sources for detection and alerting of existing and emerging infectious animal diseases (Valentin et al., 2020a; Valentin et al., 2021). PADI-Web 1.0 by Arsevska et al. (2018) used a keyword-based classification approach where corpora are classified based on existence of one or more preset list of disease outbreak-related keywords in the document. This classification process is then further enhanced in PADI-Web 2.0 (Valentin et al., 2020b) by incorporating a multilingual module and machine learning techniques based on bag-of-words and Term Frequency - Inverse Document Frequency (TF-IDF) approaches (Luhn, 1957; Jones, 1972). Later PADI-Web 3.0, Valentin et al. (2021), proposes 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.).

Even though epidemic intelligence has grown with the introduction of event based epidemiology surveillance systems, major challenges include their reliance on labeled data sets for supervised learning training. Labeling such data is relatively costly and time consuming. In order to train an epidemiology document classifier, human experts have to manually label unstructured news articles as either relevant, related or irrelevant for further disease surveillance processing (Paquet et al., 2006). Relevant corpora are those news articles that describe an infectious animal disease outbreak event, irrelevant corpora are those which are not related to dis-

ease outbreak, and related corpora contain a different subject which relates to disease outbreaks (Arsevska et al., 2018).

Beyond machine learning approaches of building language understanding models, there has been great advances in deep learning techniques based on word embeddings and language models. Word Embeddings represent the semantic meaning of lexicon encoded as vectors. Various successful models have been developed for this task. This begun with context-insensitive embeddings; Word2Vec proposed by Mikolov et al. (2013), Global Vectors for Word Representation (GloVe) Embeddings by Pennington et al. (2014), and FastText (Bojanowski et al., 2016). The works of Peters et al. (2018), introduced models of encoding contextual knowledge into word embeddings, models which have now been far improved through the use of Transformer networks in place of both Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) Networks (Vaswani et al., 2017).

This study aimed at developing a new thematic embedding based approach for epidemiological corpus classification over tagged news article sources. The classifier improves over current keyword based and machine learning based approaches. Our epidemiology document classifier learns rich thematic embeddings to discriminates between relevant and irrelevant news sources for further disease surveillance in PADI-Web system. The contributions of this study are as follows:

- Our study proposes a new way to improve epidemiological document classification by enriching epidemiological thematic features mainly disease, host, location and date contained in animal disease news articles that forms our train set corpus. We achieve this by proposing EpidBioBERT model whose architecture is two staged; pre-trained on BioBERT (Lee et al., 2019) language model that we use to learn thematic embeddings for our features and fine-tuned to learn a classifier model that discriminates between relevant and irrelevant news corpora for epidemic intelligence tasks.
- We show that fine tuning a biomedical language model improves epidemiological corpus classifier more significantly than fine tuning a general purpose pre-trained language model such as BERT (Devlin et al., 2018).
- We experiment the impact of each individual thematic feature in the overall epidemiological classifier showing that both host and disease thematic features carry vital information on corpus relevance for epidemic intelligence.
- Our model achieves State-Of-The-Art performance with a smaller data set without using thesaurii to enrich thematic features [1].
- We improve over false positive alerts in epidemiology surveillance brought about by missclassifying news articles that mention disease-free countries and those describing aftermaths of a disease outbreak.

The rest of the paper is organized as follows: Section 2 outlines related works majorly reviewing PADI-Web document classifier and our major contributions; Section 3 introduces our model its architecture and pipeline; Section 4 discusses empirical results from experiments comparing the baselines with our model; Section 5 presents ablation studies findings on key model and data aspects and Section 6 summarizes the entire paper.

## 2. Background

**PADI-Web Relevance Classification.** This work is majorly an extension of the epidemiology document classification by Arsevska et al. (2018), Valentin et al. (2020b) and Valentin et al. (2021). The PADI-Web classifier takes two approaches. Firstly disease corpus is tagged using predetermined list of disease outbreak related keywords. In the second approach, a more accurate supervised binary relevance classifier that takes manually tagged news articles encoded using TF-IDF as inputs is developed. A bag-of-words representation combined with machine learning techniques are implemented in order to identify relevant and irrelevant documents. As an extension to improve document classification, PADI-Web further incorporates a more fine-grained topic classification task to refine article relevance. This helps set priority for news articles that declare disease outbreak as opposed to those that describe outbreak consequences or aftermaths such as the economic effects caused by infectious disease outbreak.

