Automatic Information Extraction and Relevance Evaluation of Epidemiological Texts Using Natural Language Processing

Auss Abbood, Alexander Ullrich, Rüdiger Busche, Stéphane Ghozzi

ROBERT KOCH INSTITUT

SIGNALE

UNIVERSITÄT OSNABRÜCK

“More than 60% of the initial outbreak reports come from unofficial sources […].”

- https://www.who.int/csr/alertresponse/epidemicintelligence/

Surveillance

Indicator-based Surveillance

- Notifiable diseases
- Laboratory confirmations
- Weekly, monthly reporting

Surveillance

Indicator-based Surveillance

- Notifiable diseases
- Laboratory confirmations
- Weekly, monthly reporting

Event-based Surveillance (EBS)

- Rumors of outbreaks
- Clusters of diseases
- Immediate reporting

http://www.wpro.who.int/emerging_diseases/documents/docs/eventbasedsurv.pdf, p. 4

Possible Examples of EBS Sources

Informal networks

Media

Environmental disaster

Private sectors

Alert networks

NGOs

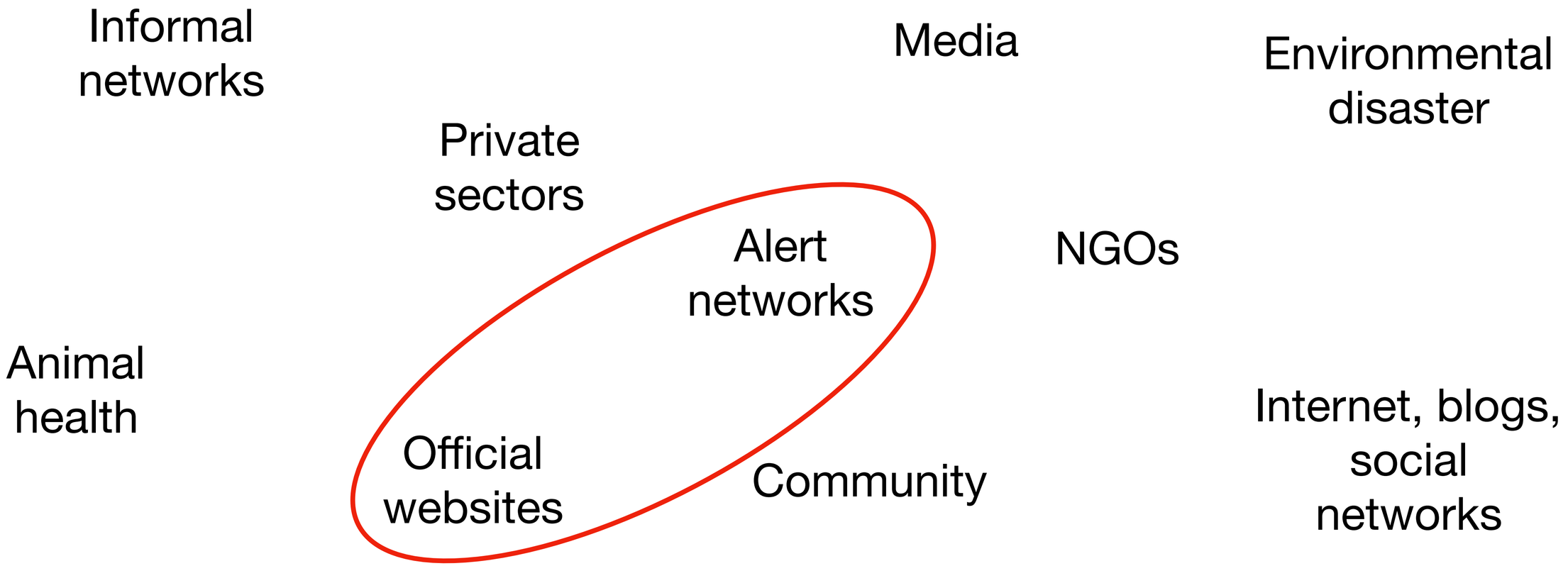
Animal health

Official websites

Community

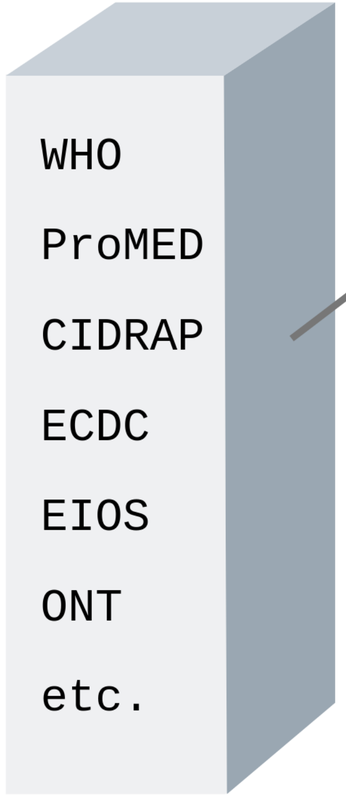
Internet, blogs, social networks

Possible Examples of EBS Sources



How can algorithms help process this data?

Online Articles



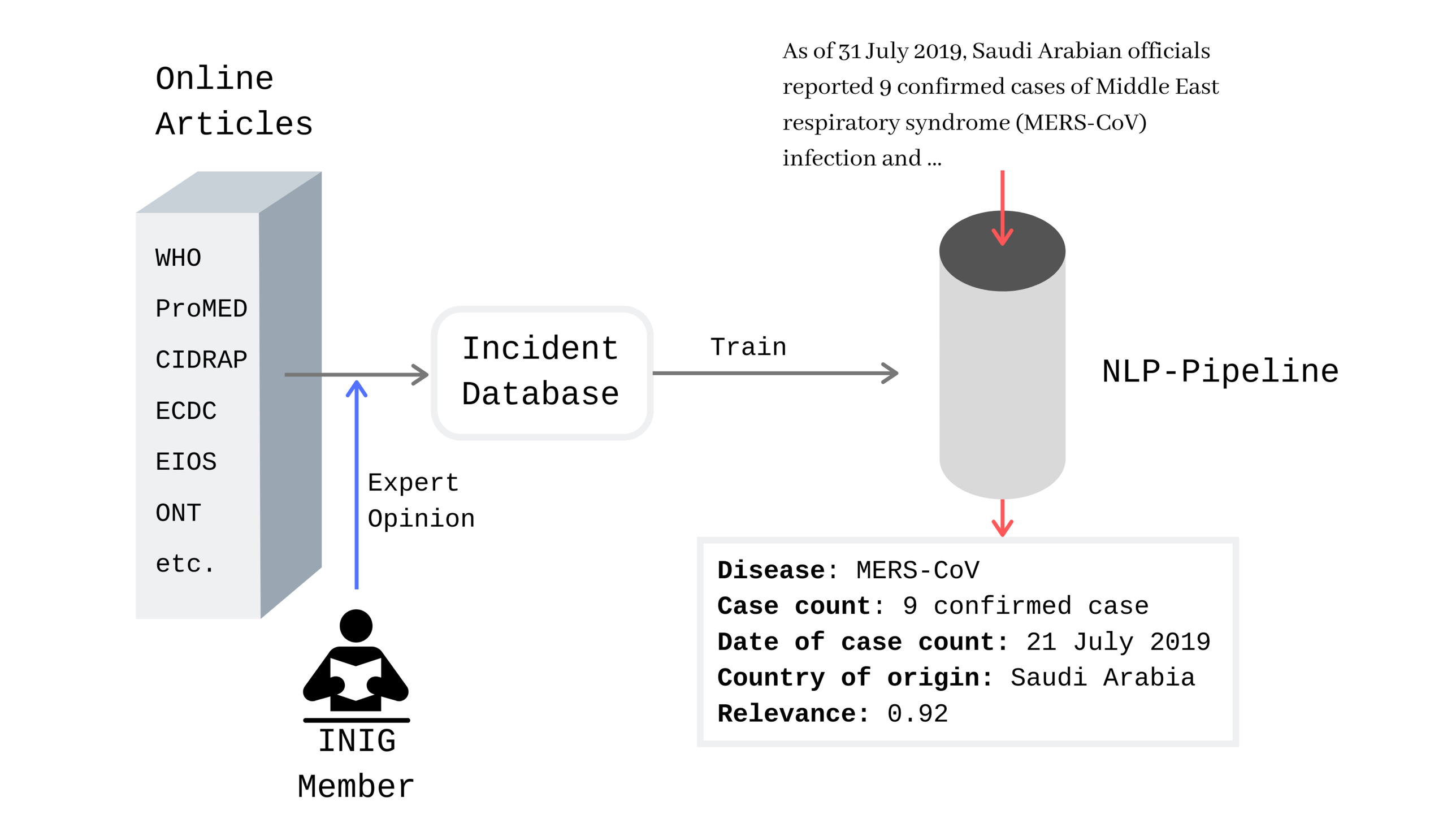
As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and ...

