DATA, TOOLS AND RESOURCES FOR MINING SOCIAL MEDIA DRUG CHATTER

Abeed Sarker
Health Language Processing Lab
Institute of Biomedical Informatics
Email: abeed@upenn.edu
Overview

• Social media data and pharmacovigilance
• NLP for pharmacovigilance
  – Data collection and annotation
• Drug-related chatter from Twitter
• Data and resources for mining
• Utilities and future tasks
Social media and pharmacovigilance

- Over **770,000** people are injured or die each year in hospitals from ADRs\(^1\)
  - Improved reporting mechanisms recommended
- 30% of adults are likely to share health-related information on social media\(^2\)
  - Provides access to large volumes of drug-related chatter

1 Agency for Healthcare Research and Quality

Twitter drug-related chatter sample

- i hate how this **firbo** and **gabapentin** robs me if my life ... **i just hate feeling so useless and worthless feeling tired**
- Off to see the gi consultant this week. Hope theres something other than **humira** to try as **not working** also **hair falling out**.
- The 100mg tabs of **trazodone** my gp prescribed are too much, now that I don't take them every night. Still **zombieish** after an hour awake
- Gone from 50mg to 150mg of **Serequel** last night. **Could barely wake up** this morning and I feel like my **body is made of lead**
- **snorted** 2 15mg **oxycodone** ($24)
- **can't sleep, temazepam** myself into a coma, pass out for hours on end. finally wake up, **feel like shite** for days. Oh I love my life! :-/
- just got retested for jcv. **tecfidera did not work out well** for me, so i'm onto **tysabri**. #ms #multiplesclerosis
Social media mining pipeline

Data collection

Annotation
Resource creation (e.g., lexicons)

Classification
Information extraction
Normalization
Case studies

- Data collection from Twitter uses drug names and their common misspellings as keywords

- Approach is semi-automatic. Available at: [http://diego.asu.edu/Publications/ADRSpell/ADRSpell.html](http://diego.asu.edu/Publications/ADRSpell/ADRSpell.html)

- Annotations performed for classification, information extraction and normalization of ADRs, and other tasks

Unannotated data

- About 4 million posts mentioning medications (in-house)
- Recently released a set of unannotated data over 4 months
  - Data is accompanied by sequential and distributed language models
Annotated data

• Binary annotations indicating ADR/nonADR posts.
  – Over 20,000 annotations. 10,000+ are publicly available
• Span annotations indicating ADRs and indications/symptoms
  – ~2000 full annotations available
• Normalization annotations with ADR mentions mapped to UMLS CUIs
  – ~2000 annotations available
• Other annotations (e.g., prescription medication abuse) available
Data, resources and tools

- Annotated and unannotated data are available with various machine learning resources (e.g., lexicons and topics)
- Tools include pre-trained classifiers and source codes
- Available at:
  - [https://healthlanguageprocessing.org/software-and-downloads/] (https://healthlanguageprocessing.org/software-and-downloads/)
- Earlier tools available at:
  - [http://diego.asu.edu/] (http://diego.asu.edu/)
Utilities

- Pharmacovigilance
  - Normalization and signal generation are still important tasks
- Medication abuse
- Sentiment analysis of users of drugs
- Estimating effectiveness of drugs
- Learning entity associations from unlabeled data
Contacts and research updates

• Follow us on Twitter to obtain data releases: @UPennHLP

Abeed Sarker
Research Scientist
Institute of Biomedical Informatics
Perelman School of Medicine
abeed@upenn.edu
Twitter: @sarkerabeed
Data, tools and resources for mining social media drug chatter