A more recent approach focuses on a fine-grained classification of sentences dealing with animal disease surveillance based on Naive Bayes, Random Forest and Surpoort Vector Machines (SVM) algorithms (Valentin et al., 2021). Our approach proposes a pre-trained BioBERT language model to enrich such epidemiological information to extend this work. These learnt epidemiological thematic embeddings form the thematic features that are fed into our fine tuned model.

**BioMedical Language Model Fine Tuning.** As demonstrated by Broscheit (2020) large pre-trained language models such as ELMo and BERT contain entity linking knowledge within their trained architectures (Peters et al., 2018; Devlin et al., 2018). Other research has looked into such large pre-trained models for specific domains. For instance, in the Biomedical domain, Lee et al. (2019) introduced BioBERT, an original BERT architecture trained over biomedical corpora; other biomedical language models such as ClinincalBERT and BioELECTRA have been proposed in recent years (Huang et al., 2020; Kanakarajan et al., 2021). Original ways to fine tune such

[1] https://github.com/menya-edmond/EpidBioBERT

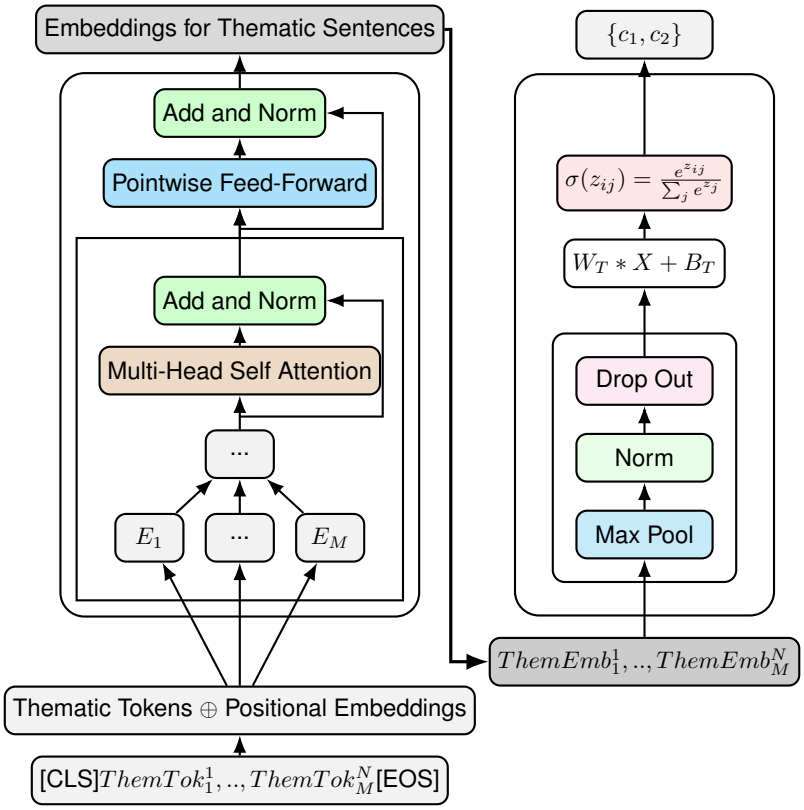


Figure 1: EpidBioBERT Transformer Architecture with fine tuned deep layers on top of pretrained BioBERT. [CLS]$ThemTok_1^1, .., ThemTok_M^N$[EOS] are the $N$ thematic feature tokens from $M$ sentences in the annotated train corpus that are inputs to the model. [CLS] and [EOS] are the tokenizers' tags for start and end of sentence respectively. A probability distribution over document classes *relevant* and *irrelevant* represented as $\{c_1, c_2\}$ are the output labels.

large pre-trained models can then be defined, in order to improve their effectiveness in various biomedical related NLP tasks, e.g., epidemiological surveillance document classification. Recently, Ruder (2021) has investigated different fine tuning techniques to improve such domain specific tasks. Our technique uses the behavioural fine tuning approach over PADI-Web dataset, with the aim of enriching thematic features.

**Behavioural Fine Tuning.** This technique, as demonstrated by (Ruder, 2021), adopts an intermediate task approach in its training pipeline. The pre-trained language model is first trained end-to-end on the intermediate task in order to improve over a related specific task. The work in Phang et al. (2018) shows that this approach is beneficial for the downstream tasks that depend on language understanding. Sun et al. (2020) showed how pre-trained language models benefits from few shot learning for tasks that have smaller training data sets. In this study, we extend PADI-Web classifier using a behavioural few shot learning approach. In our approach, BioBERT language model is first fine tuned for our epidemiology task, then used to enrich epidemiology thematic features with a focus of improving disease corpora classification.

