AN INITIATIVE OF THE NETHERLANDS RED CROSS

Text Mining of Loss Data

Jacopo Margutti, PhD jmargutti@redcross.nl
Marc van den Homberg, PhD mvandenhomberg@redcross.nl

04 May 2020

CC BY

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

OUTLINE

WHY TEXT MINING

HOW IT WORKS

HOW WELL IT WORKS

WHO WORKS ON IT

CONCLUSIONS & OUTLOOK

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

OUTLINE

WHY TEXT MINING

HOW IT WORKS

HOW WELL IT WORKS

WHO WORKS ON IT

CONCLUSIONS & OUTLOOK

WHY TEXT MINING

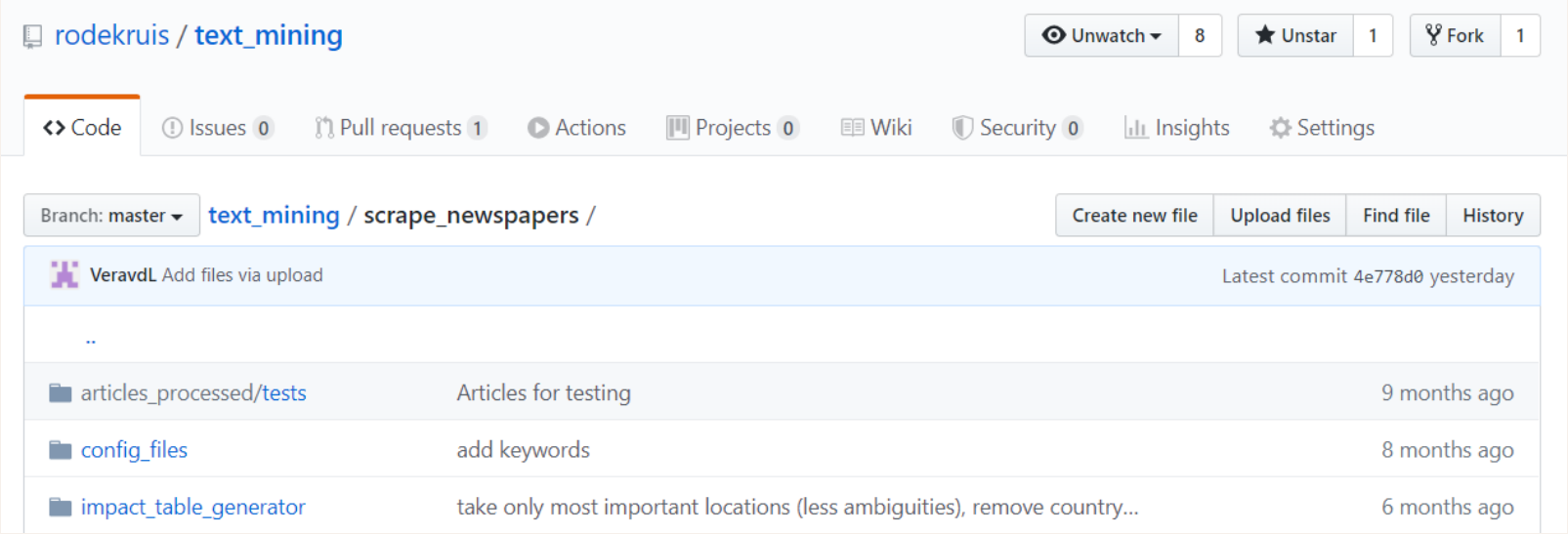
- **Loss and damage data**, i.e. data on the impact of natural disaster, is key information for **disaster preparedness programs** and **risk assessments.**
- Data mostly collected **manually** by agencies and governments
  - Governments: data often not accessible (e.g. only in paper format)
  - Humanitarian organizations: often only about intervention areas, no standard format
  - Private sector (e.g. insurance companies): proprietary data, possible biases
- Lots of information available in **local** and national **newspapers**, but highly **unstructured**

→ **TEXT MINING: extract structured loss data from news articles**

THE NEWS SCRAPER

510, the data team of the Netherlands Red Cross, has been developing an **open-source text mining framework** for loss data: the "***news scraper***"

https://github.com/rodekruis/text_mining/tree/master/scrape_newspapers



WHAT IS THE NEWS SCRAPER

The news scraper:

- input: disaster type and country
- output: loss data in tabular format, from local and national newspapers
- multilingual support (using spaCy's NLP models)
- mostly automated

Search:
"floods in Uganda"

type	where	when	...
flood	West Pokot	12 Aug 2002	...
drought	Karamoja	Sept 2005	...

