Social Media Mining for Public Health

Speaker

Abeed Sarker
Research Scholar
Department of Biomedical Informatics
Arizona State University

Overview

- The DIEGO Lab and text mining in biomedical informatics
- Natural language processing, machine learning, social media
- Social media for public health
- Social media based pharmacovigilance
- Other applications of social media
- Future of social media for public health

DIEGO lab

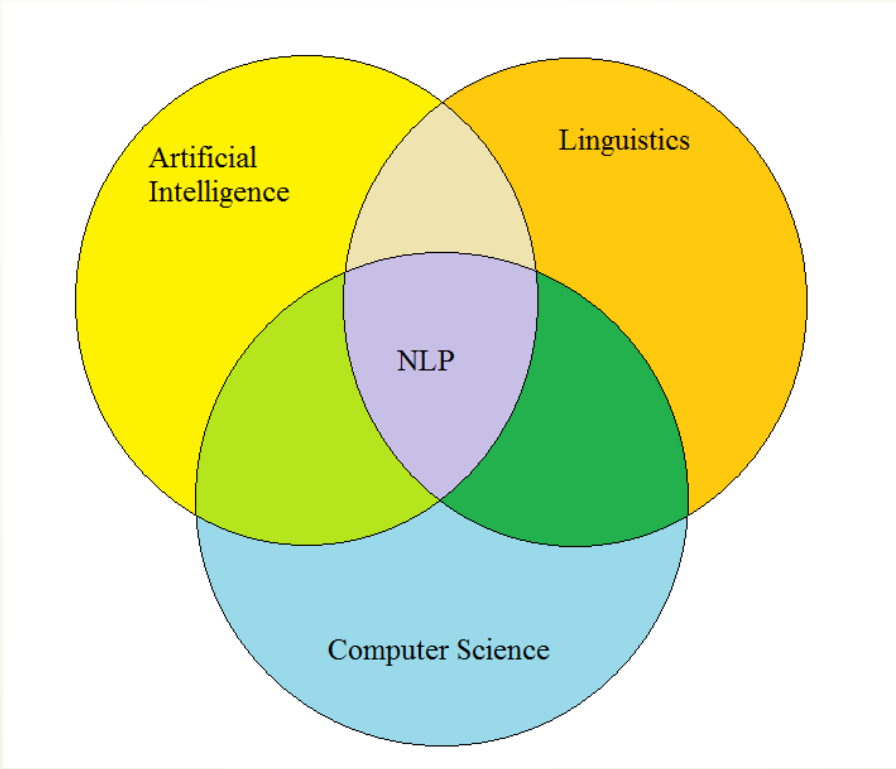
- Research areas
  - Social media mining
  - Pharmacovigilance
  - Named entity recognition of biomedical entities
  - Entity identification (normalisation)
  - Extraction of binary associations
  - Gene target prioritisation
  - Natural language processing and applied machine learning
- Website: diego.asu.edu

Members and collaborators

- Over 20 members including researchers from
  - Medicine
  - Pharmacology
  - Computer Science
  - Data Science
  - Biomedical Informatics
- Collaborators
  - Mayo Clinic, NACTEM (University of Manchester), DSV (Stockholm University), Regis University, University of Arizona
  - Pharmaceutical companies

Natural language processing (NLP)

