EventEpi — A Natural Language Processing Framework for Event-Based Surveillance

Auss Abbood[1], Alexander Ullrich[1], Rüdiger Busche[2], Stéphane Ghozzi[1],

[1]Robert Koch Institute, Berlin, Germany, {abbooda, ullricha, ghozzis}@rki.de, [2] Osnabrück University, Osnabrück, Germany, rbusche@uos.de

Introduction

According to the World Health Organization (WHO), around 60% of all outbreaks are detected using informal sources[3]. In many public health institutes, including the WHO and the Robert Koch Institute (RKI), dedicated groups of epidemiologists sift through numerous public articles and newsletters to detect relevant articles, which is part of event based surveillance (EBS)[4]. To support EBS, we developed a framework that is able to learn how to automatically perform key information extraction of the **disease**, **country**, **date** and **confirmed-case count**, and score the relevance of online articles. We wrapped these functionalities into a web application called *EventEpi.*

[3]https://www.who.int/csr/alertresponse/epidemicintelligence/en/
[4]http://www.wpro.who.int/emerging_diseases/documents/docs/eventbasedsurv.pdf

Methods

- Internal documents were used to extract expert labels.
- Key information extraction was performed by choosing the mode of the country and disease entities. For the date and key entities, a learning-based approach was used.
- We compared tf-idf vectorized and document embedded text document using different classification algorithms.
- We applied class weights and ADASYN upsampling due to class imbalance.
- Word embedding were trained using Wikipedia corpus and 20.000 epidemiological articles.