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

OUTLINE

WHY TEXT MINING

HOW IT WORKS

HOW WELL IT WORKS

WHO WORKS ON IT

CONCLUSIONS & OUTLOOK

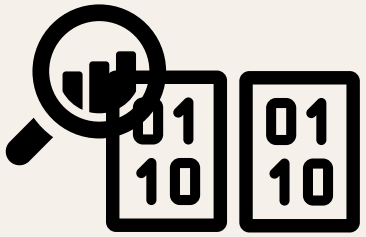
HOW IT WORKS

Three steps:

1) SCRAPE ARTICLES

2) ASSESS ARTICLE RELEVANCE

3) EXTRACT LOSS DATA



SCRAPE ARTICLES

1) SCRAPE ARTICLES:

- get a list of local and national news agencies (source: ABYZ)
- access all news agencies' websites and
  - search for articles about disasters
  - download results

ASSESS ARTICLE RELEVANCE

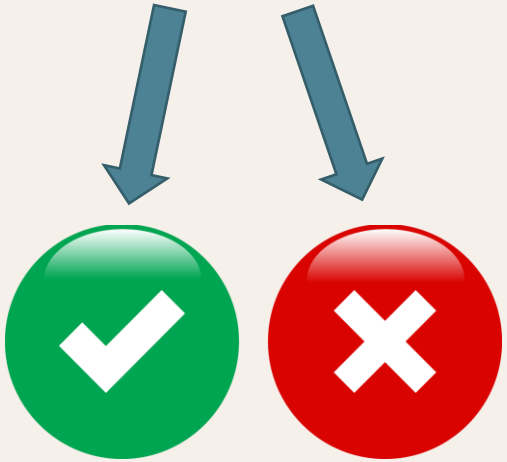
2) ASSESS ARTICLE RELEVANCE:

- to determine if an article is relevant or not
- two possible approaches:
  1. automated classification based on keywords
  2. manual tagging with an interactive prompt

```
Article #1/295: CANIVEAUX BOUCHES : Des risques d'inondations dans la capitale
Is it topical? t (True), f (False), i (Inspect text), q (Quit)  i
CANIVEAUX BOUCHES : Des risques d'inondations dans la capitale

La ville de Bamako reçoit ces jours-ci de fortes pluies alors que presque tous
 ses caniveaux sont obstrués. Un risque d'inondations pointe à l'horizon avec
le mois d'août ou « thio-thio kalo« (mois de grandes précipitations) qui s'ann
once. Il y a urgence.

L'obstruction des caniveaux est un problème sérieux à Bamako surtout en cette
période hivernale. Le peu de caniveaux dans la ville sont bouchés par des ordu
res de toutes sortes. La ville est plus que jamais exposée à un risque d'inond
ations.
```



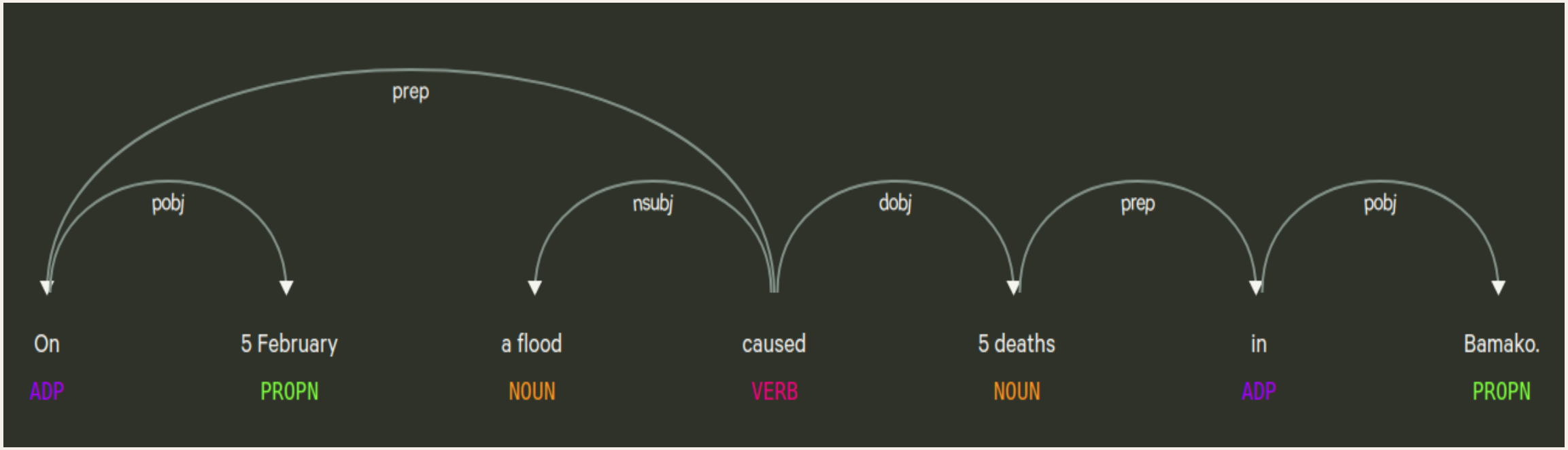
EXTRACT LOSS DATA

3) EXTRACT LOSS DATA

- based on **Natural Language Processing** (NLP) algorithms and models
- Built in python on top of the Spacy framework spaCy
- Step-wise approach:
  1. Tokenize text (divide in words)
  2. Create **syntax trees** (a.k.a. *parse trees*)
  3. Navigate syntax trees and extract loss data

SYNTAX TREE: EXAMPLE

"On 5 February a flood caused 5 deaths in Bamako"



EXTRACT LOSS DATA: EXAMPLE

EXAMPLE ARTICLE

Title: Floods in Bamako, Date: 12/04/2001

Last week, floods in the capital affected 500 people and killed 7.

