Identifying Potential Adverse Drug Events in Tweets Using Bootstrapped Lexicons

Eric Benzschawel
Brandeis University
ericbenz@brandeis.edu

Abstract

Adverse drug events (ADEs) are medical complications co-occurring with a period of drug usage. Identification of ADEs is a primary way of evaluating available quality of care. As more social media users begin discussing their drug experiences online, public data becomes available for researchers to expand existing electronic ADE reporting systems, though non-standard language inhibits ease of analysis. In this study, portions of a new corpus of approximately 160,000 tweets were used to create a lexicon-driven ADE detection system using semi-supervised, pattern-based bootstrapping. This method was able to identify misspellings, slang terms, and other non-standard language features of social media data to drive a competitive ADE detection system.

1 Background

Pharmacovigilance is tasked with detecting, assessing, understanding, and preventing adverse effects or other drug-related medical problems (Organization, 2002). Adverse effects in post-approval drugs constitute a major public health issue, representing the fourth leading cause of death in the United States and an overall treatment cost higher than those of cardiovascular and diabetes care combined (Chee et al., 2011). In the United States alone, over 700,000 yearly hospital admissions—between 2 and 5% of total admissions—result from moderate to severe adverse effects (Honigman et al., 2001), underscoring the need to identify and prevent these serious medical complications.

Adverse effects from drug usage are broken down further into two categories—*adverse drug events* (ADEs) and *adverse drug reactions* (ADRs). ADRs are a subset of ADEs, where causality between a drug treatment program and negative medical reaction has been established such that the negative reactions occur in within standard dosages (Organization, 1972). ADEs more loosely define any overlapping period of drug treatment and adverse medical effects. Importantly, ADEs do not imply causation between the drug use and co-occurring negative event (Eriksson et al., 2013). Timely, accurate identification of these medical complications therefore facilitates improvements in patient health and helps decrease both manpower and monetary costs to a healthcare system and is considered a key quality of medical care (Honigman et al., 2001).

Existing systems of ADE documentation typically rely on automatic reporting systems hosted by national or international public health organizations, electronic health records, or data from other high-quality resources. Social media was an untapped resource until recently, despite evidence that suggests nearly 31% of patients suffering from chronic illness and 38% of medical caregivers consult drug reviews posted online to various social media sites (Harpaz et al., 2014).

Twitter has recently been used as an ADR detection resource in numerous research studies within the last five years, with methodologies ranging from lexicon matching to supervised machine learning (Sarker et al., 2015). Tweets can be used to supplement existing electronic ADE/ADR monitoring systems by providing real-world, real-time clinical narratives from users posted in the public domain. Because many electronic monitoring systems underreport the prevalence of minor ADEs/ADRs, typically due to their absence from medical records and clinical studies (Eriksson et al., 2013), Twitter presents a valuable resource for providing data on a wide range of negative medi-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics – Student Research Workshop, pages 15–21,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

cal events.

Social media data presents unique challenges to clinical NLP studies in ways analogous to electronic medical records—non-standard syntax, jargon, and misspellings handicap many existing NLP systems (Eriksson et al., 2013). Handling these areas of non-standard language usage complicates lexicon-based attempts at retrieving Tweets containing potential ADEs/ADRs. Many recently published systems handle this by mapping annotated ADEs/ADRs to entries in medical ontologies (Sarker et al., 2015). Annotation is a time-consuming process and limits the size of training data sets. Many problems with non-standard language usage can be addressed with semi-supervised, pattern-based bootstrapping which, after sufficient analysis, yields high-quality lexicons with competitive ADE/ADR detection capabilities.

2 Data

The largest existing publicly available dataset for this domain is Arizona State University's DIEGO Lab data, containing over 7,500 tweets annotated for presence or absence of an ADR (Ginn et al., 2014; Nikfarjam et al., 2015). Roughly 2,000 of the tweets contain annotated ADR relations. This data set has been used in both machine learning and lexicon-based approaches to ADR detection in social media (O'Connor et al., 2014).

In order to take advantage of semi-supervised learning methods and real-time data, TwitterDrugs, a new corpus of 166,551 tweets, was generated from public tweets mined from mid-January to mid-February 2016 using 334 different drugs. Drugs were compiled from those used in the DIEGO data and supplemented with the New York State Department of Health's *150 Most Frequently Prescribed Drugs*[1], and those listed in Chemical and Engineering News' *Top 50 Drugs of 2014*[2].