## 3. EpidBioBERT

In this section, we introduce our epidemiological corpus classifier model (EpidBioBERT). EpidBioBERT adopts a two step framework transfer learning approach using a pre-trained Biomedical Language Model followed by fine tuning to improve epidemiological document classification in PADI-Web disease surveillance system.

### 3.1. Model Architecture

EpidBioBERT model architecture uses a base BioMedical Language model (i.e., BioBERT) with fine tuned disease surveillance deep layers above its architecture as described in Fig. 1. BioBERT is a deep self-attention model pretrained over biomedical corpora achieving State-of-The-Art levels in biomedical text mining on BLURB leader-board tasks (Gu et al., 2022). Three pretrained versions of BioBERT are introduced in Lee et al. (2019), namely BioBERT(+PubMed), BioBERT(+PMC) and BioBERT(+PubMed+PMC). These versions differ in architecture size since they are pretrained on different datasets. Our architecture is based on the BioBERT(+PubMed) model (Devlin et al., 2018; Lee et al., 2019).


Figure 2: EpidBioBERT input representation. The Thematic embeddings $\vec{w}$ are enriched by BioBERT embeddings from word piece tokenized annotated tokens from training copora.

In related research, Broscheit (2019) fine tuned BERT for named entity recognition and showed that different layers of BERT contain different encoded information. For instance, the lower layers of BERT are rich in named entities information which are related to our epidemiological thematic features. In our approach, we first fine tune the whole BioBERT model by unfreezing all the weights and using the last state of the pretrained optimizer to train end to end on PADI-Web corpus. The second fine tuning stage tunes for classification similar to the baselines. We adopt a cross entropy loss objective function over the two target classes. Our model learns to maximize probability of the correct classes. The output is a document tag with a *relevant/irrelevant* label.

### 3.2. Model Definition

Epidemiological document relevance classification task takes in a set of $N$ *disease outbreak* news articles which we denote as $D = \{d_1, ..., d_N\}$. The task can formally be defined as: given a disease outbreak news article $d_j \in D$ which contains $n$ epidemiological thematic features denoted $F = \{f_1, ..., f_n\}$, output a probability distribution classifying the article as either of the document classes $C = \{c_1, c_2\}$ where $c_1 = relevant, c_2 = irrelevant$ for epidemiological surveillance. Our model learns to maximize the probability $p(c_i|d_j)$ where $c_i \in C$ and $d_j \in D$ by minimizing the models' objective function:

$$L = \frac{1}{N_b} \sum_{i}^{|C|} \sum_{j}^{|N_b|} -\{y_{ij} * \ln \sigma(z_{ij})\}$$

Where $b$ is the batch size set as hypeparameter described in 4.2 and $y_{ij}$ is the true label vector for training input corpus $d_j$ signifying its true class label $c_i$. $z_{ij}$ is the output from our last linear layer such that $z_j = W_T * X + B_T$ where $W_T$ is the layers' weight matrix that is multiplied with the epidemiology thematic features embedding matrix $X$ learnt from the pre-trained model and added to the bias matrix $B_T$. $z_{ij}$ is then *softmaxed* such that $\sigma(z_{ij}) = \frac{e^{z_{ij}}}{\sum_j e^{z_j}}$ followed by taking of natural logs to get predicted output vector $\hat{y_{ij}} = p(c|d_j)$, for $c \in \{c_1, c_2\}$.

We achieve EpidBioBERT by adopting a deep pretrained transformer based model and deep fine tuned network layers as explained in Fig. 1.

### 3.3. EpidBioBERT Training Data

Our training corpora is derived from PADI-Web dataset which is supplied in JSON format (Rabatel et al., 2017). We prepare our corpus from the original data set, by running a cascade of regex rules to extract epidemiology thematic features,their annotated labels and document gold labels.