INIG Member

Incident Database

Disease: MERS-CoV
Case count: 9 confirmed case
Date of case count: 21 July 2019
Country of origin: Saudi Arabia

Around 30 articles need to be read every day.


Online Articles
WHO
ProMED
CIDRAP
ECDC
EIOS
ONT
etc.
Incident Database
Train
NLP-Pipeline
Expert Opinion
INIG Member
As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and ...
Disease: MERS-CoV
Case count: 9 confirmed case
Date of case count: 21 July 2019
Country of origin: Saudi Arabia
Relevance: 0.92

Contribution

- Automatically extract key information from epidemiological texts, namely: **disease**, **country** of outbreak, confirmed **case counts**, and **date** of those counts
- Give a relevance score to articles learned from former assessments
- Serve tools in a web application

Key Information Extraction

```
As of 31 July 2019,
Saudi Arabian officials
reported 9 confirmed
cases of Middle East
respiratory syndrome
(MERS-CoV) infection
and four associated
deaths. Those cases
were reported from
Riyadh. One case, who
is confirmed to have
diabetes, seems to have
its origin in Abu
Dhabi.
```

- We used the Python library EpiTator to find important named-entities

Key Information Extraction

As of 31 July 2019, **Saudi Arabian** officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and four associated deaths. Those cases were reported from **Riyadh**. One case, who is confirmed to have diabetes, seems to have its origin in **Abu Dhabi.**

- We used the Python library EpiTator to find important named-entities

Key Information Extraction

As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and four associated deaths. Those cases were reported from Riyadh. One case, who is confirmed to have diabetes, seems to have its origin in Abu Dhabi.

- We used the Python library EpiTator to find important named-entities

Key Information Extraction

As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and four associated deaths. Those cases were reported from Riyadh. One case, who is confirmed to have diabetes, seems to have its origin in Abu Dhabi.

- We used the Python library EpiTator to find important named-entities

Key Information Extraction

As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of **Middle East respiratory syndrome (MERS-CoV)** infection and four associated deaths. Those cases were reported from Riyadh. One case, who is confirmed to have diabetes, seems to have its origin in Abu Dhabi.

