Detecting Personal Medication Intake in Twitter: An Annotated Corpus and Baseline Classification System

Ari Z. Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, Graciela Gonzalez
Department of Biostatistics, Epidemiology, and Informatics
Perelman School of Medicine
University of Pennsylvania
{ariklein,abeed,mrou,karoc,gragon}@upenn.edu

Abstract

Social media sites (e.g., Twitter) have been used for surveillance of drug safety at the population level, but studies that focus on the effects of medications on specific sets of individuals have had to rely on other sources of data. Mining social media data for this information would require the ability to distinguish indications of personal medication intake in this media. Towards that end, this paper presents an annotated corpus that can be used to train machine learning systems to determine whether a tweet that mentions a medication indicates that the individual posting has taken that medication (at a specific time). To demonstrate the utility of the corpus as a training set, we present baseline results of supervised classification.

1 Introduction

Social media allows researchers and public health professionals to obtain relevant information in large amounts directly from populations and/or specific cohorts of interest, and it has evolved into a useful resource for performing public health monitoring and surveillance. According to a Pew report (Greenwood et al., 2016), nearly half of adults worldwide and two-thirds of all American adults (65%) use social media, including over 90% of 18-29 year olds. Recent studies have attempted to utilize social media data for tasks such as pharmacovigilance (Leaman et al., 2010), identifying user behavioral patterns (Struik and Baskerville, 2014), analyzing social circles with common behaviors (Hanson et al., 2013b), and tracking infectious disease spread (Broniatowski et al., 2015).

A large subset of the public health-related research using social media data, including our prior work in the domain, focuses on mining information (e.g., adverse drug reactions, medication abuse, and user sentiment) from posts mentioning medications (Korkontzelos et al., 2016; Hanson et al., 2013b; Nikfarjam et al., 2015). Typically, these and similar studies focus on information at the population level, but processing and deriving information from individual user posts poses significant challenges from the natural language processing (NLP) perspective. Researchers attempt to overcome the noise and inaccuracies in the data by relying on large amounts of data. For example, Hanson et al. (2013b; 2013a) attempted to estimate the abuse of Adderall® using Twitter by detecting the total number of mentions of the medication. The authors did not attempt to assess if a mention represented personal intake or not.

While such a strategy may suffice for deriving estimates "by proxy" at the population level (e.g., higher volume of chatter means higher rates of use), it has at least two limitations: (i) the actual number of tweets representing personal intake within a given sample of tweets is unknown, and (ii) it is not possible to assess the effects of medication intake on subsets of users of interest who take the medication. Studies focusing on specific subsets of individuals rely on other sources of data, such as electronic health records and published literature from clinical trials, where information about the individuals' medication intake is explicit (e.g., Akbarov et al., 2015; Zhou et al., 2016; Romagnoli et al., 2017). Harnessing social media for studying the effects of medications on specific cohorts would require developing systems that can automatically distinguish posts that express personal intake from those that do not.

Due to the very recent incorporation of social media data in healthcare systems, published research on our target task of creating a corpus for automatic detection of personal medication intake

Proceedings of the BioNLP 2017 workshop, pages 136–142,
Vancouver, Canada, August 4, 2017.

information is scarce. The study by Alvaro et al. (2015) is perhaps the most closely related work to ours. The authors annotated 1,548 tweets for whether they contain "first-hand experiences" of adverse drug reactions (ADRs) to prescription medications, and they used this annotated data in a supervised classification framework aimed at automatically identifying tweets that report personal usage. As far as we are aware, however, they have not made their annotated data public; nonetheless, we do not believe that it would have been exactly the right training set for our classification task. Because our focus is to help set the groundwork for using social media data in medication-related cohort studies, we included a subtle but key factor in our criteria for identifying personal intake: *when* the medication was taken. We will discuss this factor in more detail in the next section. In this paper, we present (i) an analysis of medication-mentioning chatter on Twitter, (ii) a publicly available, annotated corpus of tweets that can be used to advance automatic systems, and (iii) baseline supervised classification results to validate the utility of the annotated data.