There were also 3 victims in Sirakoro and 4 in Kotuba.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.1: Assign article location

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital affected 500 people and killed 7.

There were also 3 victims in **Sirakoro** and 4 in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.1: Assign article location

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital affected 500 people and killed 7.

There were also 3 victims in **Sirakoro** and 4 in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.2: Look for numbers

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital affected **500** people and killed **7**.

There were also **3** victims in **Sirakoro** and **4** in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.3: Assign impact and location

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital **affected 500** people and **killed 7**.

There were also **3 victims** in **Sirakoro** and **4** in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.3: Assign impact and location

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital **affected 500** people and **killed 7**.

There were also **3 victims** in **Sirakoro** and **4** in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.3: Assign impact and location

Title: Floods in **Bamako**, Date: 12/04/2001

Last week, floods in the capital **affected 500** people and **killed 7**.

There were also **3 victims** in **Sirakoro** and **4** in **Kotuba**.

EXTRACT LOSS DATA: EXAMPLE

STEP 3.4: Output

type	where	when	affected people	casualties
flood	Bamako	12 Apr 2001	500	7
flood	Sirakoro	12 Apr 2001		3
Flood	Kotuba	12 Apr 2001		4

Loss data categories: **people affected, casualties, houses affected, infrastructures affected**

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

OUTLINE

WHY TEXT MINING

HOW IT WORKS

HOW WELL IT WORKS

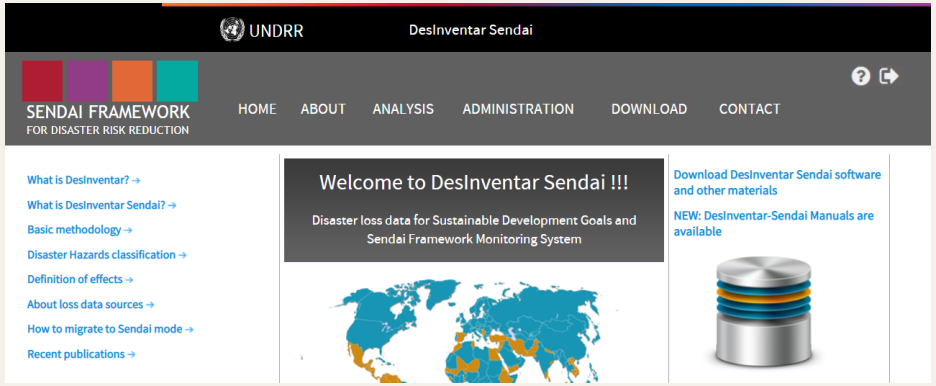
WHO WORKS ON IT

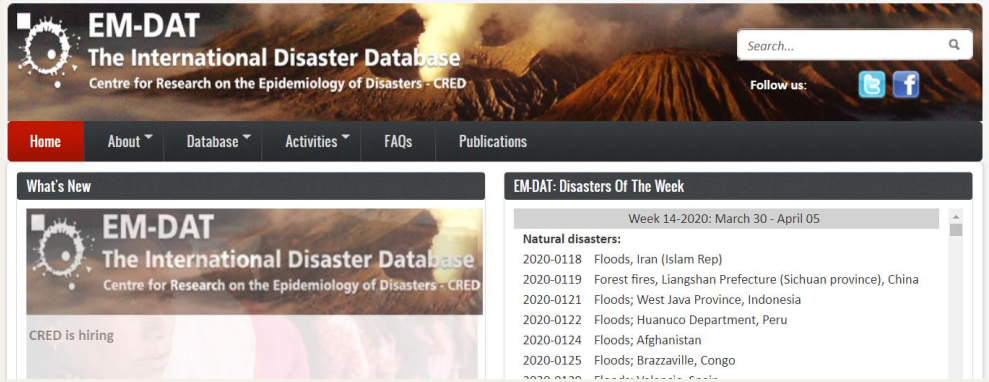
CONCLUSIONS & OUTLOOK

MODEL VALIDATION

Model validation:

- Using as reference two loss databases: DesInventar (UN) and EM-DAT (CRED / UCLouvain)