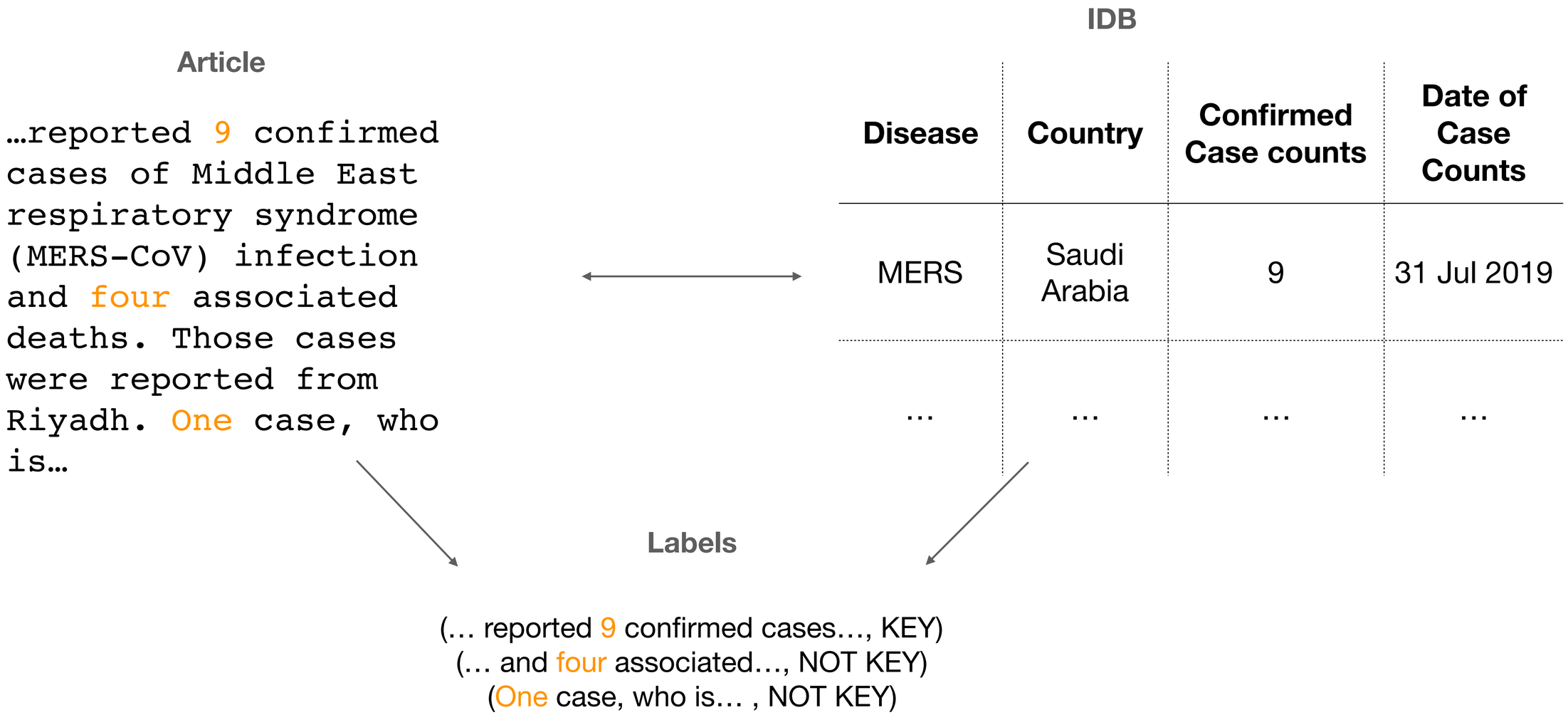
- We used the Python library EpiTator to find important named-entities

How to find the *key* entity*?*

Most-frequent approach and learning-based approach

- To find the *key* information, we used the most frequent occurrence in an entity class (**most-frequent approach**)
- This did not work well for the case count and date
- Use IDB as labels for machine learning algorithms (**learning-based approach**)

Key Information Extraction



Key Date Extraction

	Pre.	Rec.	Spec.	F1	IBA	Sup Key	Sup Not Key
Multinomial Naive Bayes	0.32	0.65	0.88	0.43	0.58	40	449
Bernoulli Naive Bayes	0.28	0.55	0.88	0.37	0.49	40	449

Key Count Extraction

	Pre.	Rec.	Spec.	F1	IBA	Sup Key	Sup Not Key
Multinomial Naive Bayes	0.50	0.44	0.87	0.47	0.39	9	39
Bernoulli Naive Bayes	0.43	0.33	0.87	0.38	0.30	9	39