Our regex cascade rules find 180 news articles (i.e., about 35%) tagged as relevant and 350 (i.e., about 65%) tagged as irrelevant. The articles contain epidemiological entities manually labeled by human experts. Each token considered as an epidemiological entity candidate is originally labelled as either *location* for location of disease outbreak, *date* for date of disease outbreak, *number* for the number of reported cases, *disease* for type of disease experienced in the outbreak, and *host* for the disease carrier species. We form our epidemiological thematic features by selecting disease, host, date and symptoms to enrich our classification, the *number* epidemiology feature is dropped. In addition, these epidemiological entity candidate are labelled in PADI-Web data as, "correct", "partial" or "incorrect" with correct label signifying relevant candidate features for information extraction task while incorrect signifying irrelevant candidate features (Arsevska et al., 2018). For example having a date candidate feature that does not fit the date of disease event occurrence would have that date feature labelled irrelevant. There are $6K$ thematic features with 66% relevant, 20% irrelevant and 14% partial relevant as captured in Fig. 3.

The resulting data set contains both relevant and irrelevant documents, which in turn contain epidemiology thematic features. For experimentation purposes, we use 60% of the documents for training, 20% as Held-Out corpus for hyperparameter tuning and the remaining 20% for Model performance Evaluation. We cross validate these splits for different K-Folds under Section 5.4, in order to understand how our approach generalizes to train datas' true distribution as the data set size increases and instances change.

### 3.4. Epidemiology Thematic Feature Engineering

To enable our model and baselines to take in significant inputs, the annotated corpus detailed in 3.3 is first preprocessed using NLTK library. The train corpus is first preprocessed by *sentence segmentation* to extract sentences from running text. This process is followed by

*tokenization* to separate words from running sentences forming tokens while maintaining their annotated labels.

Thematic tokens for each document are then selected as per their annotated labels captured in Fig. 3. These thematic labeled tokens form our model inputs as thematic features. The thematic features are then *vectorized* using various strategies, depending on the baseline model (as described below), to form thematic embeddings having each thematic feature represented by a vector $\vec{w} \in \mathbb{R}^{1*|V|}$ where $|V|$ is the size of the vocabulary set.

For the bag-of-word approach, we initially set $|V| = 30K$ and represent every thematic feature with a one hot vector $\vec{w}$ such that $\vec{w} \in \mathbb{R}^{1*|30K|}$. Since this vector is sparse, we use truncated Singular Value Decomposition (truncated-SVD) setting the number of critical values to 300 to convert the thematic embeddings to dense vectors of dimension $\vec{w}^{|1*300|}$.

For the TF-IDF approach, we also set $|V| = 30K$ thus having thematic embeddings with the same dimension as initial bag-of-words approach. However, since these embeddings are not sparse, we use Principal Component Analysis (PCA) setting number of principal components to 300 for dimensionality reduction thus we end up with thematic embeddings of dimension $\vec{w}^{|1*300|}$. For the pretrained approach, we use pretrained GloVe embeddings (Pennington et al., 2014) of the same $1 * 300$ dimension for our thematic features.

For our transfer learning approach, we first tokenize using the pretrained BioBERT word piece tokenizer, then learn thematic feature embeddings from pre-trained BioBERT transformer encoder architecture as reported in Fig. 2. Each document in the data set is embedded by a tensor of dimension $(no\_of\_sentences, sequence\_len, |V|)$ where *no_of_sentences* represents sentence count in the document, having set $sequence\_len = 128$ initially and $|V| = 768$ from the pretrained architecture. Each thematic feature is embedded by a $\vec{w}$ such that $\vec{w} \in \mathbb{R}^{1*|768|}$.

Furthermore, we experiment with different embedding dimensions to see how this improves our epidemiology corpora classification task in Section 5.

For the baseline model, words contained in a given news article form the feature set for each document, with unique words forming the Vocabulary set $|V|$. We experiment with all the above embeddings beginning with One Hot Encoded thematic vectors for the bag-of-words Approach. On the Transfer learning approach, pre-trained thematic embeddings form the features for each document. We experiment using pretrained GloVe embeddings. In the deep learning approach, we adopt pre-trained BioBert thematic embeddings after which we compare classification results presented in Section 4.

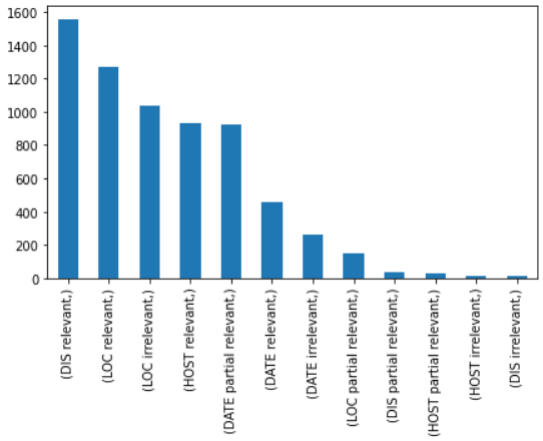


Figure 3: Epidemiology Thematic Feature Distribution in PADI-Web Dataset.