- Use of machines to process natural (human) language

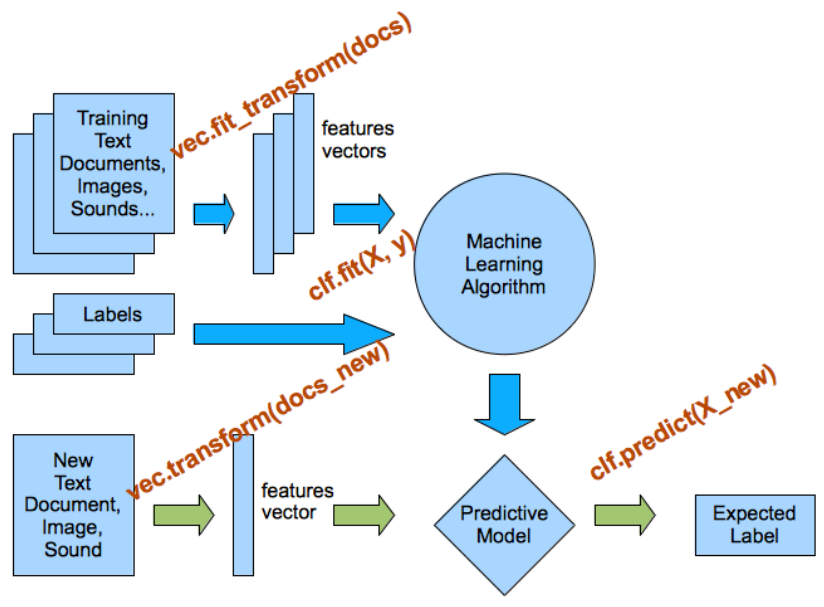


Traditional *vs.* new-age NLP

- Traditional approaches were primarily rule-based (*e.g.*, keyword based)
- New-age approaches are mostly hybrid or data-centric
- Applications:
  - Core techniques such as parsing, tokenisation, and POS tagging
  - Query formulation and searching
  - Automatic summarisation and question answering
  - Relationship extraction, sentiment analysis, predictive modeling *etc.*

Machine learning (ML)

- The construction and study of systems that learn from data
  - Rather than following explicitly programmed instructions



Source:
http://www.rosariomgomez.me/are-u-talkin-fashion-building-a-fashion-classifier-for-twitter-data/