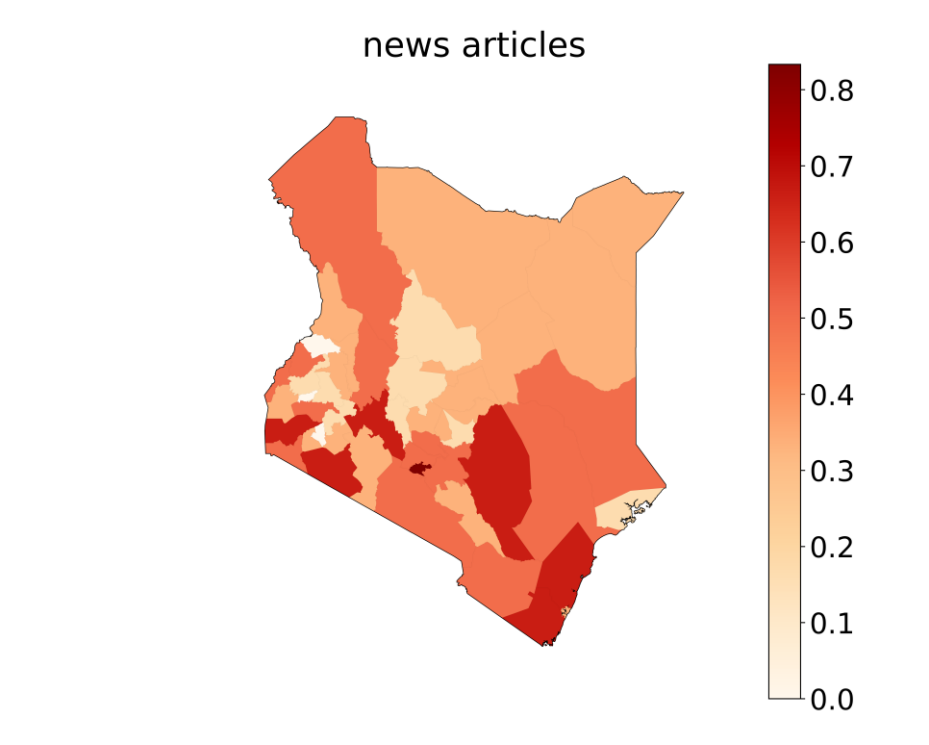


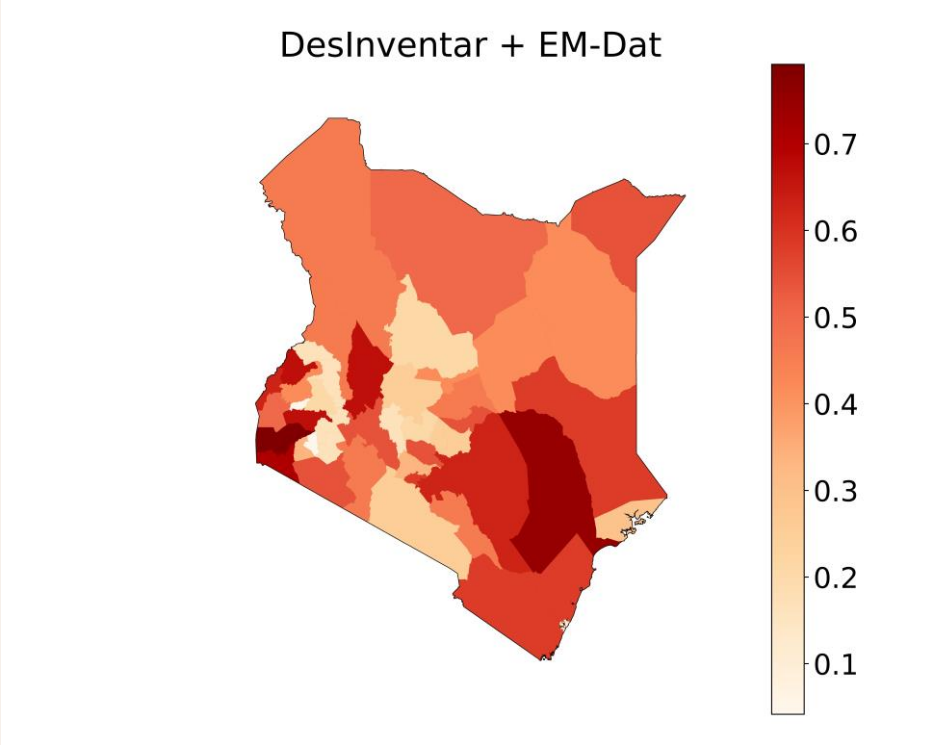


- Performance indicator: annual incidence rate of disaster at sub-national level
  - Annual incidence is defined as 1 or more occurence(s) of disaster per year
  - first administrative level considered
- Metric: root mean squared error (RMSE) on annual incidence rate

