PHARMACOVIGILANCE FROM SOCIAL MEDIA: MINING ADVERSE DRUG REACTION MENTIONS USING SEQUENCE LABELING WITH WORD EMBEDDING CLUSTER FEATURES

Authors:
Azadeh Nikfarjam , Abeed Sarker , Karen O'Connor , Rachel Ginn , Graciela Gonzalez

Presented by: Azadeh Nikfarjam

Department of Biomedical Informatics

College of Health Solutions
ARIZONA STATE UNIVERSITY

Outline

- Introduction
  - Related Work
  - Objective
- Methods
- Results
- Discussion
- Conclusion

Introduction

Adverse Drug Reaction (Lee, 2006)

"Unintended, harmful response suspected to be caused by the drug taken under normal circumstances"

Impacts

- Over 2 million serious ADRs yearly
- 100,000 deaths yearly
- ADRs are the 5th leading cause of death ahead of pulmonary disease, diabetes, AIDS, pneumonia, accidents and automobile deaths
- Cost between $30 billion and $130 billion annually.

http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/DrugInteractionsLabeling/ucm114848.htm
Institute of Medicine, National Academy Press, 2000
Lazarou J et al. JAMA 1998;279(15):1200–1205
Gurwitz JH et al. Am J Med 2000;109(2):87–94
http://www.amfs.com/resources/medical-legal-articles-by-our-experts/350/adverse-drug-reactions-and-drug-drug-interactions-consequences-and-costs

Pre-marketing clinical trials

- Clinical drug trials have limited ability to detect all ADRs due to various reasons:
  - Small sample sizes
  - Relatively short durations
  - Lack of diversity among participants
    - usually excludes specific conditions: kids, elderly, pregnant women, patients with co-morbidities

Post-marketing Drug Safety Surveillance

- Post-market drug safety surveillance is required to identify potential adverse reactions in the larger population
- Spontaneous reporting systems (SRS)
  - Submitted to national agencies
    - E.g. US FDA's MedWatch program
    - UK MHRA's Yellow Card Scheme
    - Reflects less than 10% of the adverse effect occurrences (Inman & Pearce, 1993; Yang *et al.*, 2012)

Social Media for Drug Safety Surveillance

- A relatively new resource that can augment the current surveillance systems is the user posts in:
  - social health networks
  - microblogs (e.g. Twitter)
  - disease specific communities, and etc.
- Millions of health-related messages can reveal important public health issues

Example user posts in Social Media

a) #**Schizophrenia**$_{indication}$ #Seroquel did not suit me at all. Had severe **tremors**$_{ADR}$ and **weight gain**$_{ADR}$.

b) I felt awful, it made my **stomach hurt**$_{ADR}$ with bad **heartburn**$_{ADR}$ too, **horrid taste in my mouth**$_{ADR}$ tho it does tend to clear up the **infection**$_{Indication}$.

Extraction Challenges

- Consumers do not always use terms in medical lexicons.
  - They use creative phrases, descriptive symptom explanations, and idiomatic expressions.
  - "*messed up my sleeping patterns*" was used to report "*sleep disturbance*".
- Semantic type classification
  - E.g.: ADR vs. Indications
  - *This drug prevents **anxiety** symptoms [**Indication**]*
- User postings are informal, and deviate from grammatical rules:
  - Contains misspellings, abbreviations, and phrase construction irregularities
  - Extraction is more difficult compared to other corpora

Extraction of post-marketing drug safety information (Related Work)

- Various resources:
  - electronic health records, biomedical literature, SRS
- Online user posts (initially proposed by Leaman et al. in DIEGO lab)
  - health social networking sites: DailyStrength, PatientsLikeMe, and MedHelp;
  - Twitter;
  - users' web search logs.
- Most prior studies focused on exploring existing or customized ADR lexicons to find ADR mentions in user posts.
- Limited progress on automated medical concept extraction approaches, and advanced machine learning based NLP techniques.
- Less effort in addressing the introduced challenges.