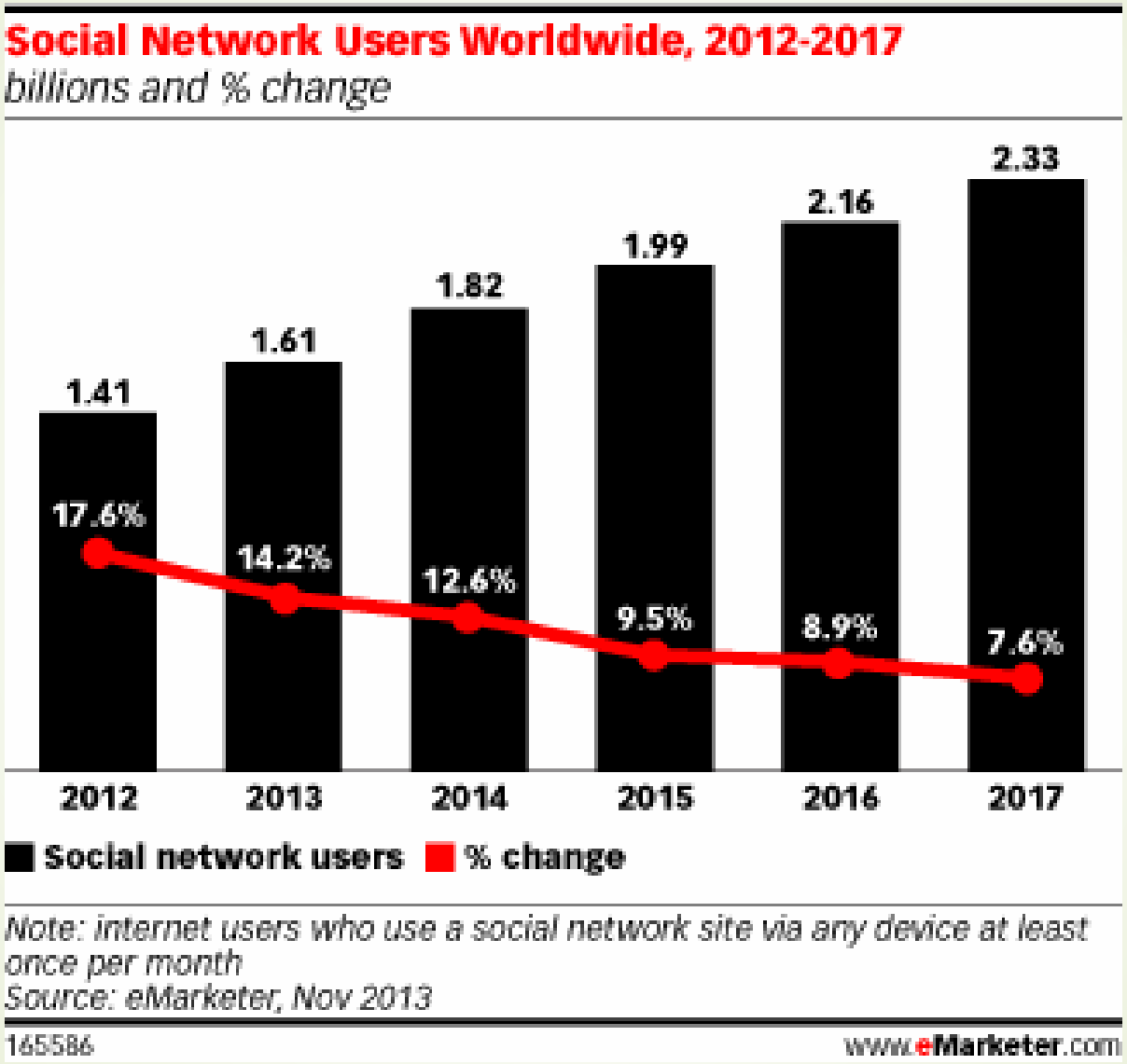
ML flavours and applications

- Supervised, semi-supervised, and unsupervised
- Neural Networks, Support Vector Machines, Bayesian Approaches (*e.g., Naïve Bayes)* and many more

- Applications (many!):
  - NLP
  - Image Processing
  - Information Retrieval
  - Bioinformatics
  - Expert Decision Making
  - Financial Market Analysis
  - Advertising

Social media

- Enormous number of users
  - Facebook: 1.28 billion monthly active users
  - Twitter: 271 million monthly active users
- Sources from which people's opinions, ideas, and experiences can be directly accessed
- Specialised social networks (*e.g.*, DailyStrength)
- Massive amounts of data are being generated
- Real-time data
- Still growing


Social Network Users Worldwide, 2012-2017
billions and % change
1.41
1.61
1.82
1.99
2.16
2.33
17.6%
14.2%
12.6%
9.5%
8.9%
7.6%
2012
2013
2014
2015
2016
2017
Social network users
% change
Note: internet users who use a social network site via any device at least once per month
Source: eMarketer, Nov 2013
165586
www.eMarketer.com

Data and the future

- Large volumes of data being constantly added and stored
  - Large component of it is language-based data
- Effective approaches required to convert
  - Raw data -> information
  - Information -> knowledge
- Current research is very much in its infancy
  - Possibilities are endless

Other challenges of social media

- Noise
- Data imbalance
- Use of colloquial language
- Out of dictionary words and misspellings
- Lack of context
  - *e.g.,* 140 characters in Twitter

Uses of social media in health

- Disease surveillance
  - Infuenza outbreaks
  - Other infectious diseases (*e.g.*, Dengue)
  - Disease prevalence (*e.g.*, current week disease prevalence)
- 'Traditional' techniques mostly rely on keyword-based approaches + location data
  - Problems of over representation of signals
- Pharmacovigilance
- Behavioral medicine
- Current approaches promise more sophisticated solutions to the health domain

Why not keyword-based approaches?

- Over estimation
  - Google Flu Trends
- Promotional material
  - Medication related studies
  - This problem is growing
- Misspellings
  - Health-related concepts are often difficult to spell
- Noise

Project: Adverse drug reaction (ADR) monitoring from social media

- Funded by NIH, the goals are to develop:
  - data collection, and annotation techniques
  - NLP techniques to identify adverse drug reaction related information
  - techniques for information extraction, and normalisation
  - medication case studies based on social media data

Adverse drug reactions

- Major public health problem
- Impacts
  - Over 770,000 people are injured or die each year in hospitals from ADRs; many of them preventable
  - ADRs can result in a number of different consequences, ranging from allergic reactions to death
  - Associated costs sum to billions of dollars
- Early detection of drug associated ADRs is a crucial research problem