## 4. Experiments

In this section we compare and report EpidBioBERT results with machine learning classifiers recently used by Valentin et al. (2021) over PADI-Web system.

### 4.1. Baselines

We first experiment with the bag-of-words approach having One Hot encoded (OHE) epidemiology thematic feature embeddings. We train an SVM classifier model boosted with Gaussian Kernel (SVM+OHE) with the thematic embeddings as input features and the document class labels *relevant* and *irrelevant* as targets. We then experiment beyond the bag-of-words approach using both TF-IDF (SVM+TF-IDF) and pre-trained GloVe thematic embeddings (SVM+GloVe) over the same SVM model. We then experiment further by training an LSTM classifier model with the GloVe thematic embeddings first by freezing the embedding layer (LSTM+GloVe$_{frozen}$) and then by training an end to end classifier (LSTM+GloVe$_{unfrozen}$). We also train a similar Bidirectional LSTM model (Bi-LSTM+GloVe$_{unfrozen}$) and compare performance of all these baselines in Table 1.

### 4.2. Hyperparameters

We fine tune BioBERT(+PubMed) model (Lee et al., 2019) with hidden embedding size of 768, 12 Attention Heads and 12 Transformer blocks (Vaswani et al., 2017). We set a small batch size of 16 and a sequence length of 128 and experiment with 50 epochs. For our higher fine tuned layers, we experiment with dropout rates of 0.2,0.3 and 0.4 to control model overfitting.
We adopt Adam optimizer with decoupled weight decay (AdamW) (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, we set $\epsilon$ =1e-8 and weight decay = 0.01. We also set small initial learning rates of 1e-5 and 2e-5 with a higher number of epochs to favour our fine tuning approach (Ruder, 2021). We then evaluate and save the best model over the held out corpus.

| Model | $F_1Score$ | $Precision$ | $Recall$ | $Accuracy$ |
|---|---|---|---|---|
| *Baselines* | | | | |
| SVM+OHE | 0.29 | **1.0** | 0.17 | 70.0 |
| SVM+TF−IDF | 0.35 | 0.83 | 0.22 | 77.12 |
| SVM+GloVe | 0.51 | 0.65 | 0.55 | 65.34 |
| LSTM+GloVe$_{frozen}$ | 0.84 | 0.84 | **0.85** | 86.13 |
| LSTM+GloVe$_{unfrozen}$ | 0.85 | 0.85 | **0.85** | 87.12 |
| Bi−LSTM+GloVe$_{unfrozen}$ | **0.86** | 0.89 | **0.85** | **88.11** |
| *Ours* | | | | |
| EpidBioBERT | **0.95** | 0.97 | **0.94** | **95.8** |

Table 1: Performance of Our Model in PADI-Web Epidemiology Feature Extraction. One hot encoded based model are denoted OHE. Best scores are in **bold**

.

### 4.3. Results and Discussions

In this section we report and discuss baseline performance as compared to EpidBioBERT.

#### 4.3.1. Performance Metrics

We evaluate our model by comparing it with all the baseline models. We compute Precision, Recall, F1 Score and Accuracy performances for all classifiers. The models are evaluated on their effectiveness to classify documents as either *relevant* or *irrelevant*, by comparing *model-predicted* versus *true* document labels as explained in Section 3.2 over an unseen test set.

#### 4.3.2. Results

In Table 1, we compare and report EpidBioBERT results with machine learning classifiers recently used by Valentin et al. (2021) over PADI-Web system. The difference between SVM+OHE and SVM+TF-IDF is that the latter takes TF-IDF thematic embeddings as inputs while the former takes one hot thematic embeddings as input features. We also train LSTM based models with GloVe thematic features as inputs to enhance baseline performance. We freeze the GloVe pretrained thematic embeddings in LSTM+GloVe$_{frozen}$ model but train end-to-end in LSTM+GloVe$_{unfrozen}$.

#### 4.3.3. Discussions

Results recorded on both SVM+OHE and SVM+TF-IDF baseline models show that they commit a lot of false negative errors leading to low recall scores. This is attributed to the low quality of both one hotted and TF-IDF epidemiology thematic embeddings. Such embeddings encode limited thematic information required for correct discrimination between relevant and irrelevant documents by the epidemiological classifier. We note a recall score improvement between 33 and 38 points when we use pre-trained GloVe embeddings on the same SVM baseline model.