Objective

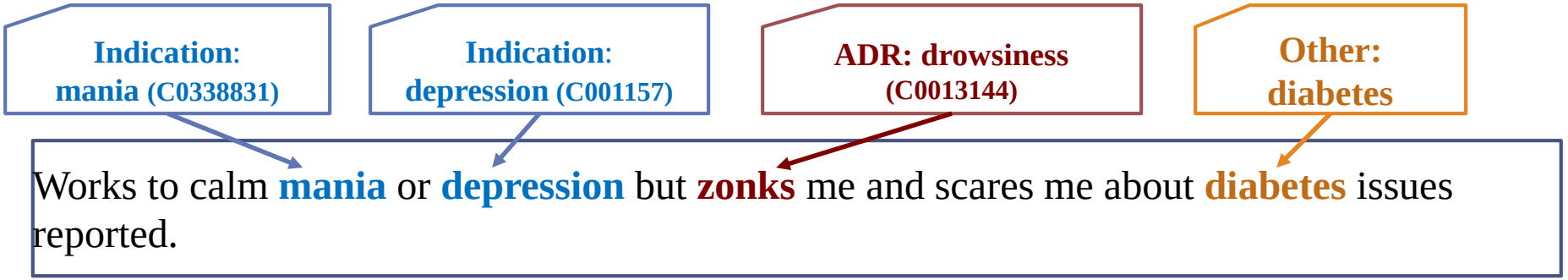
- To design a machine learning-based system (ADRMine) to extract mentions of ADRs from the highly informal text in social media
  - Hypothesis : ADRMine would address many of the abovementioned challenges, and would accurately identify most of the ADR mentions, including the consumer expressions that are not observed in the training data or in the standard ADR lexicons

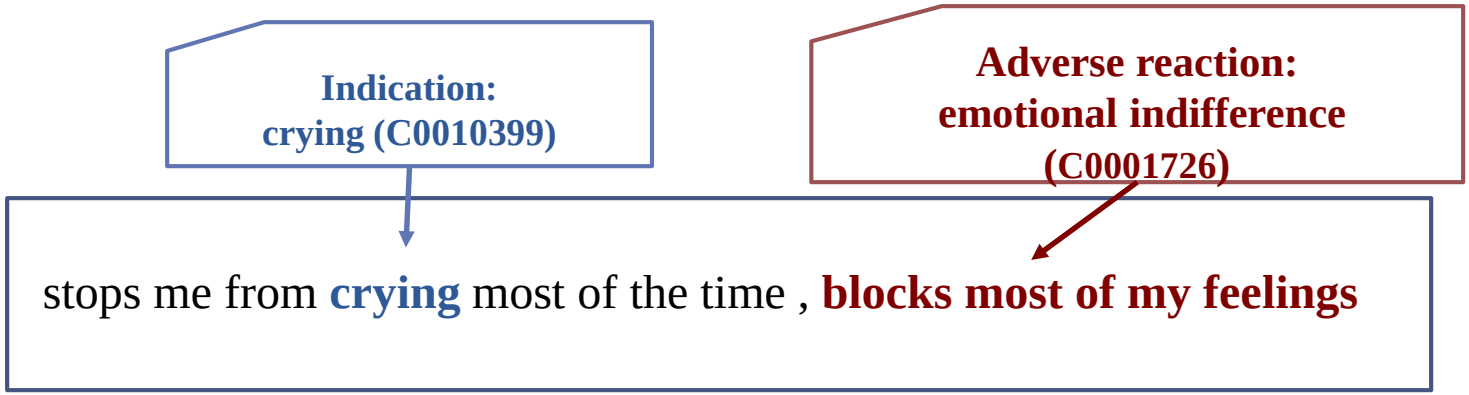
- To evaluate the effectiveness of novel semantic features (embedding cluster features) for this task
  - Hypothesis: The features would diminish the need for large amounts of labeled data.

Methods
Data Collection and Annotation

- User posts about drugs collected from two resources:
  - **DailyStrength** (http://www.dailystrength.org/)
    - The user reviews in the drug page were collected
  - **Twitter**
    - The tweets about selected drugs have been collected using Twitter API
    - the drug name (with misspelled variations) is in the search query Example: "**prozac**": "proxac", "prozacc", "prozaq", "przac", … (18 variations)
  - The list of drugs in this study is available here:
    - http://diego.asu.edu/downloads/publications/ADRMine/drug_names.txt

Corpus Annotation





Corpus Annotation (cont.)