Resources for pharmacovigilance research

- Voluntary reporting systems
  - The FDA AERS
  - Under reporting
- Electronic medical records
- Social media
  - Generic and health-related
  - Identified as a source for early detection

Data sources

- Twitter
  - Only about 0.5% of chatter that includes drug names mention ADRs
  - 58,000,000 tweets per day on average
- DailyStrength
  - Health related social network
  - 300,000+ monthly visitors
  - Discussions categorised by drugs
  - About 25%-35% of the posts discuss ADRs

Examples

HA! Not if you're on #Seroquil. EXTREMELY **vivid dreams** that stay in conscious memory. Very #Freaky ! Any idea why?

but first! Try these lovely pharmies! #zoloft **feel numbb** #paxil **hate life** more and everyone else

Was horrible experience. **Spiraled her into suicidal state** , ended up in hospital

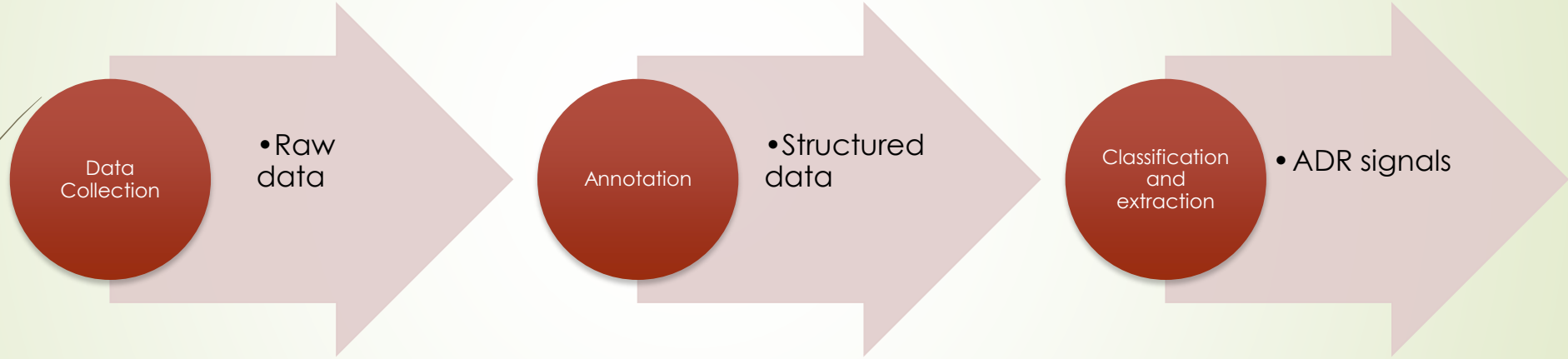
-nods- My zombie-ness when I first wake up in the morning is mostly from my nightly meds too (Seroquel).

I'd rather be on the Effexor that made me happy though I had **severe back pain** from it.. Than sad and crying for no reason on lexapro....

And then I had **horrible sleep** once I took the trazodone. I just couldn't win, haha.

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**

Pipeline


Data Collection
•Raw data
Annotation
•Structured data
Classification and extraction
• ADR signals

Data collection

- Twitter
  - Keyword-based collection from Twitter's API (150+ drugs)
  - Different surface forms of the drug name generated using phonetic variants (Pimpalkhute et al., 2014)
  - *e.g.*, Seroquel → seroquels, seroqul, seroqual
- Preprocessing
- Removed tweets with URLs (advertisements) & retweets

Annotation

- Binary annotations to indicate adverse drug reaction assertive posts
- Span annotations
  - Exact mentions
  - Type
  - UMLS concept IDs
- Multiple trained annotators + pharmacology expert to resolve annotation disagreements
- Other annotation tasks
  - Drug abuse, nutritional product safety *etc.*

Why annotations?