CASE STUDY #1

Country: Kenya, disaster: flood, language: English



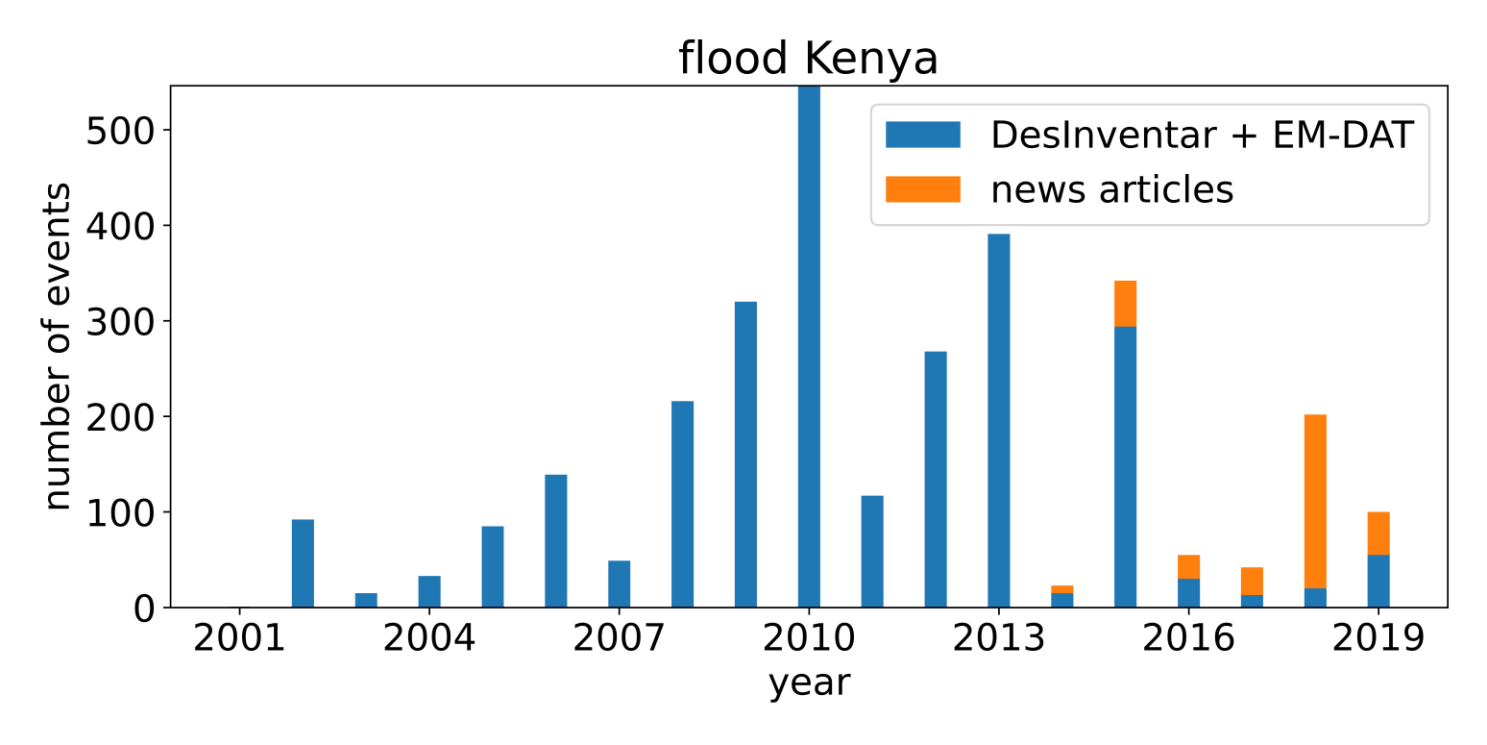


annual incidence rate

- 145 news articles, 337 unique loss data entries
- News articles vs DesInv.+EM-DAT: RMSE 0.14

CASE STUDY #1

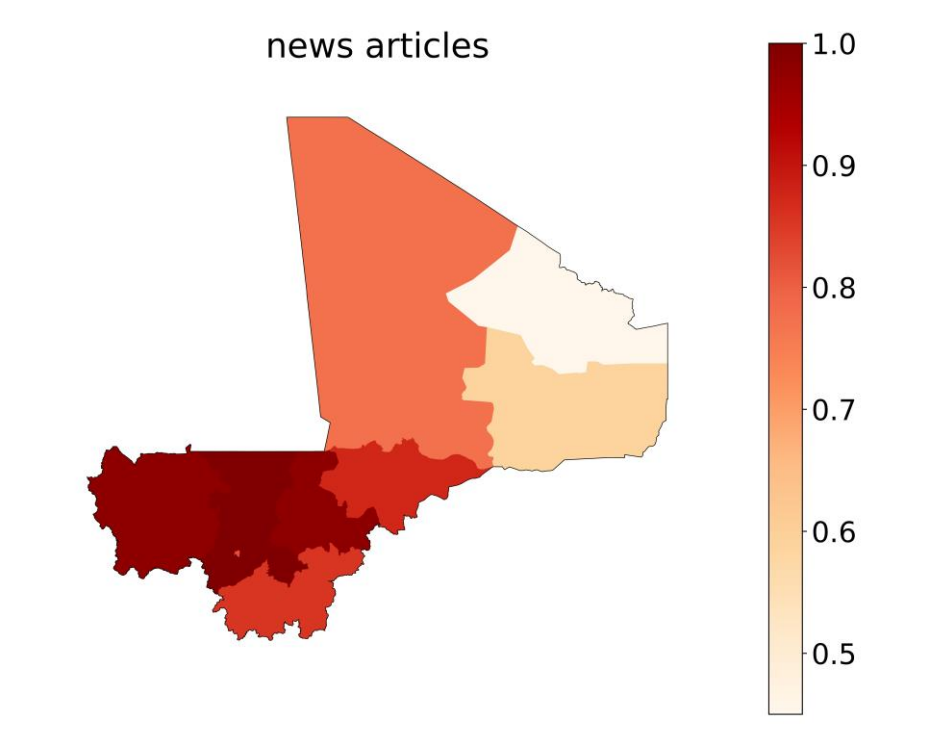
Country: Kenya, disaster: flood, language: English