After collecting approximately 700K query results, each tweet was heuristically screened for relevance. Tweets were considered irrelevant if they contained an external URL, any of a set of 16 salesmanship terms such as *promo* or *free shipping*, and whether the tweet text itself contained the queried drug string. Screening removed roughly 76.1% of mined tweets. The corpus is available online[3] for future use or expansion and represents the largest available data set for Twitter-based clinical NLP tasks.

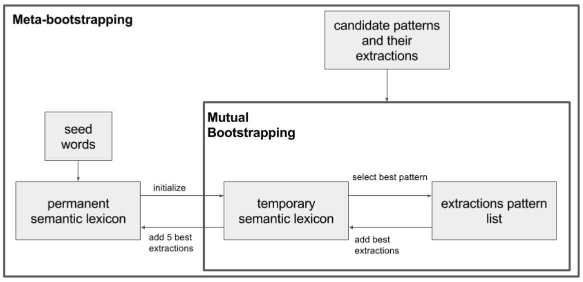

Figure 1: Riloff and Jones (1999)'s meta-bootstrapping algorithm

3 Methodology

Identifying potential ADEs required extraction of both drug mentions and negative medical events, for instance `oxycontin` and `made me dizzy`. Novel mentions were identified using a set of extraction patterns and a lexicon. Extraction patterns are flexible regular expressions capable of identifying both known and novel mentions. For instance, the pattern `took three <DRUG>` might identify `oxycontin`, `ibuprofen`, or `benzos`. `made me <REACTION>`, similarly, might identify `dizzy`, `hungry`, or `throw up`. Newly identified items are added to lexicons which are in turn used to identify new items.

Two separate lexicons for drugs and medical events were generated using the *meta-bootstrapping* algorithm detailed in Riloff and Jones (1999), which uses a pre-defined set of patterns to identify novel lexicon items occurring in similar environments as known lexicon items. To identify novel mentions, the algorithm relies on an initial set of extraction patterns and a small number of seed words to define the semantic category of interest, as seen in Figure 1. Though bootstrapped lexicons contain noise, manually screening for relevant items results in robust, automatically generated lexicons well suited to the task of identifying potential ADEs. Importantly, this method does not require expensive manual annotation and is capable of handling the colloquial terms and misspellings commonly found in social media data even though it is not specifically tailored for non-standard usage.

[1] `apps.health.ny.gov/pdpw/DrugInfo/DrugInfo.action`

[2] `CEN-supplement092014.pdf`

[3] `github.com/ericbenz/TwitterDrugs`

Meta-bootstrapping first identifies relevant extraction contexts from an input corpus and lists of known category items. This, in turn, results in a list of context patterns. Contexts were generated by taking combinations of one to three words preceding or following the known category item. Table 1 shows how known drug names and medical events present in each context pattern were anonymized with regular expressions capable of extracting one or many words.

Table 1: Extraction patterns and possible matches

Candidate Pattern	**Extracted Entities**
`took (\S+) tablet`	*ibuprofen, xanax, one, 25mg*
`made me (\S\s+)+`	*throw up, feel like dying, super happy*

Each candidate pattern was subsequently scored on the basis of how many new category items it extracts relative to the number of existing lexicon items. Scoring is initially spurred by the relatedness of extracted entities to a handful of seed words defining the semantic category of interest. Each pattern is scored with the function

$$score(pattern) = R * \log_2 F \qquad (1)$$

where F is the number of unique entities generated by the pattern which are already present in the semantic lexicon and $R = \frac{F}{N}$, where N is the total number of words the pattern extracted. R is high when patterns extract numerous items that are already contained in the semantic lexicon, as this reflects a high likelihood that all entities produced by this pattern are semantic category items.

This scoring function, however, is incapable of appropriately addressing the robustness of multi-word medical events. In some cases, an extracted entity contains multiple unique reactions, such as

`gave me vertigo and really bad nausea`

where `vertigo` and `nausea` should be considered independently. Judging the above example based on the whole string as an indivisible entity will score it too low to be considered semantically relevant. This is because the string as an indivisible whole is unlikely to ever occur again or bear strong semblance to the provided seed words or existing lexicon items. Only portions of this string are important potential category items and are likely to be included in seed words or easily identified by patterns extracting single words.