- Rule-based/keyword-based → data-centric/learning-based approaches
- Supervised learning
- DIEGO lab has the largest publicly available annotated data set on this topic (available at: http://diego.asu.edu/downloads/)
- Recent shared task attracted 11 teams with distinct approaches to solve the problem

Automatic classification

- Intent
  - Train algorithms to automatically identify posts containing ADR mentions
  - Filter out noisy user posts
- Approach
  - Extract features from text (using annotated data)
  - Utilise advanced NLP techniques to generate large numbers of features from the relatively short posts
- Portable systems--- may be customised to a wide range of other social media based text classification tasks

Classification results

Table 3
Paired classification performances (all instances) over the three data sets. ADR *F*-scores, non-ADR *F*-scores, Accuracies and 95% Confidence Intervals (CI) for each of the train-test set combinations are shown.

Test data	Training data	ADR *F*-score	non-ADR *F*-score	Accuracy (%)	95% CI
ADE	ADE	0.812	0.914	88.2	87.3–89.1
	ADE + DS_{ALL}	0.789	0.904	86.9	85.9–87.8
	ADE + TW_{ALL}	0.800	0.912	87.7	86.8–88.7
TW	TW	0.538	0.919	86.2	84.7–87.6
	TW + ADE_{ALL}	0.545	0.941	88.6	87.2–89.7
	TW + DS_{ALL}	0.597*	0.943	90.1	88.7–91.3
DS	DS	0.678	0.890	83.8	82.2–85.0
	DS + ADE_{ALL}	0.674	0.891	83.5	81.6–84.8
	DS + TW_{ALL}	0.704*	0.899	85.0	83.3–86.5

* Statistically significant improvement in performance over the highest score achieved in the binary classification task.

- Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J Biomed Inform 2015;53: 196–207.

Automatic extraction

- Traditional lexicon-based approaches have known limitations
  - they cannot differentiate between symptoms and adverse reactions
  - they cannot detect unknown expressions
- Supervised machine learning
  - Sequence patterns combined with other features (*e.g.*, the semantic similarities of words)
  - Contextual information

Examples

ID	Start index	End index	Type	Text
51b39ef85378b9555a2f0828	108	118	ADR	zap noises
51b7a4025378b9555a2f1934	80	92	ADR	pancreatitis
51b7a4025378b9555a2f1934	97	115	ADR	digestive problems
517ef17eac6af778c203db0a	62	78	ADR	sleep no problem
517ef17eac6af778c203db0a	88	95	Indication	anxiety
517ef17eac6af778c203db0a	107	115	ADR	hangover
51d13d2253785f584a9aec37	107	123	ADR	hair falling out
51c75fcf53785f584a9aabc9	89	94	ADR	sleep
51d0e2aa53785f584a9ae8d9	86	93	ADR	smelled
5176e99bac6af778c2036160	47	69	ADR	want to eat my own arm

ADR extraction results

Table 3: Comparison of ADR classification precision (P), recall (R), and *F*-measure (*F*) of ADRMine with embedding cluster features ($ADRMine_{WITH\_CLUSTER}$) and the baselines systems on two different corpora: DS and Twitter

Method	DS			Twitter		
	P	R	F	P	R	F
$MetaMap_{ADR\_LEXICON}$	0.470	0.392	0.428	0.394	0.309	0.347
$MetaMap_{SEMANTIC\_TYPE}$	0.289	0.484	0.362	0.230	0.403	0.293
Lexicon-based	0.577	0.724	0.642	0.561	0.610	0.585
SVM	**0.869**	0.671	0.760	0.778	0.495	0.605
$ADRMine_{WITHOUT\_CLUSTER}$	0.874	0.723	0.791	**0.788**	0.549	0.647
$ADRMine_{WITH\_CLUSTER}$	0.860	**0.784**	**0.821**	0.765	**0.682**	**0.721**