- Every annotation includes:
  - Span, semantic type, (i.e. ADR, indication, drug interaction, Beneficial effect, other), drug name, UMLS ID — using ADR lexicon
- ADR lexicon
  - We compiled exhaustive list of ADR concepts and their corresponding UMLS IDs
  - Includes concepts from: SIDER, a subset of CHV (Consumer health vocabulary) and COSTART
  - Available for download:
    - http://diego.asu.edu/downloads/publications/ADRMine/ADR_lexicon.tsv

Corpus information

Dataset	# of user posts	# of sentences	# of tokens	# of ADR mentions	# of Indication mentions
DS train set	4,720	6,676	66,728	2,193	1,532
DS test set	1,559	2,166	22,147	750	454
Twitter train set	1,340	2,434	28,706	845	117
Twitter test set	444	813	9,526	277	41

- The annotated Twitter corpus is available for download:
  - http://diego.asu.edu/Publications/ADRMine.html

Concept extraction approach: sequence labeling with CRF

- ADRMine uses supervised sequence labeling CRF to extract mentions of ADR and indications from user sentences
- We use the IOB (Inside, Outside, Beginning) encoding
- Every token can be the beginning, inside, or outside of a semantic type. Therefore, it learns to distinguish 5 different labels: *B-ADR*, *I-ADR*, *B-Indication*, *I-Indication* and *Out*.

Gave	me	electric	shocks	and	caused	me	to	gain	Almost	40	POUNDS	in	3	WEEKS.
O	O	B-ADR	I-ADR	O	O	O	O	B-ADR	I-ADR	I-ADR	I-ADR	O	O	O

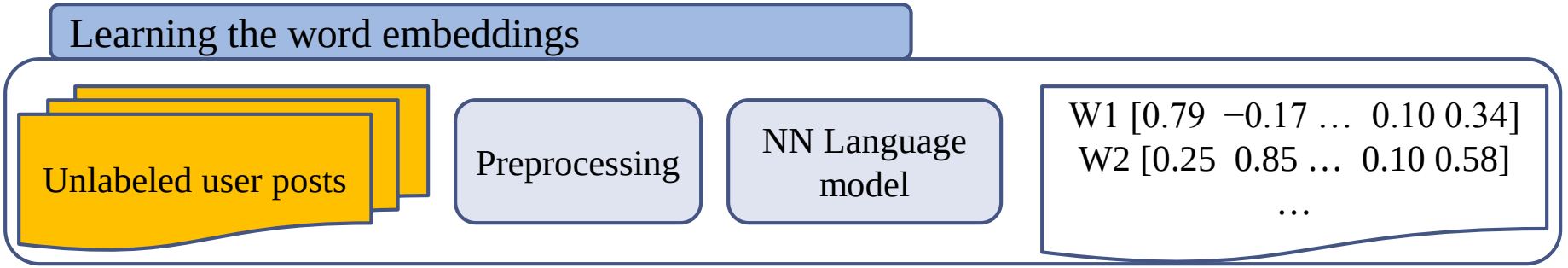
CRF Features

- Context features (t_{i-3}, t_{i-2}, t_{i-1}, t_i ,t_{i+1}, t_{i+2}, t_{i+3}).
- Lexicon feature (binary)
- POS: Part of speech of the token
- Negation: This feature indicates whether the token is negated or not.
- Embedding cluster features

Word Embedding Representations

- A word representation is a mathematical object (often a vector) associated with each word (Turian 2010).
- Conventionally NLP systems use one-hot representation which is a sparse vector.
- One-hot representations do not model the similarity between the words.
- The classifiers struggle with correctly estimating the rare or unseen words in the test sets.
- Word embedding representations are dense real-valued vectors generated by neural network-based language models. (Bengio et al., 2001;Mikolov, 2013)

Embedding cluster features



- We utilize more than one million unlabeled user sentences from both Twitter and DS.
- The word categorized into 150 distinct clusters (examples next slide)
  - Woed2vec tool (https://code.google.com/p/word2vec/) is used for generating the embeddings and the clusters using K-means algorithm.
- Seven features are defined: the cluster number for the current token and every neighbor token in a window of 7 tokens.

