insight into value
ADVERSE DRUG REACTIONS ON SOCIAL MEDIA: BIAS AND LIMITATION

PhUSE EU Connect 2018 – Paper DH05
AGENDA

Context
- Context
- Previous paper
- Goal

Method
- Data sources
- Languages
- Metadata

Results & Discussion
- Results
- Discussion
AGENDA

Context
- Context
- Previous paper
- Goal

Method
- Data sources
- Languages
- Metadata

Results & Discussion
- Results
- Discussion
CONTEXT – PREVIOUS PAPER

“Patient-generated Health Data (Social Media) is a Potential Source for ADR Reporting”

**Goal**
- Data extraction
- ADR* coding
- Storage

*Adverse Drug Reaction

**Sources**
- **Twitter®**, a micro-blogging social network
- **Reddit®**, a widely used forum

ADR detection possible but subject to limitation:
1) type of sources  
2) languages  
3) method
PROCESS FLOW

Drugs to search

Data extraction

Database storage

Data cleaning

ADRs and drugs coding
CONTEXT - GOAL

Find & Estimate

Bias

Limitations

Of doing ADRs detection on social media

By collected data using:

Source: classic social media and health related websites

Additional language support: using more than just English

Better understand extracted data
Prepare machine learning methods
AGENDA

Context

• Context
• Previous paper
• Goal

Method

• Data sources
• Languages
• Metadata

Results & Discussion

• Results
• Discussion
DATA SOURCES 1/2

- **Twitter®,** a micro-blogging social network
- **Reddit®,** a widely used forum
- **Carenity®,** an online patient community for patient and caregiver

Are all data sources equal?
DATA SOURCES 2/2

- **12th most used website in the world**

- Used to send short messages (280 characters)

- Used to share news, pictures and everyday life

- **9th most used website in the world**

- Used for online discussion

- Contains various sections from video games to drugs usage

- **Health-related forum**

- Used to share disease experience with other patient

- Contains 1,000+ pathologies
Languages used for extraction:

- English
- French
- Spanish
- Portuguese

32% of the world's population

Language influence on ADRs detection
METADATA

**Metadata**: data that provides information about other data.

**Carenity®**:

**Gender**

**Age**

**Reddit®**:
- *None available* for anonymization purpose

**Twitter®**:
- *Not trustworthy* and difficult to exploit

How can we use metadata for ADRs detection
AGENDA

Context

• Context
• Previous paper
• Goal

Method

• Data sources
• Languages
• Metadata

Results & Discussion

• Results
• Discussion
# RESULTS

<table>
<thead>
<tr>
<th>Period</th>
<th>Twitter</th>
<th>Reddit</th>
<th>Carenity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>1 year</td>
<td>2007 - 2018</td>
<td>2011 - 2018</td>
</tr>
<tr>
<td># records</td>
<td>352,285</td>
<td>163,426</td>
<td>12,637</td>
</tr>
</tbody>
</table>

**Disclaimer:**
The data contained in this paper was obtained for demonstration purposes only using the techniques presented. Anyone analyzing data from social media sites, either public or membership based, should investigate the source terms and conditions relevant to data reuse and obtain any necessary permissions if required.

*Detection ~40% of records with potential ADR*

*Potential ADR = ADR and other medical events*
• **Not similar** data repartition

• **Aspirin** and **Ibuprofen** over representation on Twitter

• **Better** drugs representation on forums

All data sources are not equal
# RECORDS AND POTENTIAL ADR RATIO ON TWITTER®

## All drugs

- **35%**

## Without Aspirin and Ibuprofen

- **40%**
# RECORDS AND POTENTIAL ADR RATIO ON REDDIT®

![Graph showing the relationship between the number of extracted records and potential ADRs on Reddit. The graph includes several medications like aspirin, ibuprofen, clonazepam, gabapentin, baclofen, levothyroxine, duloxetine, and adalimumab. The graph highlights a 74% ratio.](image-url)
# RECORDS AND POTENTIAL ADR RATIO ON CARENITY®

108%
# RECORDS AND POTENTIAL ADR RATIO

- **Ratio** very different depending on source
- **Forums** = higher ratio
- Twitter noisier

Forums carry more valuable information
# POTENTIAL ADR RATIO ON TWITTER® (FRENCH & PORTUGUESE RECORDS)

**Low ratio** (potential ADR/records) in French: ~5 to 35%

- Language complexity
- French market oriented

**Low ratio** (potential ADR/records) in Portuguese: ~25 to 40%

Languages difference
**Detected ADR are not similar:**

- French first potential ADR: Pregnancy
- First potential ADR for other languages: Pain

→ Cultural and not grammatical characteristic
MEDIA IMPACT IN FRENCH

- Twitter: *lots of Levothyroxine* records
- Carenity: *high peak* of levothyroxine mention in 2017

→ Media impact

All sources subject to media impact
POTENTIAL ADRS DETECTION BY AGE AND GENDER ON CARENITY

Close but *not similar* distribution between *men* and *women*

ADRs and pathologies not similar between gender
AGE & GENDER REPARTITION ON CARENITY

- **Consistent** age distribution
- **Women**: 80% of the records
- Women ~35yo mention more ADRs

Women more willingly to share their experiences
## REAL ADR VS CONDITION (SUBSET FOR METHOTREXATE)

<table>
<thead>
<tr>
<th></th>
<th>Reddit® En</th>
<th>Twitter® En</th>
<th>Carenity® En</th>
<th>Carenity® Fr</th>
</tr>
</thead>
<tbody>
<tr>
<td># detected ADR</td>
<td>318</td>
<td>318</td>
<td>318</td>
<td>318</td>
</tr>
<tr>
<td># real ADR % of the total</td>
<td>31 10%</td>
<td>17 5%</td>
<td>54 17%</td>
<td>23 7%</td>
</tr>
<tr>
<td># ADR missed % of missed</td>
<td>6 16%</td>
<td>4 19%</td>
<td>4 7%</td>
<td>32 58%</td>
</tr>
</tbody>
</table>

- **Carenity** and **Reddit** in **English**: most accurate sources
- **Twitter**: not a lot of useful information
- **Carenity** in **french**: lot of ADR missed
  - Language barrier

**Forums seems more valuable, especially health-related ones**
CONCLUSION

Objectives

Detect bias and limitation by:
- sources number
- number of languages

Results

Sources quality
Languages specification
Gender influence

Next steps

method
false positives
false negatives
Erwan Le Covec
Lise Radoszycki
Stéphane Chollet

Keyrus Biopharma