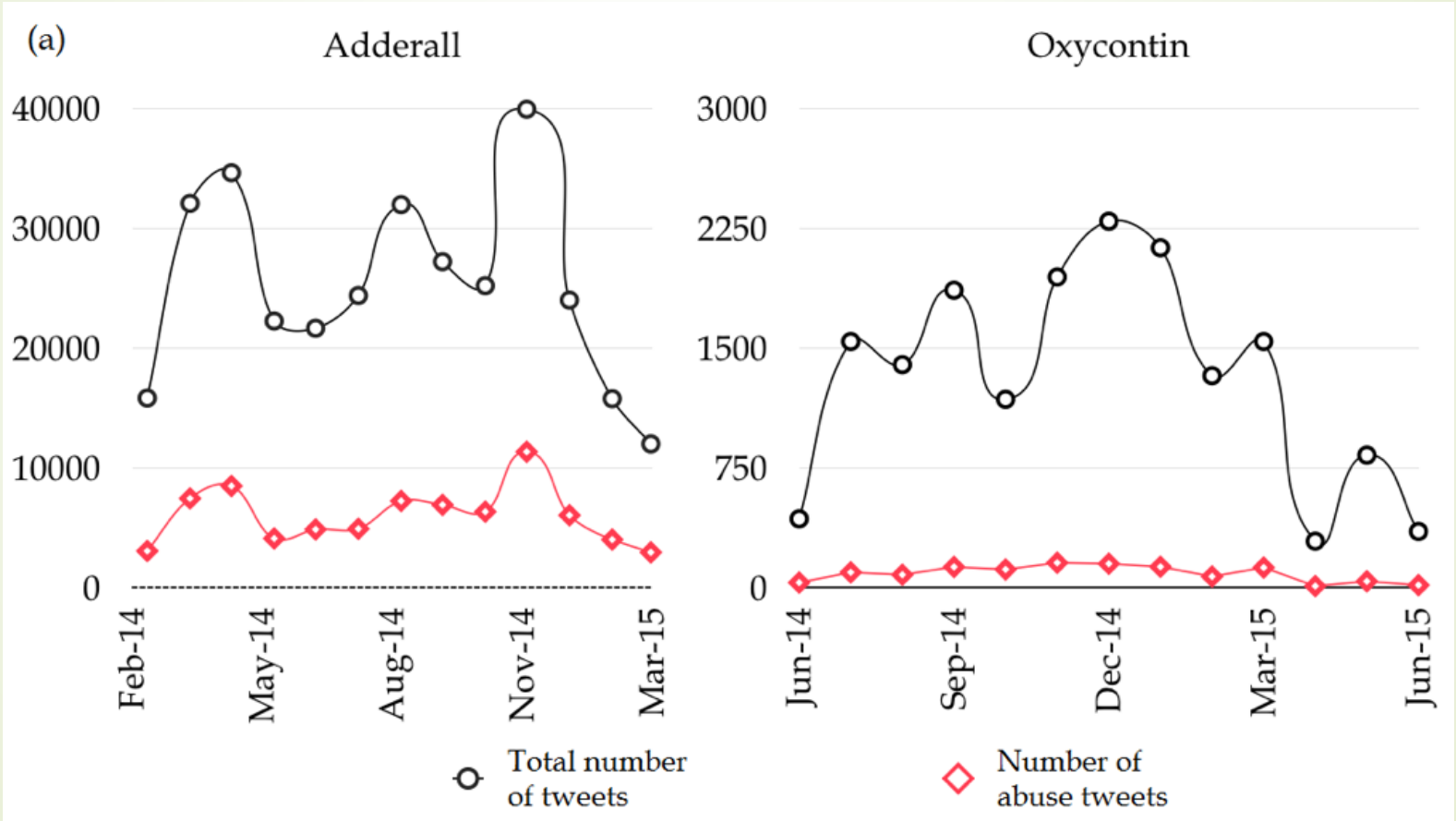
The highest values in each column are highlighted in bold.

- Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015;22(2).