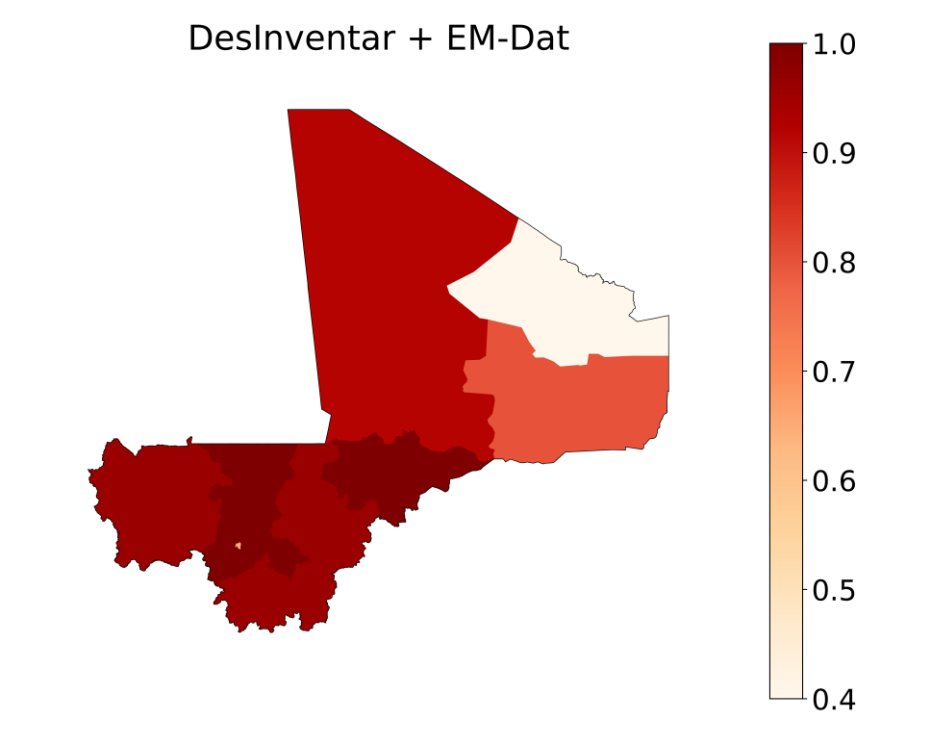


- News articles **only from 2014**
- From 2016, **more data from** news articles than from DesInv.+EM-DAT

CASE STUDY #2

Country: Mali, disaster: flood, language: French



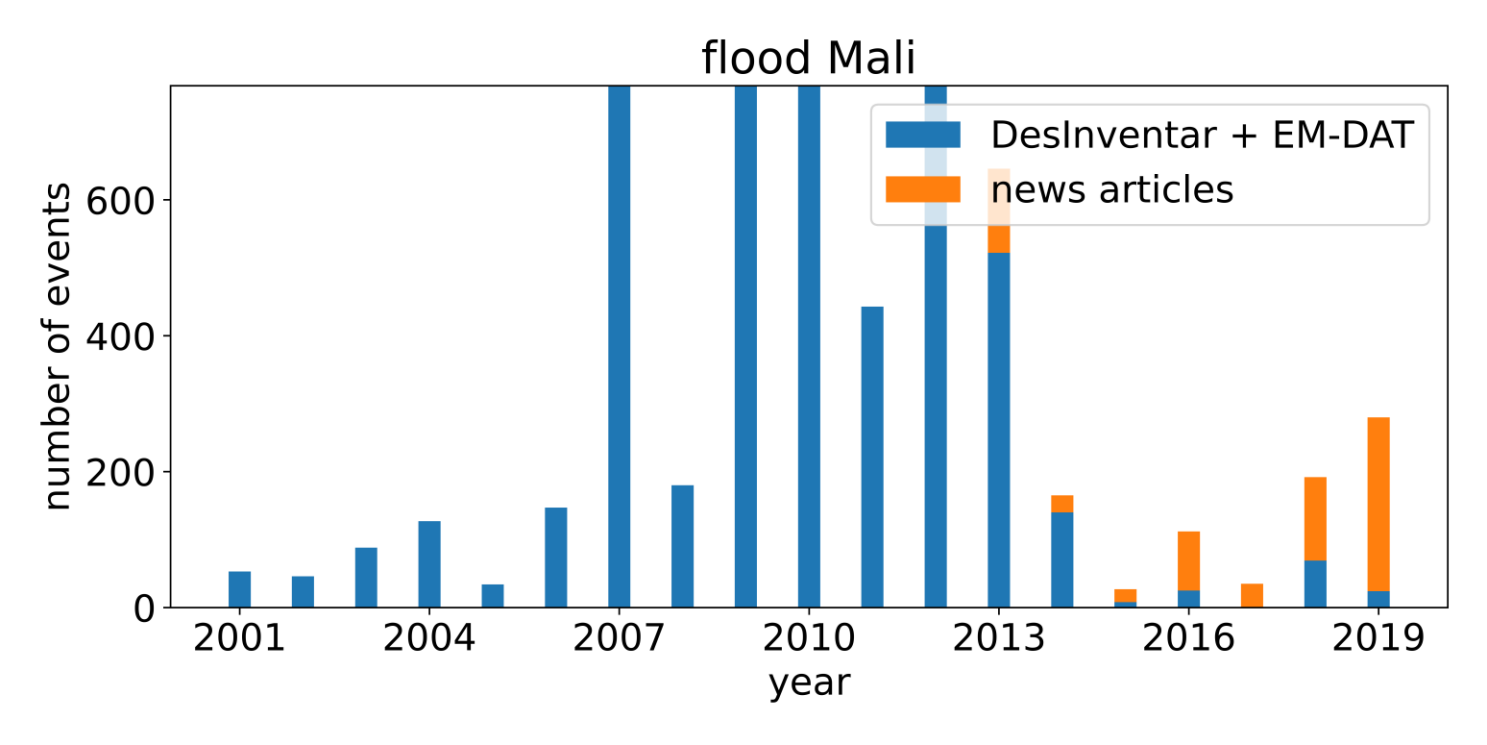


annual incidence rate

- 520 news articles, 682 unique loss data entries
- News articles vs DesInv.+EM-DAT: RMSE 0.13