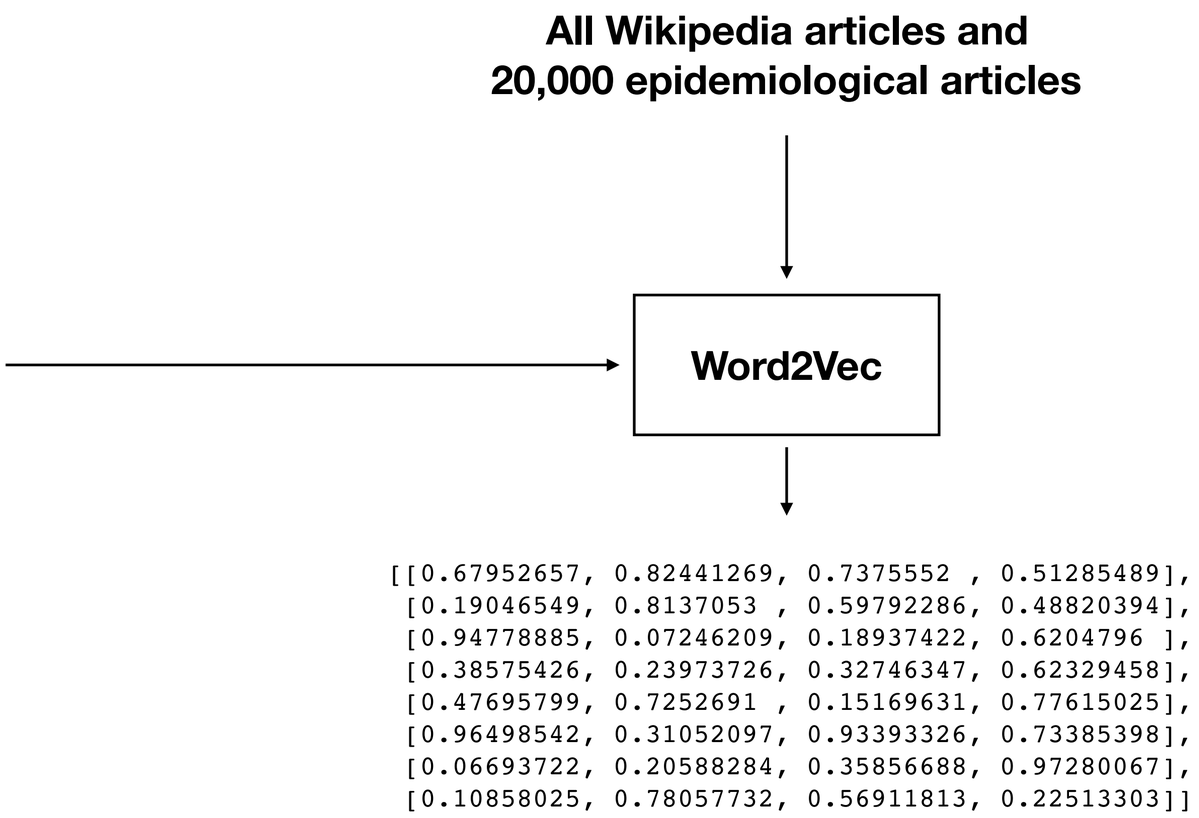
How do we find the relevance of an article?

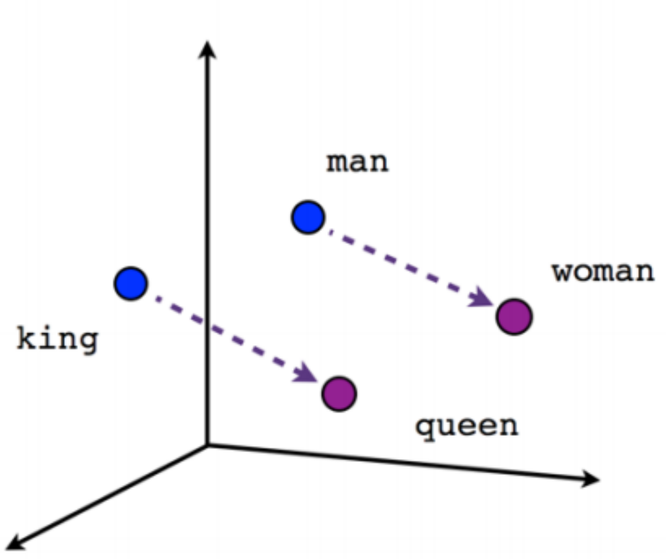
Relevance Evaluation

- Assumption: If an article is in the IDB, it is **relevant.** If it was read but not entered into the IDB it is **not relevant**
- Scrape INIGs “main” sources (WHO Disease Outbreaks News and ProMED Mail) and label them
- Train a machine learning algorithm to detect the relevance of an unseen article

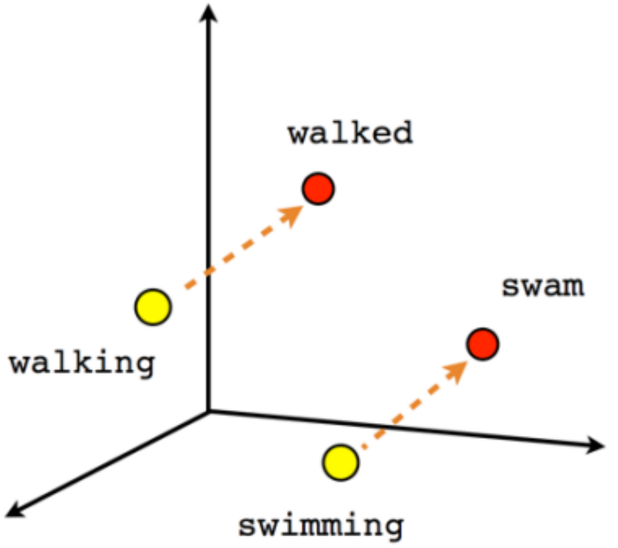
Relevance Evaluation

As of 31 July 2019, Saudi Arabian officials reported 9 confirmed cases of Middle East respiratory syndrome (MERS-CoV) infection and four associated deaths. Those cases were reported from Riyadh. One case, who is confirmed to have diabetes, seems to have its origin in Abu Dhabi.