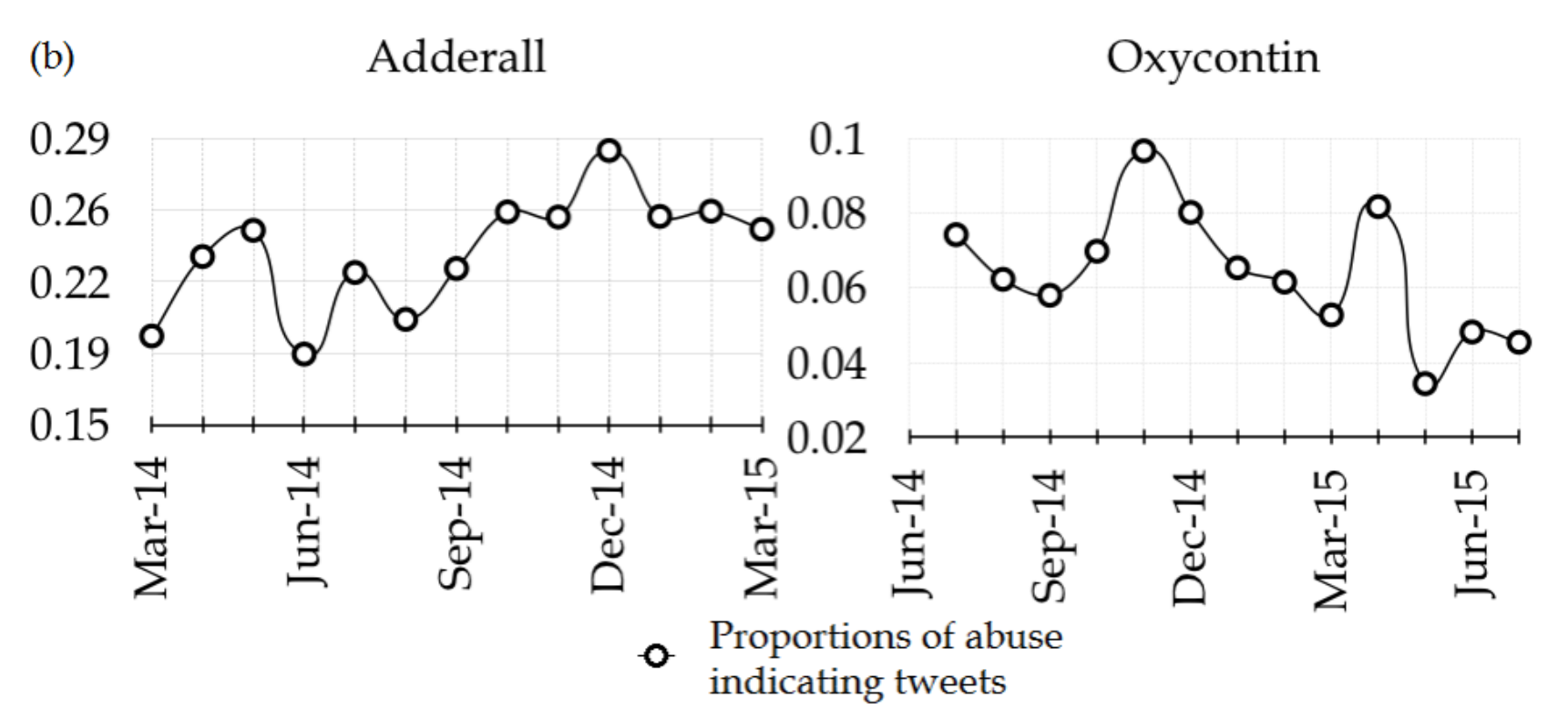
Current ADR studies

- Normalisation
  - Converting user mention of ADRs to standard IDs
  - Unsupervised, similarity based approaches
  - Spelling correction/lexical normalisation
- Comparing social media signals for ADRs with signals from other sources
- Medication case studies
- Deep learning techniques that utilise unstructured ‘big data’

Prescription drug abuse monitoring



Prescription drug abuse monitoring



Assessing the safety of dietary supplements

Product: **batch5 Extreme Thermogenic Fat Burner**	Product: **NOW Foods Bromelain**
This pills dont work at all. Its just another pill with to much caffeine and makes you cranky, edgy and nervous.	This is just fine.....not sure what it was for. I do believe it is helping with my sinus problems, at least I haven't had any lately.
I take this product before i work out and i feel more energetic and i get a feeling of well being and it last long after im done working out. I definitely recommend B4.	the product caused adverse reactions for me and could not tolerate, had back pain and right right kidney pain and decreased urine output was not good for me.
I felt awful after I took it got a terrible niacin rush would never take it again side effects are scary	This product has helped me with the pain I have in my joints due to arthritis. My knees and hands were so bad before, but after just a couple of weeks I have gotten amazing relief.