Reranking this pattern to favor extractions containing these two substrings can allow the the entire extraction to enter the medical event lexicon where each relevant bit can be manually identified in post-processing. To do this, the scoring function was modified as

$$score(pattern) = \lambda(R * \log_2 F) \qquad (2)$$

where F is re-evaluated as the number of relevant substrings, and λ is a penalty term where $\lambda = \frac{c}{\log_2(avg\_words)}$, where c is a constant and avg_words is the average number of tokens per extraction per pattern. All other terms remain the same between both scoring functions.

Because the F values grow increasingly large as the semantic lexicons grow, the λ penalty is introduced to control the balance of single and multiple word entities. Shorter strings containing more relevant entities are penalized less than longer ones potentially containing lots of noise. The c constant must grow in proportion to the number of data instances being used in the training set. Too small a c value will result in lexicons comprised mostly of single-word extractions. Too large a c value will result in lexicons comprised mostly of multi-word extractions.

Following the scoring of each pattern, each entity is evaluated with the scoring function

$$score(entity) = \sum_{k=1}^{N} 1 + (0.1 * score(pattern_k)) \qquad (3)$$

where N is the number of different patterns which found the entity being scored. This function scores each entity on the basis of how many patterns were able to identify it, as words extracted by numerous patterns are more likely to be true members of the semantic lexicon. The five highest scoring patterns are added to the semantic lexicon, serving as additional seed words for subsequent bootstrapping iterations. This process continues until end conditions are reached.

4 Results

Six different training sets were used in the bootstrapping tasks to explore the influence of unannotated data during lexicon generation. Each training set contained the full DIEGO Lab training corpus and an increasingly large amount of non-overlapping TwitterDrugs data. The bootstrapping procedure outlined above continued until lexicons contained maximally 5000+i items, where i is the number of seed words. Bootstrapping terminated early if new items were not added for five consecutive iterations.

The resulting lexicons were used to flag tweets in held-out test sets where an extracted drug co-occurred with an extracted reaction. The DIEGO test set was used to compare flagged tweets using this methodology to O'Connor et al. (2014), which utilized a different lexicon-based ADR detection algorithm on the same DIEGO data set. Tweets flagged using bootstrapped lexicons increased precision, recall, and F_1 in most cases, suggesting the viability of this method.

4.1 Generating Drug Lexicons

Drug lexicons were generated using 10-20 seed words. As the number of training instances increased, additional seed words were required to spur the bootstrapping algorithm to add lexicon items in early iterations. Seed words were taken from the most frequently occurring drugs in the DIEGO training corpus.

Using only the DIEGO training data resulted in 1907 candidate patterns, 1312 extracted entities, and 113 relevant identified drugs. The best performing training set added 5K tweets from TwitterDrugs to those in the DIEGO training set, resulting in 355 relevant extracted entities of which nearly 60% were neither in the DIEGO data nor the list of drugs used to generate the TwitterDrugs corpus. Included in these lexicons are numerous misspellings, slang terms, and hashtags.[4]

4.2 Generating Medical Event Lexicons

Due to the challenges associated with multi-word extractions, only three training sets were explored for reaction extraction. 30 seed word were used for all bootstrapping procedures, taken from the most frequent annotated ADRs in the DIEGO dataset provided they were less than five words long.

Using only the DIEGO training data resulted in 32,879 candidate patterns, producing a lexicon with 1321 items. To balance single and multi-word expressions, where $c = 0.25$ for this small dataset. Manual analysis of each lexicon item yielded 500 medical events after complex, multi-word entities were broken down. The largest lexicon contained 783 medical events extracted from 177,494 patterns generated by appending 5K tweets from TwitterDrugs to the DIEGO training set. $c = 0.75$ in this case. Over 87% of this lexicon contained novel entities.

4.3 Identifying Potential ADEs

Tweets were flagged as 'potentially containing an ADE' by identifying those in which a term from a drug lexicon co-occurred with one from a medical event lexicon. The effects of increasing the amount of training data can be seen in Table 2, which shows that an increasing proportion of tweets are flagged as the amount of training data increases. This suggests that the composition of the resulting lexicons contains drugs and reactions that more frequently co-occur.

