insight into value
PATIENT-GENERATED HEALTH DATA (SOCIAL MEDIA) IS A POTENTIAL SOURCE FOR ADVERSE DRUG REACTION REPORTING

PhUSE 2017 – Paper DH01
AGENDA

Context
- Context
- Goal
- Sources

Proposed solution
- Process flow
- Extraction
- Coding
- Storage

Results & Discussion
- Results
- Discussion
AGENDA

Context
- Context
- Goal
- Sources

Proposed solution
- Process flow
- Extraction
- Coding
- Storage

Results & Discussion
- Results
- Discussion
CONTEXT - PHARMACOVIGILANCE

Pharmacovigilance, the science related to the collection, detection, assessment, prevention, and monitoring of Adverse Drug Reactions (ADRs) with pharmaceutical products.

How to use social media as a potential data source of adverse events?
## CONTEXT - EXISTING RESEARCH

<table>
<thead>
<tr>
<th>Method</th>
<th>Data</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon-based</td>
<td>Twitter®</td>
<td>Drug representation, Time consuming</td>
</tr>
<tr>
<td>Association rule-mining</td>
<td>DailyStrength®</td>
<td>Time consuming</td>
</tr>
<tr>
<td>Machine learning</td>
<td>Twitter®, DailyStrength®</td>
<td>Normalization</td>
</tr>
</tbody>
</table>

http://creofire.com/spending-scientific-research/
GOAL

An *automated* solution for

Based on two sources:

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

*Adverse Drug Reaction

The solution should be compatible to various sources.
SOURCES

- 12th most used website in the world
- Used to send short messages (140 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

Both source have mention potential ADRs
AGENDA

Context
• Context
• Goal
• Sources

Proposed solution
• Process flow
• Extraction
• Coding
• Storage

Results & Discussion
• Results
• Discussion
PROCESS FLOW

Drugs to search

Data extraction

Database storage

Data cleaning

ADRs and drugs coding
EXTRACTION

Twitter® & Twitter API:
- **Freely** available
- **Easy** to use
- Return **structured** JSON

Reddit®:
- **No tool** for the needed data extraction
- **HTML** structure **harder** to analyze
CODING - DICTIONARIES

**MedDRA**: MEDical Dictionary for Regulatory Affairs
- Medical Terminology Dictionary
- Used to *detect* and *code* ADRs

**WHO-DDE**: WHO-Drug Dictionary Enhanced
- Classification of medicines for drug identification
- Used to *detect* and *code* drugs
STORAGE

Twitter and Reddit can change their data structure anytime which can imply:

- Change of data schema and/or format
- Data deletion
- Data addition

→ Need of a flexible structure

MongoDB: a document-oriented database
AGENDA

- Context
  - Context
  - Goal
  - Sources

- Proposed solution
  - Process flow
  - Extraction
  - Coding
  - Storage

- Results & Discussion
  - Results
  - Discussion
## RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>Reddit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period</strong></td>
<td>3 months</td>
<td>1 year</td>
</tr>
<tr>
<td><strong># records</strong></td>
<td>23,695</td>
<td>145,770</td>
</tr>
<tr>
<td><strong>records / month</strong></td>
<td>7,898.33</td>
<td>12,147.5</td>
</tr>
</tbody>
</table>

### Disclaimer:

The data contained in this paper was obtained for demonstration purposes only (from Twitter and Reddit) 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*
For Aspirin and Ibuprofen: ~116,000 records

For other drugs: ~3,500 records

Several sources needed
# RECORDS AND POTENTIAL ADR RATIO ON TWITTER®

**Ratio** (potential ADR/records) **constant**: ~40% for all drugs

- Dictionary size
- Expressions recognition
- Self learning

**Method limitation?**
IBUPROFEN (MIXED STRENGTH): ADR DETECTION VS LITERATURE

Not expected data repartition:
- Somnolence
- Constipation
- Wheezing
→ Social media embarrassment

Detection of all ADRs kind thus social media can be sources of new ADRs
Aspirin / Ibuprofen ratio:
- Occidental countries: ~50 / 50
- Africa, Middle East and Asia: ~75 / 25
- Oceania: ~34 / 66
- South America: ~55 / 45
→ Aspirin cheaper than Ibuprofen

Clonazepam overrepresentation on South America:
→ Recreational use
CONCLUSION

Objectives
- ADRs Detection
- Sources number
- Normalization

Results
- ~40% potential ADR ≈ literature Codification and storage
- Limitations

Next steps
- Improve detection
- Health specialized websites
QUESTIONS