Assessing the safety of dietary supplements

Product	Human Annotator 1	Human Annotator 2	ADR Score	ADR Score Category
batch5 Extreme Thermogenic Fat Burner	High Potential	Average Potential	0.336	Average Potential
BPI Sports B4 Fat Burner	High Potential	High Potential	1.0	High Potential
Buy Garcinia Cambogia Extract With Confidence	Low Potential	Low Potential	0.129	Low Potential
Cellucor D4 Thermal Shock Thermogenic Fat Burner	High Potential	High Potential	0.614	High Potential
Garcinia Cambogia Drops	Low Potential	Low Potential	0.120	Low Potential
Liporidex MAX w Green Coffee Ultra	Average Potential	Average Potential	0.371	Average Potential
Raspberry Ketones The ONLY 250 mg PURE Raspberry Ketone Liquid	Low Potential	Low Potential	0.186	Low Potential
SafSlim Tangerine Cream Fusion	Low Potential	Average Potential	0.341	Average Potential
VPX Meltdown	Average Potential	High Potential	0.685	High Potential

Other continuing studies

- Case studies of classes of medications
  - e.g., biologics
- Mine safety profiles for drugs used during pregnancy
- Normalisation (hard problem)
- Comparison of signals derived from social media *vs.* systematic reviews
- Automatic detection of promotional/fake social media posts

Conclusions

- Primary challenges have been and continue to be NLP-oriented
- Recent advances in NLP and data science have seen increasing applications in public health
- Still very much in its infancy, but will invariably see more applications
- Specialised online health communities are opening up new channels

References

- A Sarker, R Ginn, A Nikfarjam, K O'Connor, K Smith, S Jayaraman, T Upadhaya, G Gonzalez. Utilizing social media data for pharmacovigilance: A review. **Journal of Biomedical Informatics (JBI)**. 2015.
- A Nikfarjam, A Sarker, K O'Connor, R Ginn, G Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. **Journal of the American Medical Informatics Association (JAMIA).** 2015.
- A Sarker, A Nikfarjam, D Weissenbacher, G Gonzalez. DIEGOLab: An Approach for Message-level Sentiment Classification in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SEMEVAL)*. 2015.
- A Sarker, G Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. **Journal of Biomedical Informatics (JBI)**. 2015.
- K O'Connor, A Nikfarjam, R Ginn, P Pimpalkhute, A Sarker, K Smith, G Gonzalez. Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions. ***AMIA.*** 2014.
- R Ginn, P Pimpalkhute, A Nikfarjam, A Patki, K O'Connor, A Sarker, G Gonzlez. Mining Twitter for adverse drug reaction mentions: a corpus classification benchmark. ***BIOTXTM***. 2014.
- A Patki, A Sarker, P Pimpalkhute, A Nikfarjam, R Ginn, K O'Connor, K Smith, G Gonzalez. Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction. ***BiolinkSIG.*** *2014.*

To-be published studies

- R Sullivan, A Sarker *et al.* Monitoring nutritional supplements: Challenges and promises of mining user comments for adverse events. **PSB.** 2016.
- E Emadzadeh, A Sarker, A Nikfarjam, G Gonzalez. Normalizing adverse drug reaction mentions in tweets. **AMIA.** 2016.
- A Sarker *et. al.* Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from user generated content. **Drug Safety.** 2015.
- A Sarker *et* al., Domain adaptation techniques for social media based text normalization. **TBD.**

Other interesting things…

- How's social media being used for the next U.S. election

Contacts

- Abeed Sarker (abeed.sarker@asu.edu)
- Graciela Gonzalez (graciela.gonzalez@asu.edu)