Even though pre-trained thematic feature embeddings significantly improves recall and overall F1 scores we note that this largely depends on the nature of pre-training. Epidemiological thematic embeddings pre-trained on general English corpora prove not rich enough for robust epidemiology feature classification task (we study this further in Section 5.1 by comparing BERT to BioBERT as a base pretrained model for our classification task). This is the case with 5 to 12 points accuracy drop in SVM+GloVe model given the nature of GloVe embeddings (Pennington et al., 2014). We attribute this largely due to the fact that the number of Out of Vocabulary (OOV) thematic tokens unseen during pre-training are high since the thematic embeddings are not pretrained on Bio Medical data. This causes features like diseases, host and location to largely fall under OOV. In Section 5.3 we investigate if this can be improved with better generalization and higher dimensional thematic embeddings.

LSTM based baselines improve performance using GloVe thematic embedding features as inputs. Bi-LSTM+GloVe$_{unfrozen}$ achieves the best baseline scores. We attribute this first to the nature of end-to-end training that allows the model to learn epidemiology thematic feature embeddings from the PADI-Web train set data, and second to the Bi-directional (Melamud et al., 2016) nature of the Bi-LSTM model that enables it to learn both left and right contexts of thematic embeddings enriching them further than unidirectional LSTM+GloVe architecture.

EpidBioBERT outperforms Bi-LSTM+GloVe$_{unfrozen}$ on precision, recall, F1 and accuracy scores. This we attribute to a number of key architectural benefits that EpidBioBERT has over all the baselines. First the underlying pre-trained model BioBERT uses transformer (Vaswani et al., 2017) architecture which has better surrounding-features context management Attention technique that surpasses LSTM architectures. Secondly, BioBERT is based on BERT (Devlin et al., 2018) which is a bidirectional architecture thus maintaining the bidirectionality advantages aforementioned. Thirdly, we fine tune BioBERT(+PubMed) that is pre-trained on biomedical data, this enriches our epidemiological thematic features beyond capabilities of a general purpose language model. Lastly, EpidBioBERT novel fine tuning technique uses network layers and hyperparameters that favour few shot learning that coun-

| Thematic Feature | $F_1 ScoreDrop$ | Precision Drop | Recall Drop | Accuracy Drop |
|---|---|---|---|---|
| Date | $-4$ | $-3$ | $\mathbf{-5}$ | $-4.8$ |
| Location | $-1$ | $-6$ | $+2$ | $-2$ |
| Host | $\mathbf{-8}$ | $\mathbf{-18}$ | $+2$ | $\mathbf{-9.4}$ |
| Disease | $-5$ | $-12$ | $+2$ | $-5.8$ |
| *All Features* | 0.0 | 0.0 | 0.0 | 0.0 |

Table 2: Impact of each Thematic Feature on our classifier performance as captured during EpidBioBERT training and testing. Features with the most information have cause highest drop in both F-Score and Accuracy as shown in bold.

ters our small and class imbalanced train set (Sun et al., 2020; Ruder, 2021). We study further the effect of small train and test sets in EpidBioBERT and selected baselines in Section 5.4 by use of cross validated experiments.

## 5. Discussion: Ablation Studies

In this section we present the findings on ablation experiments, carried out to understand the independent effects of quality of thematic features and data size on our model.

### 5.1. Effect of Biomedical Pretraining

To test if biomedical pretraining improves our epidemiological document classification task, we evaluate our model alongside a BERT based model pretrained on general English corpora. To set a fair comparison, we fine tune distill-bert model (Sanh et al., 2020) for our classification task setting hyperparameters to be the same as those used in EpidBioBERT described in 4.2.
We present this experiment results in Fig. 4 showing the training loss over 50 epochs. We report on the training loss metric observing that the loss decreases much faster and more significantly for EpidBioBERT as compared to the BERT based model. This shows that BioBERT significantly improves training of our epidemiological classifier largely because BioBERT is pretrained on biomedical corpora thus enriching epidemiological thematic features as opposed to BERT.

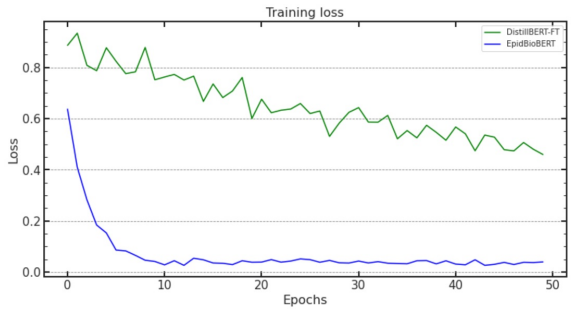


Figure 4: Train Loss scores of EpidBioBERT compared to BERT-FT for epidemiology document classification.

