Topic Modeling for Discovering Drug-related Adverse Events from Social Media

Mengjun Xie (Ph.D.)[1], Jiang Bian (Ph.D.)[2], and Umit Topaloglu (Ph.D.)[2]

[1]University of Arkansas at Little Rock, [2]University of Arkansas for Medical Sciences

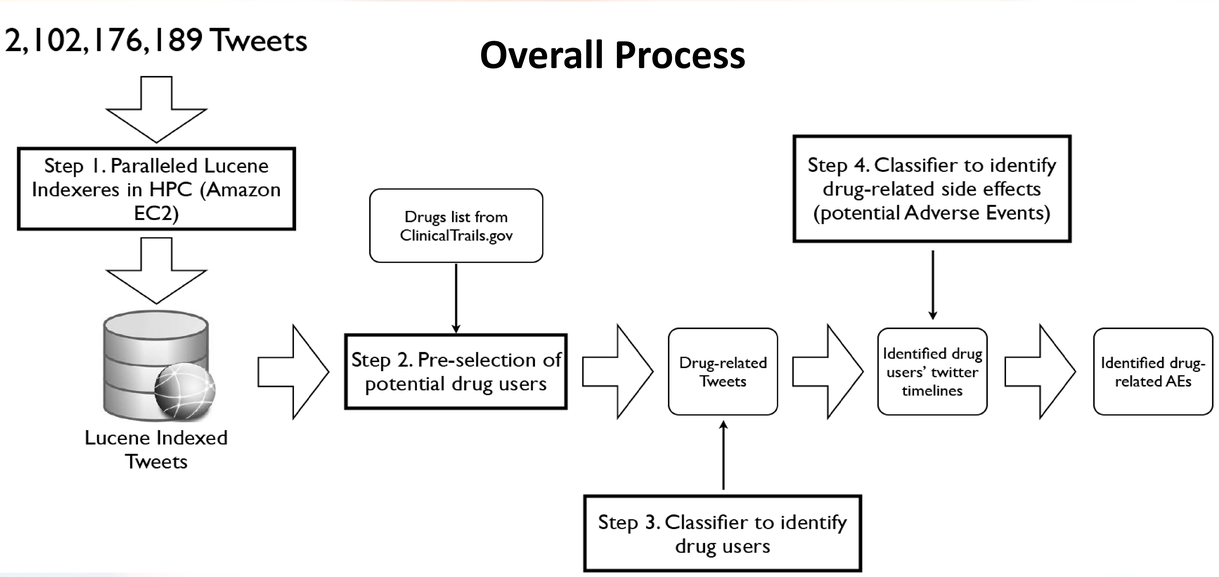
Introduction

- Goal: discover adverse events (AEs) of post-market or investigational drugs.
- Why Matters: Drug-related adverse events pose substantial risks to patients and Current self-reports mechanism is slow and AEs it detects may be incomplete.
- Approach: A data-driven approach to early detection of AEs through mining Tweeter messages.
- Intuition: From such a Tweeter message "this warm weather + tamoxifen hot flushes is a nightmare!", we can infer a possible drug use (tamoxifen) and side effect (hot flushes).

Method

- The process of discovering AEs from tweets has two subprocesses: 1) identifying the users of the drug of interest, and 2) finding possible side effects attributed to the use of the drug.
- Both subprocesses involve building and training classification models based on features extracted from the users' Twitter messages (tweets).
- In this work, we use a topic model based method to extract features while in the previous work the features are predefined.
- The full process consists of four steps, illustrated in the figure blow.

Overall Process



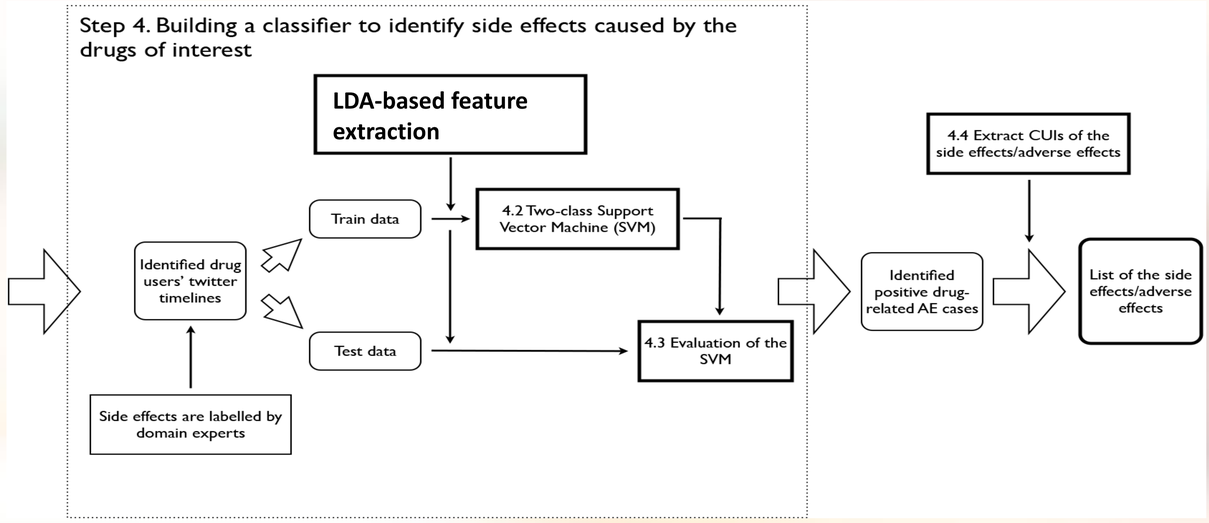
Step 1:

- Raw data: over two billion tweets (from 5/2009 to 10/2010).
- We used 15 Amazon EC2 high-memory double extra large instances (13 EC2 compute units, 34.2 GB memory) to parallelize Lucene indexing of tweets, which took 2 days to finish.
- The size of the Lucene indexes is 896 GB.

Step 2:

- five cancer drugs were selected.

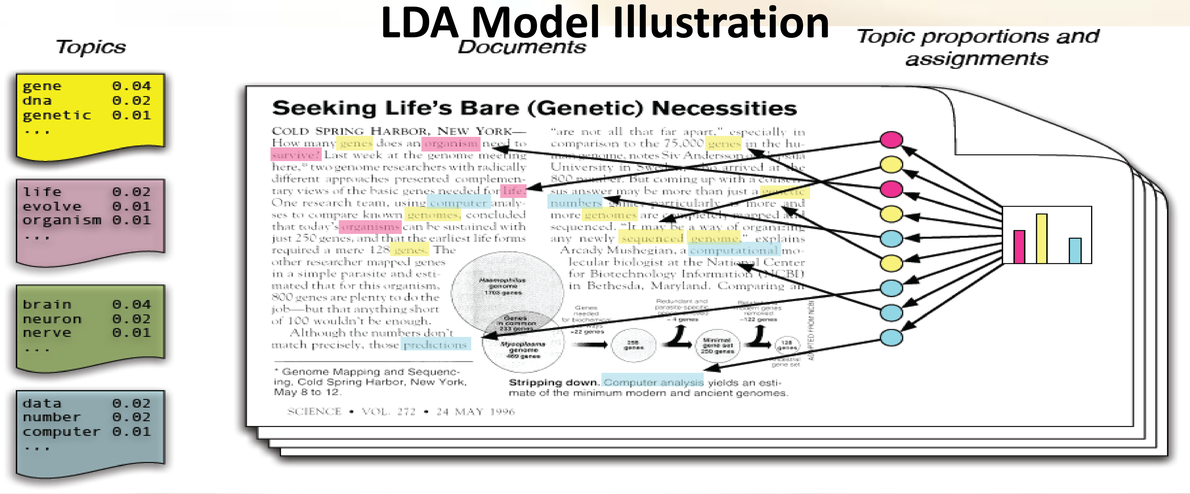
Drug Name	Synonym (s)	# of tweets	# of users
Avastin	Bevacizumab	264	236
Melphalan	ALKERAN	23	15
Rupatadin	Rupafin, Urtimed	10	10
Tamoxifen	Nolvadex	147	124
Taxotere	Docetaxel	45	39



Steps 3 and 4:

- Share a similar process in which an SVM model is built based on features extracted from tweets.
- We apply the Latent Dirichlet allocation (LDA) to categorize the collection of tweets into latent topics.
- We then use the probability distribution of topics as features to train the SVM prediction models.

LDA Model Illustration



In the LDA model, each document (tweet(s) in our study)—treated as a vector of word counts using the bag-of-words approach—is viewed as a mixture of probabilities over the topics, where each topic is represented as a probability distribution over a set of words.

Result

	Accuracy	ROC-AUC
Drug user identification (LDA model)	**0.79**	**0.87**
Drug user identification (old method)	0.74	0.82
AE detection (LDA model)	**0.81**	**0.86**
AE detection (old method)	0.74	0.74

Summary

- We believe that the performance improvement is mainly due to the improved features.
- As a data-driven method, LDA based feature extraction requires neither prior knowledge of the topics nor explicit "understanding" of the language. Thus, it is more suitable for our special tweet mining task.

References

[1] J. Bian, U. Topaloglu and F. Yu. Towards large-scale twitter mining for drug-related adverse events. In Proceedings of ACM SHB 2012.

[2] M. Blei D., Ng A. Y. and Jordan M. I. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003. ISSN 1532-4435.