The low proportion of flagged tweets is unsurprising, as most Twitter users rarely discuss the physical effects of their drug use. It is important to emphasize that the proportion of true ADEs is not identical to the proportion flagged. Discussion of drug indications—why a drug was taken—and beneficial effects are much more common than ADEs or ADRs. Of the proportion of flagged tweets, roughly 25.2% contained obvious ADEs. This is roughly 16% more than the 9.3% captured in the O'Connor et al. (2014) study which used only the DIEGO data.

In order to better evaluate the composition of flagged tweets using the bootstrapped lexicons, results were directly compared to the O'Connor et al. (2014) study using the 317 tweets distributed in the DIEGO Lab test set. O'Connor et al. (2014) reported precision, recall, and F_1 scores of 0.62, 0.54, and 0.58, respectively. In nearly all cases, bootstrapped lexicons have higher precision and F_1 score as Table 3 shows.

Adding small amounts of data helped increase performance mostly through increases in precision. Larger datasets hurt performance because the bootstrapped lexicons were tuned more appropriately to the composition of drugs and reactions present in the TwitterDrugs corpus which are not guaranteed to overlap exactly with the DIEGO data despite the shared source (Twitter).

Flagged tweets must be manually reviewed for potential ADE/ADR relations. Because flagged tweets were simply captured by mere co-

[4] Twitter permits the use of the # 'hashtag' to prefix strings for searching, indexing, and statistical analysis such as in `#adderall` or `#mighthaveaheartattack`

Table 2: Proportions of flagged tweets as 'potentially containing ADE relation' increases as larger amounts of TwitterDrugs data is used for bootstrapping.
(*—lexicon generated from DIEGO +5K TD dataset)

Training Corpus	Held-out Test Set	#Drugs	# ADEs	Num. Flagged	% Flagged
DIEGO	TwitterDrugs (TD)	113	500	7,993/166,551	4.80%
DIEGO +1K TD	165K TD	235	702	22,981/165,868	13.85%
DIEGO +5K TD	160K TD	355	783	25,135/161,868	15.53%
DIEGO +10K TD	155K TD	343	783*	24,661/156,868	15.72%
DIEGO +25K TD	140K TD	311	783*	22,668/141,868	15.98%
DIEGO +50K TD	115K TD	287	783*	19,091/116,868	16.34%

Table 3: Precision, recall, and F_1 score for bootstrapped lexicons using different training set combinations using larger portions of TwitterDrugs data, best results in **bold**

		Drug Train Set					
		DIEGO	+1K	+5K	+10K	+25K	+50K
Med.Event Train Set	DIEGO	$P = .7321$	.7182	.7297	.7156	.7170	.7142
		$R = .5125$	.4938	.5062	.4875	.4750	.4688
		$F_1 = .6029$	.5852	.5978	.5799	.5714	.5660
	+1K	**.7419**	.7177	.7280	.7154	.7167	.7167
		.5750	.5563	.5688	.5500	.5375	.5373
		.6479	.6267	.6386	.6219	.6143	.6142
	+5K	.7368	.7130	.7241	.7105	.7112	.7112
		.5250	.5125	.5250	.5063	.4938	.4938
		.6131	.5964	.6087	.5912	.5830	.5830

occurrence of terms, numerous captured tweets contain discussions of beneficial effects or why a drug was taken. For instance, no obvious ADE/ADR exists in the flagged tweet

That ibuprofen 800 knocked my
headache right out

Contrast this with

took this vicodin and it is
seriously hard to breathe all of
a sudden

which clearly documents a potentially dangerous co-occurrence of Vicodin® and breathing difficulties. Untangling beneficial effects and drug indications remains a problem area for automatic ADE/ADR detection especially given that similar language is used for both.

5 Discussion

Though social media represents a rich source of data, ADE detection with lexicon-based methods remains vulnerable to data sparsity—a low percentage of tweets containing drug names actually include ADEs. However, the results discussed above show that bootstrapping can increase the proportion of true ADEs in returned datasets. Meta-bootstrapped lexicons do not require extensive manual annotation unlike other recent lexicon-based systems. Because of the scoring function, bootstrapped lexicons are able to easily capture variations in spelling and slang phrases provided they occur in contexts similar to the words present in the growing semantic lexicon.