### 5.2. Effect of Thematic Feature Enrichment

To understand how much impact each thematic feature has in our epidemiology corpus classifier model, we experiment by running EpidBioBERT model on different data sets each having one epidemiology thematic feature dropped. We focus on four thematic features namely Disease, Host, Location and Date and prepare four train, validation and test sets. We present results in Table 2 showing the F1-Score, precision, recall and accuracy drops as compared to EpidBioBERT model results in Table 1.
The Host thematic feature causes the highest drop in classifier accuracy and F-Score of $-9.4$ and $-8$ respectively. Training the classifier without Date thematic information reduces models recall value increasing false negative classification error. Location, Disease and Host thematic features roughly contain the same information to equally influence recall measure meaning less false negatives. The results signifies that both Host and Disease are the key thematic features influencing our classifier followed by Date and lastly Location.
We find this interesting since the distribution of thematic features in the data follows a different distribution order with Disease, Location, Host and Date, in that order, having high representations as presented in Fig. 3.

### 5.3. Effect of High Dimension Thematic Embeddings on Baselines

We explore the effects of having deeper embeddings as inputs to our model. High dimensional word embeddings can be thought as being rich since they contain more dimensions to carry context information. We start by training different versions of SVM+OHE model with increasing OHE embedding dimensions and report results in Table 4.
We find that for the bag-of-words based approach, increasing embedding dimension alone does not necessarily improve performance as seen between the baseline models SVM+OHE$_{50dims}$, SVM+OHE$_{100dims}$ and SVM+OHE$_{300dims}$. However, beyond the bag-of-words approach performance can slightly be improved in this way as shown between LSTM+TF-IDF$_{100dims}$ LSTM+TF-IDF$_{300dims}$. We achieve a competitive classifier using LSTM+GloVe$_{300dims}$.

| EpidBioBERT+CV folds | Avrg Acc | Acc Std Dev | Avrg F-Score | F-Score Std Dev |
|---|---|---|---|---|
| 2 | 0.6670 | ±0.0465 | 0.8083 | ±0.0583 |
| 3 | 0.8024 | ±0.0534 | 0.8685 | ±0.0464 |
| 4 | 0.7131 | ±0.028 | 0.8680 | ±0.0542 |
| 5 | **0.8492** | ±0.0398 | 0.8239 | ±0.0393 |
| 6 | 0.8139 | ±0.0368 | 0.8167 | ±0.0423 |
| 7 | **0.8447** | ±0.0352 | **0.8784** | ±0.0513 |

Table 3: EpidBioBERT+CV K-Folds Test-set Average Accuracy, F1 Scores and their Standard Deviations showing folds 2 upto 7. The model demonstrates a small deviation uniformly among the data folds. The best scores are in **bold** having considered ones with smaller deviations.

| Model | $F_1 Score$ | Accuracy |
|---|---|---|
| *OHE Thematic Features* | | |
| $\texttt{SVM+OHE}_{50dims}$ | 0.34 | 0.71 |
| $\texttt{SVM+OHE}_{100dims}$ | 0.31 | 0.62 |
| $\texttt{SVM+OHE}_{300dims}$ | 0.29 | 0.70 |
| *TF-IDF Thematic Features* | | |
| $\texttt{SVM+TF-IDF}_{50dims}$ | 0.30 | 0.71 |
| $\texttt{LSTM+TF-IDF}_{100dims}$ | 0.31 | 0.72 |
| $\texttt{LSTM+TF-IDF}_{300dims}$ | 0.35 | **0.77** |
| *GloVe Thematic Features* | | |
| $\texttt{SVM+GloVe}_{300dims}$ | 0.64 | 0.69 |
| $\texttt{LSTM+GloVe}_{300dims}$ | **0.74** | 0.75 |