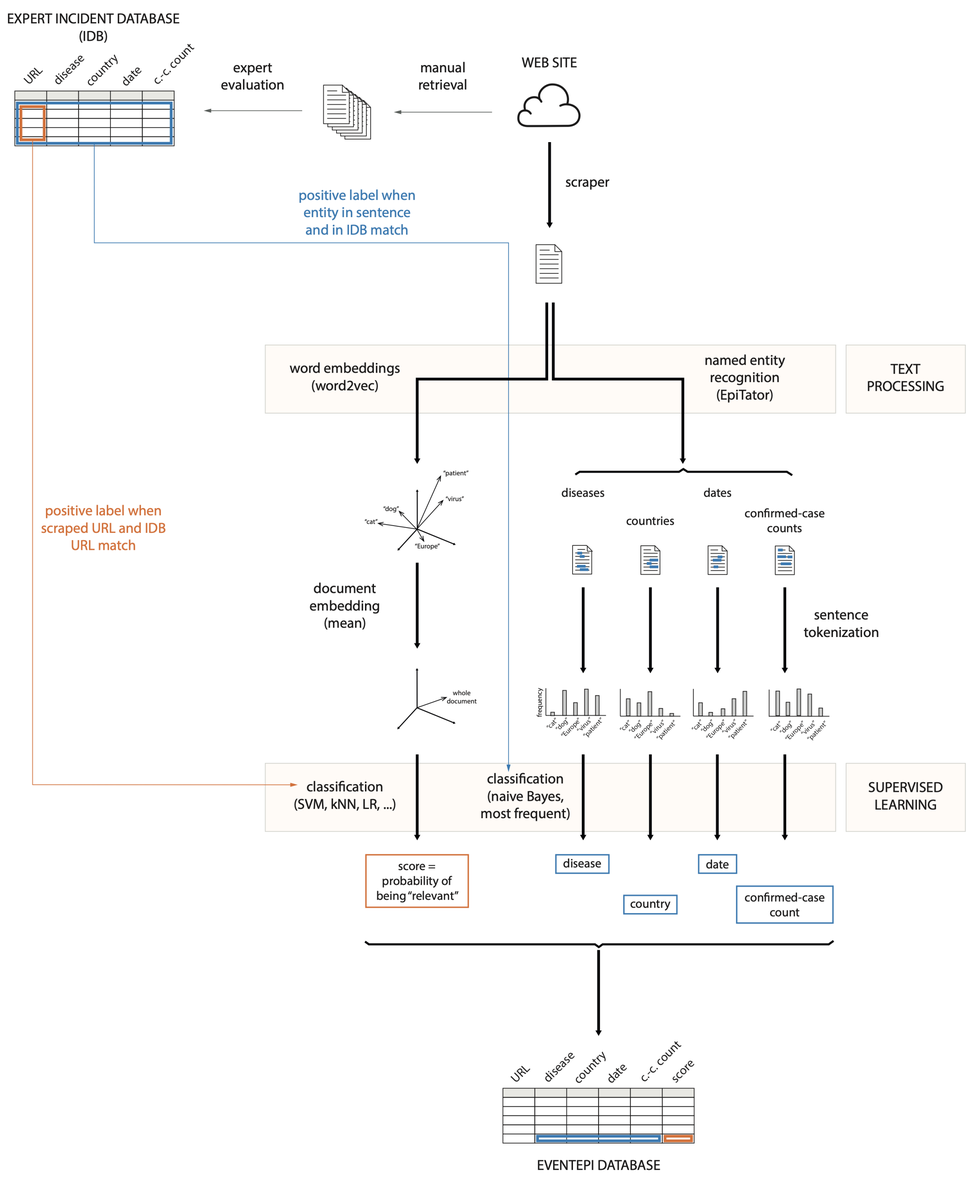

Figure 1: An illustration of the *EventEpi* architecture. Orange refers to relevance scoring, blue to key information extraction.

Results

- Results were acquired using 20% of the data for testing.
- Key information extraction was trained on tf-idf transformed bag-of words.
- The relevance estimation was trained on the average of the document's word embeddings.

Table 1: Performance of the key information extraction using multinomial Naive Bayes classifier.

Entitiy	Prec.	Rec.	Spec.	F1	IBA
Date	**0.50**	0.44	0.87	**0.47**	0.39
Key	0.32	**0.65**	**0.88**	0.43	**0.58**

Table 2: Performance of the relevance classification of articles.

Method	Prec.	Rec.	Spec.	F1	IBA
Multinomial naive Bayes	0.26	0.19	0.97	0.22	0.20
Logistic regression	0.10	0.62	0.72	0.18	0.45
Support-vector machine	0.06	**0.88**	0.30	0.11	0.25
Multilayer perceptron	0.19	0.50	0.89	**0.28**	**0.46**
Convolutional neural network (CNN)	**1.00**	0.14	**1.00**	0.24	0.15

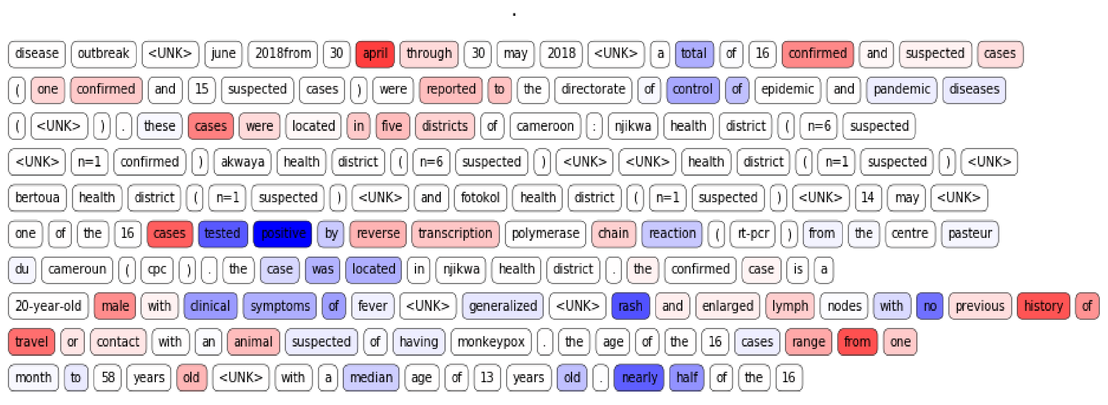

Figure 2: Explainability through layer-wise relevance propagation[5]: Contribution of tokens to the correct classification of a text by a CNN (contributing words are red, dissuading ones are blue).

[5]https://doi.org/10.1371/journal.pone.0181142

Conclusion

We have shown that novel methods of natural language processing can be used to support EBS and public health. However, *EventEpi's* performance needs to be improved. More labeled data, e.g., would be helpful to improve performance for which manual labeling could be an option. Low precision could be due to missed or duplicate relevant articles. Thus, it would be important to revisit false-positives to (if necessary) update their label (relevant or not). Finally, the most interesting sources for EBS are not written in English. Therefore, a future task would be to include also non-English languages to *EventEpi*. Furthermore, we successfully applied a similar approach for the global EIOS platform of WHO.

Acknowledgment and preprint:
We would like to thank Sandra Beerman, Sarah Esquevin and Raskit Lachman for providing data and consulting us on epidemiological topics. More information to this work can be found in our preprint: https://doi.org/10.1101/19006395