2 Method

We chose Twitter as the data source for this study because of its growing popularity in public health research, and its easy-to-use public APIs. We discuss the three primary tasks—data collection, annotation, and classification—in the following subsections.

2.1 Data Collection

To build the corpus, we queried 73,800 Twitter user timelines (that we collected for related work) for 55 medication names, including both prescription and over-the-counter medications, brand and generic names, and types of medications (e.g., *steroid*). Using a tool that was developed by Pimpalkhute et al. (2014), we generated frequent misspellings of the medications in order to expand the query. We then tokenized all of the tweets, using the ARK Twokenizer (O'Connor et al., 2010; Owoputi et al., 2013), and identified 35,075 tweets containing a target medication. To account for the linguistic idiosyncrasies of how Twitter users might express their medication intake, we randomly selected one medication tweet from the 18,033 timelines that included such a tweet, and we prepared them for annotation. For this paper, 10,260 tweets were annotated, with overlapping annotations for 1,026 (10%).

[1] The annotation guidelines and a sample of the annotated data are available at: https://healthlanguageprocessing.org/twitter-med-intake/

2.2 Annotation

In order to control for studying the effects of medication intake on subsets of individuals in a social media setting, we decided that tweets of interest should not only represent the author's *personal usage* of the target medication in the tweet; they should also indicate the *specific instance* in which the user took the mentioned medication, since researchers using social media data cannot physically observe and record when medications were taken. Only if the tweets provide this additional information about the time of intake can we potentially use Twitter data to assess causal associations between users' health information (also mined from social media data) and the usage of particular medications. As we mentioned earlier, the way that time factors into our definition of "intake" marks an important distinction between our annotated data and Alvaro et al.'s (2015).

We found that, under minimal guidance, intuitively agreeing on what constituted a personal intake of medication, given the above criteria, was very difficult. We attribute this difficulty to the wide range of linguistic patterns in which we found medication mentions occurring. In an effort to obtain high inter-annotator agreement and address the human disagreement that Alvaro et al. seek to overcome, we analyzed linguistic patterns in samples of the data and used this analysis to inform the development of annotation guidelines;[1] in addition, we limited the number of annotation classes to the three high-level classes that we thought were most directly relevant to the classification task at hand: *intake*, *possible intake*, and *no intake*.

We will summarize our analysis of the three classes of tweets here. *Intake* tweets indicate that (i) the medication was actually taken, (ii) the author of the tweet personally took the medication, and (iii) the medication was taken at a specific time. To illustrate (i), consider the following tweets:

(a) Migraine from hell… **Took** 6 Motrin and nothing's touching it
(b) I've **been off** adderall about a month now and I'm so much happier, but COMPLETELY useless. I'm like a child again.

(c) A lot of people hate on prednisone but **I feel better already**. #stuffworksforme
(d) this ibuprofen still ain't **kicked in** my head poundin

While only (a) uses a verb phrase that explicitly indicates intake (*took…*), we can infer from features of the other tweets that the medication was taken: (b) *being off* the medication, (c) experiencing the effects of the medication, and (d) waiting for the medication to *kick in* all entail that the medication was taken.

Moreover, *intake* tweets should indicate that the *author* of the tweet took the medication:

(e) Sorry for this rant thingy, **I** took my Vyvanse today lol
(f) Sick and only had a Tylenol PM at work so now **i** feel better but **i** am fighting sleep😊
(g) Just threw back these Xanax
(h) In soooo much pain tonight and Tylenol just isn't cutting it. Literally hurting all over

Through the use of the first-person reference *I*, (e) explicitly states that the author took the medication, and (f) explicitly attributes the experiential effects of the medication (*feel better, but fighting sleep*) to the author. While (g) and (h) do not explicitly reveal that the *author* took the medication (*threw back*) or is (not) experiencing the effect of the medication (*isn't cutting it*), respectively, the high degree of self-presentation in social media (e.g., Kaplan and Haenlein, 2010; Papacharissi, 2012; Seidman, 2013) allows us to infer that the authors are writing about their own intake and experiences.