Examples of the unsupervised learned clusters with the subsets of the words in each cluster

Cluster#	Semantic category	Examples of clustered words
c_1	Drug	abilify, adderall, ambien, ativan, aspirin, citalopram, effexor, paxil, …
c_2	Signs/Symptoms	hangover, headache, rash, hive, …
c_3	Signs/Symptoms	anxiety, depression, disorder, ocd, mania, stabilizer, …
c_4	Drug dosage	1000mg, 100mg, .10, 10mg, 600mg, 0.25, .05, ...
c_5	Treatment	anti-depressant, antidepressant, drug, med, medication, medicine, treat, …
c_6	Family member	brother, dad, daughter, father, husband, mom, mother, son, wife, …
c_7	Date	1992, 2011, 2012, 23rd, 8th, april, aug, august, december, …

Example for CRF Features

Sentence: I had the side effect of a bloody nose$_{ADR}$ and hated it.

Token	CRF Features	Class
bloody	t_{i-3} = effect; t_{i-2} = of; t_{i-1} = a; t_i = bloody; t_{i+1} = nose; t_{i+2}= and; t_{i+3}= hate; $cluster_{i-3}$ = 77; $cluster_{i-2}$ = 49; $cluster_{i-1}$ = 49; $cluster_i$ = 147; $cluster_{i+1}$ = 116; $cluster_{i+2}$ = 43; $cluster_{i+3}$ = 51; is_negated = 0; is_in_lexicon = 1; POS = JJ (adjective)	B-ADR
nose	t_{i-3} = of; t_{i-2} = a; t_{i-1} = bloody; t_i = nose; t_{i+1} = and; t_{i+2} = hate; t_{i+3} = it; $cluster_{i-3}$ = 49; $cluster_{i-2}$ = 49; $cluster_{i-1}$ = 147; $cluster_i$ = 116; $cluster_{i+1}$ = 43; $cluster_{i+2}$ = 51; $cluster_{i+3}$ = 85; is_negated = 0; is_in_lexicon = 1; POS = NN (noun)	I-ADR
and	t_{i-3} = a; t_{i-2} = bloody; t_{i-1} = nose; t_i = and; t_{i+1} = hate; t_{i+2} = it; t_{i+3} = .; $cluster_{i-3}$ = 49; $cluster_{i-2}$ = 147; $cluster_{i-1}$ = 116; $cluster_i$ = 43; $cluster_{i+1}$ = 51; $cluster_{i+2}$ = 85; $cluster_{i+3}$ = 101; is_negated = 0; is_in_lexicon = 0; POS = CC (coordinating conjunction)	Out

Baseline ADR Extraction Techniques

- We aimed to analyze the performance of ADRMine relative to the following baseline techniques:
  - Lexicon-based technique for candidate ADR phrase extraction
  - An SVM (support vector machine) classifier for candidate phrase classification
  - Two MetaMap baselines

Lexicon-based Candidate phrase extraction

- Apache Lucene index used for indexing and retrieval of ADR lexicon entries.
- Every lexicon entry is lemmatized and the stop words are removed before indexing.
- To identify the ADR concepts in a post, a Lucene search was generated from preprocessed tokens in the tweet
  - String comparisons using regular expressions for concept identification
- Example: "… *I **gained** an excessive amount of **weight** during six months*." extracted: Weight gain

SVM Semantic Type Classifier

- A multiclass SVM classifier is trained
- Classification candidates: phrases that are matched with ADR lexicon (e.g. *gained an excessive amount of weight)*
- Semantic types: ADR, Indication, other
- SVM features: the phrase tokens, three preceding and three following tokens around the phrase, the negation feature, and the embedding cluster number for the phrase tokens and the neighbor tokens.

MetaMap baselines

- We use MetaMap to identify the UMLS concept IDs and semantic types in the user posts
  1. $\text{MetaMap}_{\text{ADR\_LEXICON}}$
     - ADRs are all extracted concepts by MetaMap existing in ADR lexicon
  2. $\text{MetaMap}_{\text{SEMANTIC\_TYPE}}$
     - Accepted ADRs are concepts with the following semantic types:
       - injury or poisoning, pathologic function, cell or molecular dysfunction, disease or syndrome, experimental model of disease, finding, mental or behavioral dysfunction, neoplastic process, signs or symptoms, mental process

Evaluation and Results

- We evaluate the performance of the extraction techniques using precision (p), recall (r) and F-measure (f):

$$p = \frac{tp}{tp+fp} \quad r = \frac{tp}{tp+fn} \quad f = \frac{2*p*r}{p+r}$$

- The proposed methods are evaluated on two different corpora: DailyStrength(DS) and Twitter