CASE STUDY #2

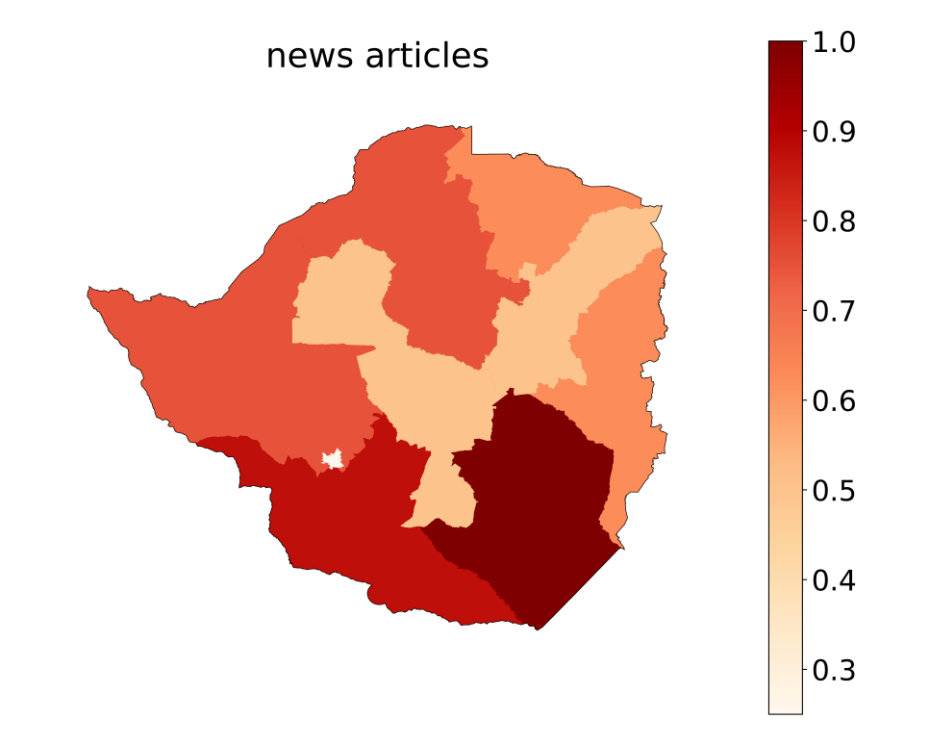
Country: Mali, disaster: flood, language: French

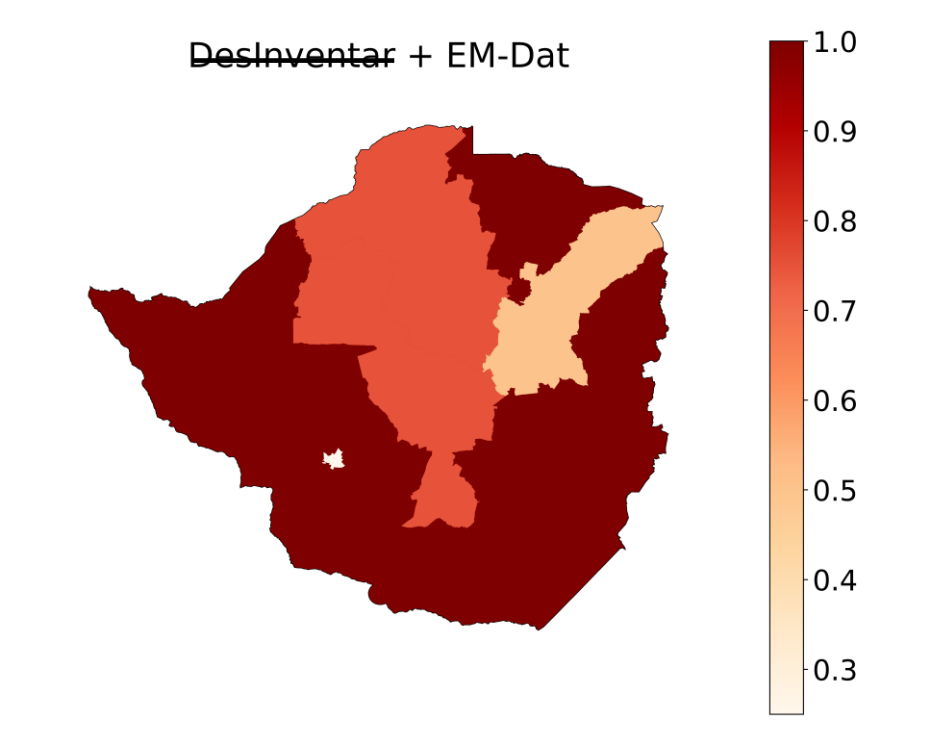


- News articles **only from 2012**
- From 2014, **more data from** news articles than from DesInv.+EM-DAT