Finally, *intake* tweets also specify *when* the medication was taken:

(i) I've been sick **for the last 3 days** taking Ibuprofen just feel better and to fight Infection "swelling"
(j) Tylenol is my bestfriend **at the moment**
(k) maybe i'm tired as had 2 tramadol my bk is sore sore sore... #scoliosis
(l) Prednisone headache! Ahhhh

Tweet (i) uses a temporal marker that explicitly specifies an instance of intake, and, similarly, (j) explicitly indicates when the effect of the intake occurred. Although (k) and (l) do not explicitly specify instances of intake, Twitter's real-time nature (Sakaki et al., 2010) gives us reason to believe that the author of (k) recently *had* the medication and that the effect in (l) is being currently experienced, which represents an intake in the recent past (i.e., a specific instance).

Unlike *intake* tweets, some tweets do not specify that the author actually took the medication or when the medication was taken, but, unlike *no intake* tweets, are generally about the author's intake. Consider the following tweets:

(m) I want to cry it's that painful 😩**gonna** take codeine this morning for sure
(n) 800 mg of Advil **cause this headache is real**
(o) **I need** a Xanax like right now
(p) Codeine is **one hell of a drug**. 😔😔😔
(q) 😩😩😩 I never understood why I get so angryyyy omg **I was so mellow** on Xanax 😏
(r) I pretty much eat Advil **like it's candy**.💊🍬

We consider a tweet to be a *possible intake* if it expresses the intake as a future event (m); it contains merely a purpose for intake (n); it expresses a present-tense need for the medication (o); it abstractly praises (or criticizes) the medication without describing a concrete effect (p); it indicates that the author has used the medication in the past, but does not specify when (q); or, similarly, it indicates that the author uses the medication frequently, but does not specify an instance of intake (r). We decided to distinguish *possible intake* tweets because they can direct us to a user's timeline for manual probing, where we may find, for example, that a series of tweets aggregate to form a sort of composite *intake* tweet.

In contrast to *intake* and *possible intake* tweets, *no intake* tweets are not about the author's intake of the medication. While some *no intake* tweets are not about intake at all, some may be about the intake by others, not the author:

(s) @[Username redacted] Mine hurt for days last year!! **Take** some paracetamol **hun** ☺
(t) **Gave James** 2 ibuprofen pm and I'm being repaid by the sound of him snoring penetrating through my earplugs

The act of suggesting a medication (s) or giving someone a medication (t) might be interpreted as implying that the author has taken the medication in the past (i.e., a *possible intake*), but, because the

tweets are not primarily about the author's intake, we consider this inferential leap to be too large to warrant the same classification as other *possible intake* tweets.

While (s) and (t) are explicitly not about the author's intake, other tweets may not be as obvious, such as tweets that contain merely the name of a medication:

(u) @[Username redacted] @[Username redacted] @[Username redacted] @[Username redacted] **methadone** !

Although (u) also might be interpreted as indicating the author's use of the medication, the textual evidence does not seem to favor this interpretation over other possible ones, such as mere question-answering. We classify tweets that contain merely the name of a medication as *no intake* because, unlike *intake* and *possible intake* tweets, they do not contain enough information for us to conclude that they are about the author's intake.