In the drug lexicon, several misspellings or slang variations of numerous drugs were identified, such as bendaryl (Benadryl®) or xannies (Xanax®), addressing a problem area for social media data. If one were to simply apply existing drug lexicons against this dataset, any slang terms or misspellings would be missed without additional processing. Meta-bootstrapping can easily retrieve this data, with the only post-processing being quick manual sifting of generated lexicons for relevant category items.

Medical event lexicons tended to robustly include slang descriptions for medical issues ranging from intoxication (tweakin, turnt up, smashed) to mental states (got me up like zombies), to descriptions of body processes and fluids (barf, urine contains blood). These cannot be identified with existing medical ontologies and several are liable to change dramatically as drug users modify the ways they describe their experiences. Importantly, manual analysis can easily capture these potential ADE indications without robust medical training.

Taken together, misspellings and common slang descriptions can be used to identify potentially severe ADEs, such as

```
The ER gave me percs and
flexeril, I'm high af lmao
```

where `percs` is a slang term for Percocet®, and `high` a common generic description for any number of abnormal sensory events. Percocet® and Flexeril® have a high potential for drug interaction causing drowsiness, lightheadedness, confusion, dizziness, and vision problems[5]—all potential adverse events contained within the generic slang term. Within slang-driven social media data, this drug interaction and its associated side effect would be difficult to capture without the flexible lexicons generated by the bootstrapping procedure.

Because the bootstrapped lexicons require manual pruning of irrelevant results, meta-bootstrapping is unlikely to save large amounts of time compared to existing research methods. However, the ease at which novel, relevant, non-standard lexicon items are identified and added to the lexicon and the competitive abilities of known-ADE identification in a small test set emphasizes the applicability of this approach for this task.

6 Future Work

The lexicons generated by meta-bootstrapping provide numerous opportunities for research extension. For instance, lexicons may be easily applied across a drug class, allowing for fast identification of ADE discussion in social media across a particular class of interest, such as the ongoing crisis surrounding the abuse of prescription-only narcotic painkillers. After flagging a tweet containing an ADE/ADR resulting from opioid use, researchers could utilize tweet metadata to help crisis managers identify demographic areas of interest for more targeted care.

Outside pharmacovigilance, the lexicons can also be used to 'bootstrap' corpus generation. Because novel extractions represented roughly 60% of the generated drug lexicon, these new entries can be used to expand the search query set, returning a more diverse set of tweets than the original 334 drug names. This, in turn, is likely to lead to identification of more novel items, allowing the process to be repeated. Doing so allows for easy identification of slang terms as they are created and enter common use.

Lastly, the TwitterDrugs corpus represents a rich resource for subsequent research. It may be easily annotated for supervised techniques, or can be explored with different semi- and unsupervised methods for lexicon generation, relation extraction, or ADE/ADR classification. The bootstrapping procedure itself can be modified to include additional standardization techniques which may diminish the number of patterns by simplifying linguistic complexities. Lemmatization would be highly effective here, allowing patterns differentiated by only inflectional morphology to be combined. However, many of these standardization techniques still perform poorly on the non-standard language found in social media data.

Acknowledgments

Special thanks to Prof. Nianwen Xue, the advisor for this work, which presents a portion of a masters thesis on the same topic titled *Identifying Adverse Drug Events in Twitter Data Using Semi-Supervised Bootstrapped Lexicons* and available electronically at the Brandeis Institutional Repository[6]. Thanks also to Prof. Sanda Harabagiu (UT-Dallas), Prof. James Pustejovsky, Dr. Marc Verhagen, Dr. Keith Plaster, and Nikhil Krishnaswamy for additional help and input.

References

Brant W Chee, Richard Berlin, and Bruce Schatz. 2011. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. 2013. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. *Journal of the American Medical Informatics Association*, 20(5):947–953.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen

[5] `umm.edu/health/medical /drug-interaction-tool`

[6] `bir.brandeis.edu/handle/10192/32253`

Smith, and Graciela Gonzalez. 2014. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Safety*, 37(10):777–790.

Benjamin Honigman, Patrice Light, Russel M Pulling, and David W Bates. 2001. A computerized method for identifying incidents associated with adverse drug events in outpatients. *International Journal of Medical Informatics*, 61(1):21–32.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association.

World Health Organization. 1972. International drug monitoring: the role of national centres. *World Health Organization Technical Report Series*, 498:1–25.

World Health Organization. 2002. The importance of pharmacovigilance.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.