Comparison of ADRMine 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**

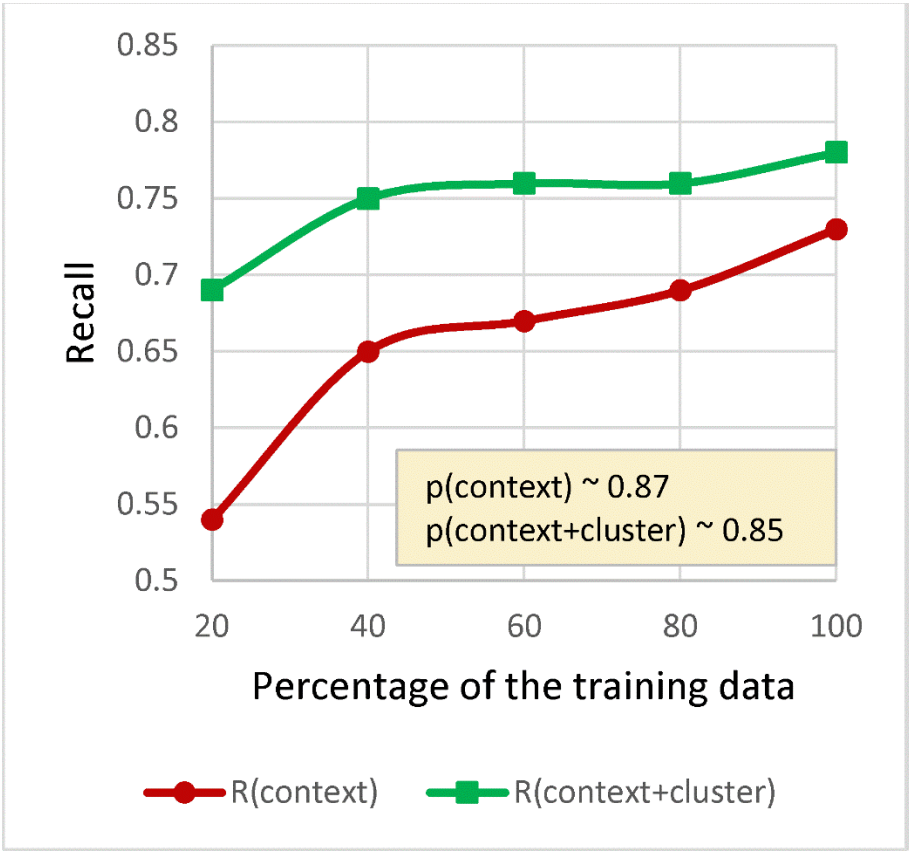
Examples of ADRMine Extracted Mentions

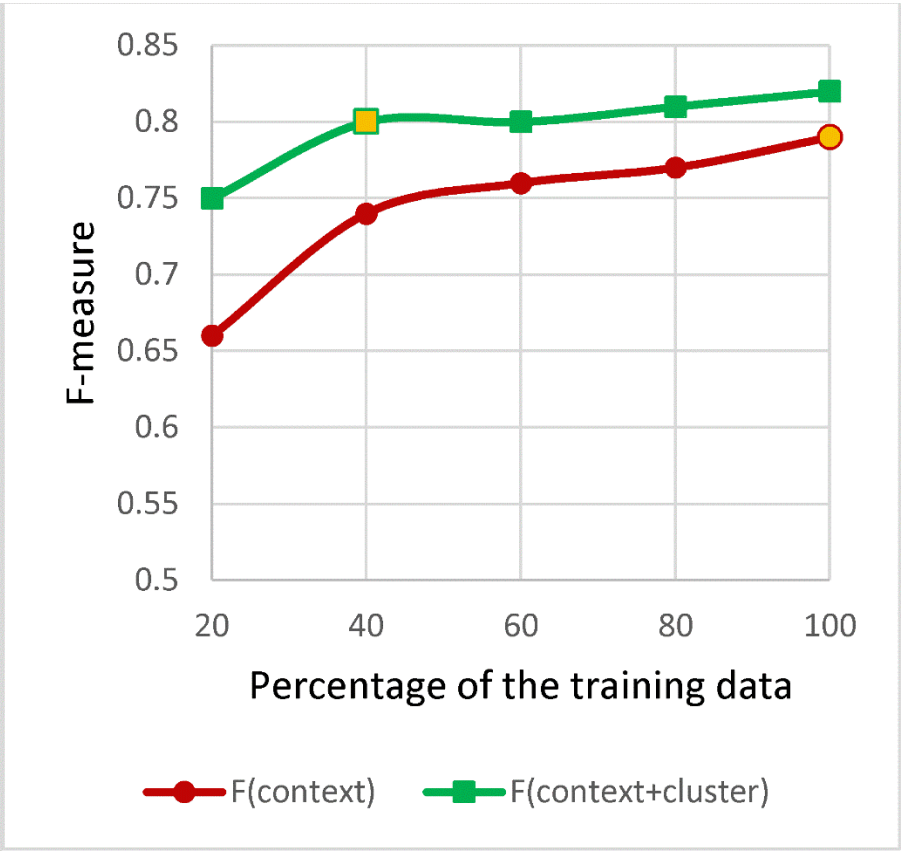
Didn't work, a ***nightmare to come off*** $_{\text{ADR}}$.