The "addressivity" (Bakhtin, 1986) markers "@" in (u) reflect the "dialogic" (Bakhtin, 1981) space of social media, wherein the linguistic data that we are mining is not only textual, but "intertextual" (Kristeva, 1980)—that is, oriented to what has already been said by others. Tweets also mark this social orientation to others through features of "reported speech" (Voloshinov, 1973). Consider the following tweets:

(v) @[Username redacted] **"**I don't either cause these Tylenol aren't doing crap!**"** Lol
(w) I just wanna give a shoutout to adderall for helping me get through the semester **- Florida State**

While (v) and (w) would otherwise be classified as *intake* tweets, the quotation marks in (v) and the hyphen in (w) mark that the authors are directly reporting the words of others—in (w), a student at Florida State—not their own medication intake.

Other cases of reported speech involve tweets that make cultural references about taking medications—for example, song lyrics or lines from movies. As our analysis of the three classes suggests, identifying indications of personal medication intake in social media required grappling with a number of annotation issues, which forecast the challenges of using this data to train classifiers.

[2] Available at: http://www.cs.waikato.ac.nz/ml/weka/. Accessed: 5/25/2017.

2.3 Classification

We performed supervised classification experiments using several algorithms. The goal for these experiments was not to identify the best performing classification strategy, but to (i) verify that automatic classifiers could be trained using this data, and (ii) generate baseline performance estimates.

We used stratified 80-20 (training/test) split of the annotated set for the experiments. As features, we used only word n-grams (n = 1, 2, and 3) following standard preprocessing (e.g., stemming using the Porter stemmer (Porter, 1980) and lowercasing). We experimented with four classifiers—naïve bayes (NB), support vector machines (SVM), random forest (RF), logistic regression (LR), and a majority-voting based ensemble of the last three. Pairwise classification (i.e., 1-vs-1) is used to adapt the SVMs to the multiclass problem. Parameter optimization for the individual classifiers was performed via 10-fold cross validation over the training set, with an objective function that maximizes the F-score for the *intake* class.

Following the classification experiments, we performed brief error and feature analyses to identify common misclassification patterns and possible future approaches for improving classification performance. To identify informative n-grams for the intake class, we applied the *Information Gain* feature evaluation technique, which computes the importance of an attribute with respect to a class according to the following equation:

$$IG(Class, Attribute) = H(Class) - H(Class|Attribute)$$

$H()$ represents the information entropy for a given state (Yang and Pedersen, 1997). We used the Weka 3 tool[2] for all machine learning and feature analysis experiments. We present the results for these experiments in the next section.

3 Results and Discussion

In this section, we present and discuss the results of annotation and the baseline classification experiments, including a brief error analysis of misclassified *intake* tweets and a feature analysis to identify informative n-grams.

3.1 Annotation

For the corpus that we present in this paper, two expert annotators have annotated 10,260 tweets,

with overlapping annotations for 1,026 (10%). Their inter-annotator agreement was $\kappa = 0.88$ (Cohen's Kappa). They disagreed on 81 tweets, which the first author of this paper resolved through independent annotation. In total, 1,952 tweets (19%) were annotated as *intake*; 3,219 (31%) were annotated as *possible intake*; and 5,089 (50%) were annotated as *no intake*. These frequencies suggest that a minority of tweets that mention medications represent personal intake, which substantiates the need for this classification when mining large amounts of social media data for drug safety surveillance.

3.2 Classification

Table 1 presents the performances of the different classifiers. The overall accuracy (Acc) over the three classes and the F-scores (F) for each of the three classes are shown. The *no intake* (NI) class has the best F-score due to the larger number of training instances. SVMs, RF and LR classifiers have comparable accuracies, and they outperform the NB baseline. SVMs have the highest F-score for the *intake* (I) class, suggesting that it might be the most suitable classifier for this task.

The voting-based ensemble of the three classifiers does not improve performance over the SVMs. Post-classification analyses revealed that this is because the individual classifiers in the ensemble, particularly the LR and SVMs classifiers, make almost identical predictions given the feature set of n-grams. The confusion matrices for the classifiers' predictions are also alike, with strong inter-classifier agreements in terms of false and true positives and negatives. The results and the analyses suggest that incorporating/generating features that are more informative is more likely to improve performance on this task, rather than combining multiple classifiers on the same feature vectors.

