Towards automating location-specific opioid toxicosurveillance from Twitter via data science methods

ABEED SARKER, GRACIELA GONZALEZ-HERNANDEZ, JEANMARIE PERRONE

PRESENTED BY

Abeed Sarker, Ph.D.
Department of Biomedical Informatics
Emory University School of Medicine
Email: abeed.sarker@emory.edu | abeed@dbmi.emory.edu
Twitter: @sarkerabeed
Outline

• Project: social media mining for toxicovigilance

• Study objectives
  • Effective data collection strategy from Twitter regarding opioids
  • Analyze distribution of opioid chatter
  • Categorize and annotate data
  • Supervised classification

• Long-term objectives
  • Geolocation-centric monitoring of opioid misuse/abuse
  • Temporal trends
Why is it important?

• The opioid crisis is having devastating impact in the U.S.
• More than 130 Americans die everyday from opioids
• Traditional monitoring mechanisms are slow
  • Considerable lag

Source: The National Institute on Drug Abuse (NIDA) website: [https://www.drugabuse.gov/](https://www.drugabuse.gov/)
Do people really talk about opioids on Twitter?*

- @username i shouldn't have done all that heroin this morning
- its on the news.. kensington oxys on the loose
- an average of 130 people a day die from heroine overdose in USA.
- i remember when the teacher was on heroin in class... Really!!
- i'm on this codeine because the weed made me coughing...
- saw him shooting some china white over by kensington

*Tweets have been modified in an attempt to preserve anonymity
Tasks and experiments

- Automatic, data-centric misspelling generation
- Four categories
  - Misuse/abuse (A), Information (I), Unrelated (U), non-English (N)
- Iterative annotation for improving inter-annotator agreement (IAA)

Workflow

- Spelling variant generation
- Noisy keyword removal
- Data collection
- Analysis
- Supervised classification
- Annotation
- Geolocation-based filtering
### Misspellings [1] and data distributions

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Generated Misspellings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tramadol</td>
<td>trammadol tramadal tramdol tramadols tramado tramedol tramadoll tramadole tramidol tamadol tranadol tramadol tremadol</td>
</tr>
<tr>
<td>Heroin</td>
<td>herione heroine heroins heroine heroin heorin herion</td>
</tr>
<tr>
<td>Methadone</td>
<td>methadones methodose methodone mehtadone metadone methodon methdone</td>
</tr>
<tr>
<td>Oxycontin</td>
<td>oxicontin oxcotin oycotin oxycotins oxycotin oxycotins oxycotin oxycotinine ocycontin</td>
</tr>
<tr>
<td>Codeine</td>
<td>codiene coedine codine codene codein</td>
</tr>
<tr>
<td>Dilaudid</td>
<td>delaudid dialudid dilaudad diluadid diaudid dilauadin dilauaded diluadid dillauid</td>
</tr>
</tbody>
</table>

Classifiers and features

• Classifiers
  • Naïve bayes, support vector machines, random forests and deep convolutional neural networks

• Features
  • Traditional classifiers:
    • n-grams, word clusters & abuse-indicating terms
    • 10 fold CV over training set to optimize parameters
  • d-CNN:
    • Separate training and validation set for parameter optimization
    • Dense vector representations for terms
## Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>53.9</td>
<td>51.6-56.3</td>
</tr>
<tr>
<td>Random Forest</td>
<td>70.1</td>
<td>67.9-72.2</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>69.9</td>
<td>67.8-72.1</td>
</tr>
<tr>
<td>Deep Convolutional Neural Network</td>
<td>70.4</td>
<td>68.2-72.5</td>
</tr>
</tbody>
</table>

- Moderate IAA $\kappa = 0.75$ (Cohen’s kappa)
- Particularly difficult (for annotators) to distinguish between A and I tweets
  - Lack of context in tweets
  - Ambiguous expressions
What’s next?

• Improving classification performance

• Can misuse/abuse related chatter predict other opioid related metrics?
  • County-level opioid overdose death rates? (Graves et al. (2018) showed weak but significant correlations via unsupervised approaches)

• Establish social media based near real-time monitoring system

Funding

• Research reported in this publication was supported in part by the National Institute on Drug Abuse of the National Institutes of Health under Award Number R01DA046619. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

• The data collection and annotation efforts were partly funded by a grant from the Pennsylvania Department of Health.

• The Titan Xp used for this research was donated by the NVIDIA Corporation.
Questions?

CONTACT
ABEED@DBMI.EMORY.EDU
https://sarkerlab.org
Twitter: @sarkerabeed

We are hiring!
- Postdocs
- Ph.D. Students
- Research Scientists
- Developers