Too many negative side effects ***hard time staying awake*** $_{ADR}$ *even at very low doses.*

I didn't feel ***depressed*** $_{Indication}$ *at all anymore, but the problem was that I* ***didn't feel ANYTHING*** $_{ADR}$*!*

The effectiveness of Embedding cluster Features





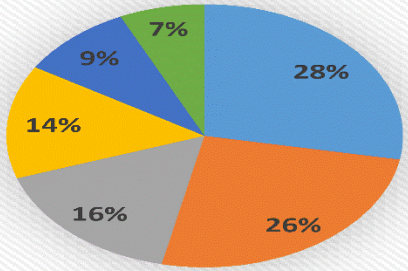
Discussion

- The extraction performance for DS is much higher than Twitter
  - Analysis of the tweets is more challenging
    - DailyStrength is a health-focused site but Twitter is a general networking site
    - Shorter text, more ambiguous
      - *E.g. Hey not sleeping. #hotflashes #menopause #effexor*”
  - More DS annotated data

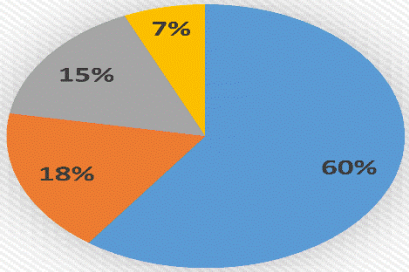
Discussion
The Effectiveness of Classification Features

- Context features are the most contributing features in performance improvement.
- lexicon, POS and negation features added no significant contribution to the results when larger number of training instance were used.
- Embedding cluster features significantly improve the recall.

Error Analysis



- Too descriptive/vague-- explained with general words: loved it , except for [not being able to be woken up at night].
- Lack of context (too short phrases or irrelevant context): [Ecstasy] side effects
- Misclassified to indications: I have terrible [pain in joints]
- Annotation guideline: Used to work , [does not anymore]
- Spelling error: ... Started [hillucinating] ... NOT cool !!!
- Idiomatic expressions: didn't work I [pack the fat on] too; My [hair seems to be shedding], ...



- Indications/beneficial effects: He is no longer in pain and [vomiting] all the time.
- Negative modifiers: Finding the right dose is a [nightmare]; Its [annoying] but the benefits are worth it.
- Non-ADR general clinical terms: [Tired] of the side effects; I am very [chemical sensitive].
- Non-ADR symptom descriptions: I have really bad spasms that [keep me up all times] of day and night.

Conclusions

- We proposed ADRMine for automatic extraction of ADR mentions from user posts in social media.
- Outperformed all baseline techniques (F-measure of 0.82 for DS, and 0.72 for Twitter)
- The embedding cluster features were highly effective in rising the recall and the overall F-measure.
- The method diminished the dependency on large numbers of annotated data.
  - Particularly effective when large volumes of unlabeled data and relatively small labeled data is available (e.g. social media posts)
- Future work
  - Further exploring the effectiveness of training deep learning techniques for automatic learning of classification features
  - Concept normalization

Acknowledgements

- This work was supported by NIH National Library of Medicine grant number NIH NLM 1R01LM011176.

- The authors would like to thank Dr. Karen L Smith, for supervising the annotation process and Pranoti Pimpalkhute, Swetha Jayaraman and Tejaswi Upadhyaya for their technical support.

Questions?

- Azadeh Nikfarjam
  - anikfarj@asu.edu