CASE STUDY #3

Country: Zimbabwe, disaster: drought, language: English



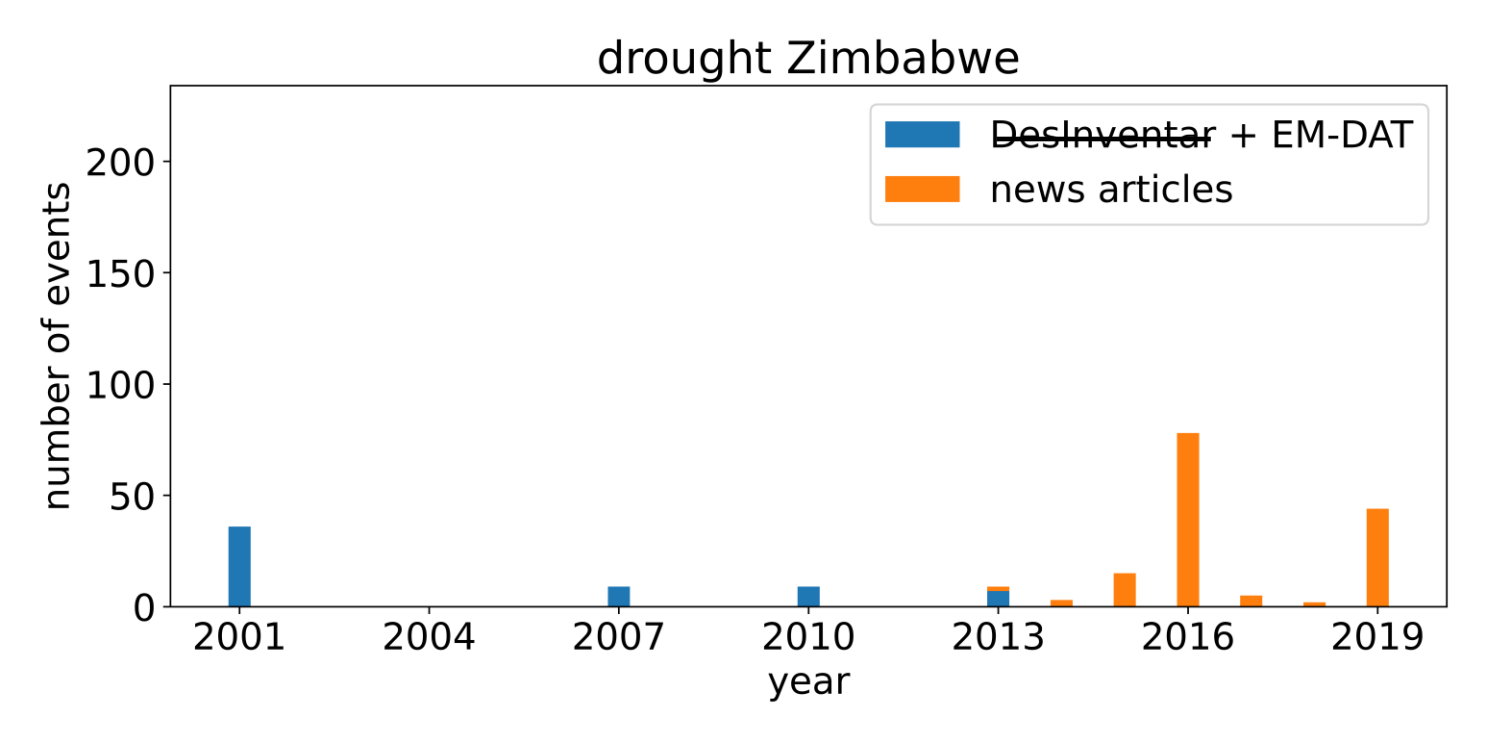


annual incidence rate

- 176 news articles, 171 unique loss data entries
- News articles vs EM-DAT (no data on DesInventar): RMSE 0.22

CASE STUDY #2

Country: Zimbabwe, disaster: drought, language: English



- News articles **only from 2013**
- From 2013, **more data from** news articles than from EM-DAT

DISCUSSION

Results:

- Reasonable **agreement** between loss data from news articles and other databases*
- **Large amount** of data available from news, but **only from the last 6-8 years**

Main **limitations**:

- Multiple articles about same event → possible double-counting
- Bias towards major cities: news agencies are more likely to report disasters in cities than rural areas
- Article date used as event date, not always accurate
- Location not always accurate in complex sentences

*measured as RMSE on annual incidence rate

CONCLUSIONS AND OUTLOOK

- We developed an **open-source framework** for **text mining** of loss data: the ***news scraper***
- Mostly automated, highly customizable, multilingual support
- **Valid complement** to existing loss data sources
- Next steps:
  - Extend validation
  - Automated classification of relevance
  - Improve loss data extraction: merge multiple events, date parsing, solve location ambiguities

CONTRIBUTORS

The *news scraper* is a **collaborative** work of professional **volunteers** and **students** at 510 Netherlands Red Cross!

Many thanks to

- Bonnie van Vuure
- Monica Turner
- Wessel de Jong