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

EpidBioBERT, is a deep biosurveillance epidemiological document tagger for disease surveillance over PADI-Web system. The 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. Our approach proposes a new way to perform epidemiological document classification by enriching 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. We compare our biomedical pre-trained approach with a general language model based model. **EpidBioBERT** achieves an F1-score of 95.5% over an unseen test set, with an improvement of +5.5 points on F1-Score.

Introduction

- 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.2020, Valentin et al.2021].
- 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.
- This study aimed at developing a new thematic embedding based approach for epidemiological corpus classification over tagged news article sources.

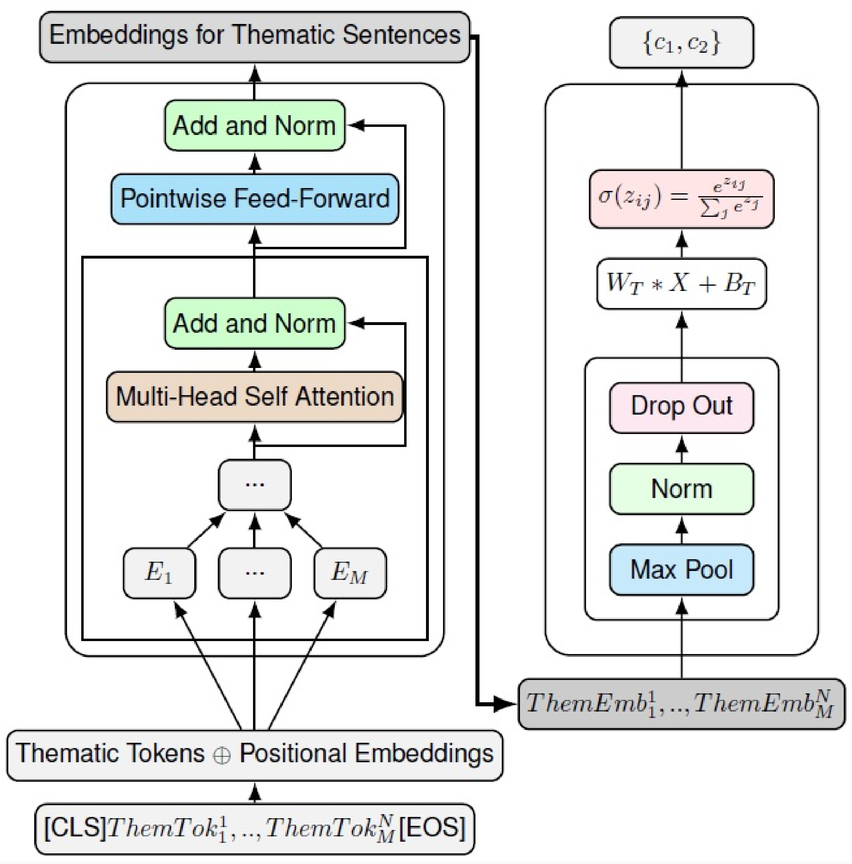
EpidBioBERT Description

- EpidBioBERT model architecture uses a base BioMedical Language model (i.e., BioBERT) with fine tuned disease surveillance deep layers above its architecture.
- EpidBioBERT takes in a set of N *disease outbreak* news articles which we denote as $D = \{d_1, ..., d_N\}$. It operates 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 $c_1 = relevant, c_2 = irrelevant$.
- 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:

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

Model Code: `https://github.com/menya-edmond/EpidBioBERT`

EpidBioBERT Architecture



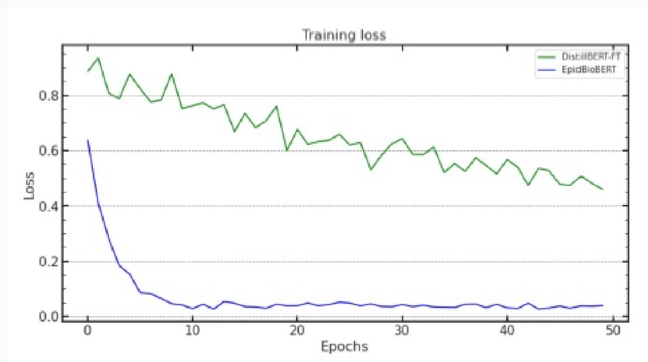
EpidBioBERT Transformer Architecture with fine tuned deep layers on top of pretrained BioBERT. [CLS]$ThemTok_1^1, .., ThemTok_M^N$[EOS] are the N thematic feature tokens from M sentences in the annotated train corpus that are inputs to the model. [CLS] and [EOS] are the tokenizers' tags for start and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

Experimental Results

Model	$F_1 Score$	*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**

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

EpidBioBERT Vs BERT For Epid Classification



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

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

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.

References

References

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

[Valentin et al.2021] 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.

Acknowledgment

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.