Text Mining of Loss Data

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

04 May 2020
OUTLINE

WHY TEXT MINING
HOW IT WORKS
HOW WELL IT WORKS
WHO WORKS ON IT
CONCLUSIONS & OUTLOOK
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
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
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

<table>
<thead>
<tr>
<th>type</th>
<th>where</th>
<th>when</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>flood</td>
<td>West Pokot</td>
<td>12 Aug 2002</td>
<td>…</td>
</tr>
<tr>
<td>drought</td>
<td>Karamoja</td>
<td>Sept 2005</td>
<td>…</td>
</tr>
</tbody>
</table>
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
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
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
3) EXTRACT LOSS DATA

- based on **Natural Language Processing** (NLP) algorithms and models
- Built in python on top of the **Spacy** framework
- 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
"On 5 February a flood caused 5 deaths in Bamako"
LAST WEEK, FLOODS IN THE CAPITAL AFFECTED 500 PEOPLE AND KILLED 7.

THERE WERE ALSO 3 VICTIMS IN SIRAKORO AND 4 IN KOTUBA.
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.
STEP 3.1: Assign article location

Title: Floods in \textbf{Bamako}, Date: 12/04/2001

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

There were also 3 victims in \textbf{Sirakoro} and 4 in \textbf{Kotuba}.
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.
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.
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.
**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**.
**STEP 3.4: Output**

<table>
<thead>
<tr>
<th>Type</th>
<th>Where</th>
<th>When</th>
<th>Affected People</th>
<th>Casualties</th>
</tr>
</thead>
<tbody>
<tr>
<td>flood</td>
<td>Bamako</td>
<td>12 Apr 2001</td>
<td>500</td>
<td>7</td>
</tr>
<tr>
<td>flood</td>
<td>Sirakoro</td>
<td>12 Apr 2001</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Flood</td>
<td>Kotuba</td>
<td>12 Apr 2001</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

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

WHY TEXT MINING

HOW IT WORKS

HOW WELL IT WORKS

WHO WORKS ON IT

CONCLUSIONS & OUTLOOK
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 occurrence(s) of disaster per year
  • first administrative level considered

• Metric: root mean squared error (RMSE) on annual incidence rate
Country: Kenya, disaster: flood, language: English

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

- 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

- 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
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.
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