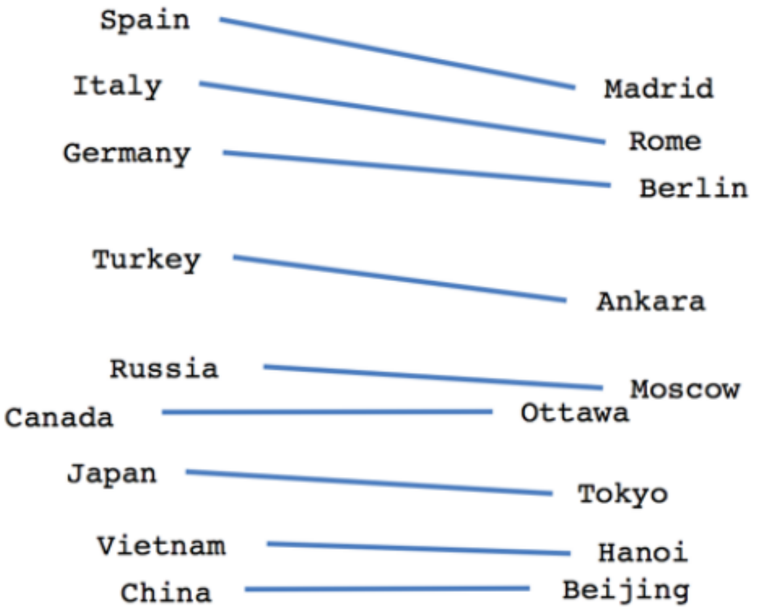




Male-Female



Verb tense



Country-Capital

- https://towardsdatascience.com/deep-learning-4-embedding-layers-f9a02d55ac12

	Pre.	Rec.	Spec.	F1	IBA	Sup, relevant	Sup. irrelevant
Multinomial naive Bayes	0.26	0.19	0.97	0.22	0.20	32	615
Complement naive Bayes	0.26	0.19	0.97	0.22	0.20	32	615
Logistic regression	0.10	0.62	0.72	0.18	0.45	32	615
k-nearest neighbor classifier	0.07	0.69	0.55	0.13	0.38	32	615
Support-vector machine	0.06	0.88	0.30	0.11	0.25	32	615
Multilayer perceptron	0.19	0.50	0.89	0.28	0.46	32	615
Convolutional neural network	1.00	0.14	1.00	0.24	0.15	36	611

EventEpi

Enter an URL : SUMMARIZE

Get WHO DONs | Get Promed Articles | Copy | CSV | Excel | PDF | Print

Search:

Disease	Country	Confirmed Cases	Date Of Case Count	Relevance	Input Date	Source
Ebola hemorrhagic fever	Democratic Republic of the Congo	312	05, March, 2019	0.44	2019-Mar-14	https://www.who.int/csr/don/7-march-2019-ebola-drc/en/
Lassa fever	Federal Republic of Nigeria	5	14, February, 2019	0.72	2019-Mar-7	https://www.who.int/csr/don/14-february-2019-lassa-fever-nigeria/en/
poliomyelitis	Independent State of Papua New Guinea	1369	01, January, 2005	0.9	2019-Mar-7	https://www.who.int/csr/don/27-february-2019-polio-indonesia/en/

Showing 1 to 3 of 3 entries

Previous 1 Next

Thanks to Sandra Beermann, Sarah Esquevin, and Raskit Lachmann from the INIG unit who provided us with data and consultation!

ROBERT KOCH INSTITUT

SIGNALE

UNIVERSITÄT OSNABRÜCK

Contact: abbooda@rki.de

Source code: https://github.com/aauss/EventEpi

Preprint: https://doi.org/10.1101/19006395

What we do: www.rki.de/signale-project

ROBERT KOCH INSTITUT

SIGNALE

UNIVERSITÄT OSNABRÜCK