Table 4: Effect of Increasing Thematic Embedding Dimensions

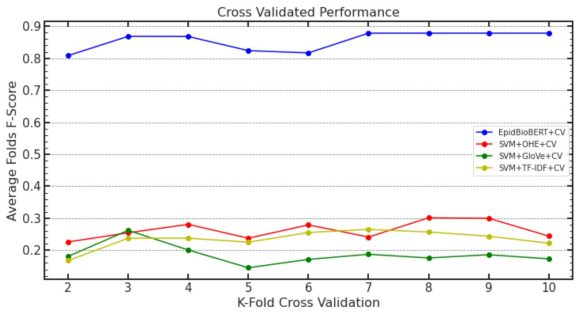


Figure 5: Cross Validated Models

### 5.4. Effect of Epidemiology Data Size

To understand how increase in training data size would impact EpidBioBERT model, we train the cross validated model version (EpidBioBERT+CV) over a series of K-Fold cross validated data set. We set the range values of $K = [2, 3, 4, 5, 6, 7, 8, 9, 10]$ and run our model for 5 epochs using the same architecture and hyperparameters of EpidBioBERT from 2-fold up to 10-fold cross validations. To compare cross validation effect, we also train the cross validated SVM-based baseline versions (SVM+OHE+CV, SVM+GloVe+CV, SVM+TF-IDF+CV) over the same values of $K$ and present average F-Score per fold measures over an unseen test set in Fig. 5.

The results show that higher folds of cross validation improves EpidBioBERT though with a small margin. We report EpidBioBERT+CV standard deviations over different folds in Table 3. The results show the slight variation of both accuracy and F-measure when the data is changed from fold to fold. However we observe that the SVM baselines SVM+OHE+CV and SVM+TF-IDF+CV improve beyond SVM+GloVe+CV for higher K-Fold values. This highlights the ability of EpidBioBERT to learn a lot from a small epidemiology data set as compared to the baseline models as demonstrated by smaller deviations that uniformly holds in all K-Folds.

## 6. Conclusion and Future Work

This paper presents an epidemiological document tagger EpidBioBERT over the infectious animal disease biosurveillance system; PADI-Web. Our contribution improves the classification accuracy by enriching epidemiological thematic features in disease news articles using transfer learning approach on knowledge contained in BioBERT. We found out that biomedical language models contains encoded knowledge for enriching and improving disease surveillance systems.

We however also found out that there is limited annotated news corpora for this task. As future work, we propose an unsupervised pipeline that can take in unlabelled news articles which are much available. To further improve epidemiological relevance corpus classification, we propose an integration of semantic knowledge from relevant sources such as Medical Subject Headings (Mesh) to further enrich thematic features. We also propose to research further on individual impact of thematic features in disease surveillance systems.

## 7. Acknowledgements

This study was partially funded by EU grant 874850 MOOD and is catalogued as MOOD032. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission.

## 8. Bibliographical References


Arsevska, E., Valentin, S., Rabatel, J., de Goër de Hervé, J., Falala, S., Lancelot, R., and Roche, M. (2018). Web monitoring of emerging animal infectious diseases integrated in the french animal health epidemic intelligence system. *PLOS ONE*, 13:1–25, 08.

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

Broscheit, S. (2019). Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China, November. Association for Computational Linguistics.

Broscheit, S. (2020). Investigating entity knowledge in bert with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, Jan.

Huang, K., Altosaar, J., and Ranganath, R. (2020). Clinicalbert: Modeling clinical notes and predicting hospital readmission.

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

Kanakarajan, K. r., Kundumani, B., and Sankarasubbu, M. (2021). BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, Online, June. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization.

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.

Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61.

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

Paquet, C., Coulombier, D., Kaiser, R., and Ciotti, M. (2006). Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Eurosurveillance*, 11(12):5–6.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Phang, J., Févry, T., and Bowman, S. R. (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2020). How to fine-tune bert for text classification?

Valentin, S., Arsevska, E., Falala, S., de Goër, J., Lancelot, R., Mercier, A., Rabatel, J., and Roche, M. (2020a). Padi-web: A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169:105163.

Valentin, S., Arsevska, E., Mercier, A., Falala, S., Rabatel, J., Lancelot, R., and Roche, M., (2020b). *PADI-web: An Event-Based Surveillance System for Detecting, Classifying and Processing Online News*, pages 87–101. 12.

Valentin, S., Arsevska, E., Rabatel, J., Falala, S., Mercier, A., Lancelot, R., and 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*, 13:100357.

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

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

## 9. Language Resource References

Rabatel, Julien and Arsevska, Elena and de Goër de Hervé, Jocelyn and Falala, Sylvain and Lancelot, Renaud and Roche, Mathieu. (2017). *PADI-web corpus: news manually labeled*. CIRAD Dataverse.