	I (F)	PI (F)	NI (F)	Acc (%)	95% CI
NB	0.59	0.58	0.73	64.4	62.4-66.3
SVM	0.67	0.69	0.80	73.4	71.5-75.1
RF	0.60	0.68	0.80	72.2	70.4-74.0
LR	0.65	0.68	0.79	72.5	70.7-74.3
Ensemble	0.67	0.69	0.80	73.3	71.4-75.1

Table 1: Class-specific F-scores and accuracies for four classifiers and ensemble

The promising results obtained from automatic classification verify that our annotated dataset may indeed be used for training automated classification systems. Including more informative features is likely to further improve performance, particularly for the smallest (*intake*) class.

3.3 Error and Feature Analyses

An analysis of the false negative results of the *intake* class from the SVM classifier suggests that the majority of the errors (62%) could be attributed to the *implicit* indication that (i) the medication was taken, (ii) the author of the tweet personally took the medication, or (iii) the medication was taken at a specific time. In 69% of these cases, the *intake* tweet did not explicitly state (i), that the medication was taken. The next largest set of misclassified *intake* tweets comprised instances where the *intake* tweets contain lexical features that seem to frequently occur in the other classes (e.g., negation). Incorporating semantic features into the SVM classifier is likely to improve classification of the *intake* tweets.

Table 2 presents the 15 most informative n-grams for distinguishing the *intake* class from the others, as identified by the information gain measure. The table suggests that certain personal pronouns and explicit markers of personal consumption (e.g., *I took*), information about effectiveness (e.g., *not working*), and expressions indicating the need for a medication (e.g., *need a*) are useful n-grams for the classification task.

i	*not helping*	*i ve taken*
took	*i need*	*not working*
i took	*ve been taking*	*still in*
took some	*took two*	*need a*
to kick in	*i ve taken*	*just took*

Table 2: Most informative n-grams that distinguish the *intake* class from the others

4 Conclusion

In this paper, we presented a brief analysis of what we consider to be linguistic representations of personal medication intake on Twitter. This linguistic analysis informed our manual annotation of 10,260 tweets. We presented baseline supervised classification results that suggest that this annotated corpus can be used for training automated classification systems to detect personal medication intake in large amounts of social media data, and we will seek to improve the performance of our classifiers in future work. We believe that this classification is an important step towards broadening the use of social media for surveillance of drug safety.

References

Artur Akbarov, Evangelos Kontopantelis, Matthew Sperrin, Susan J. Stocks, Richard Williams, Sarah Rodgers, Anthony Avery, Iain Buchan, and Darren M Ashcroft. 2015. Primary care medication safety surveillance with integrated primary and secondary care electronic health records: A cross-sectional study. *Drug Safety*, 38(7):671–682, July.

Nestor Alvaro, Mike Conway, Son Doan, Christoph Lofi, John Overington, and Nigel Collier. 2015. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. *Journal of Biomedical Informatics*, 58: 280-287, December.

Mikhail M. Bakhtin. 1981. *The Dialogic Imagination*. University of Texas Press, Austin, TX.

Mikhail M. Bakhtin. 1986. Speech Genres & Other Essays. University of Texas Press, Austin, TX.

David Andre Broniatowski, Mark Dredze, Michael J. Paul, and Andrea Dugas. 2015. Using social media to perform local influenza surveillance in an inner-city hospital: A retrospective observational study. *JMIR Public Health Surveill 2015; 1(1):e5 https://publichealth.jmir.org/2015/1/e5/*, 1(1):e5.

Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. PEW Research Center Social Media Update 2016.

Carl L. Hanson, Scott H. Burton, Christophe Giraud-Carrier, Josh H. West, Michael D. Barnes, and Bret Hansen. 2013a. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62, April.

Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. 2013b. An exploration of social circles and prescription drug abuse through Twitter. *Journal of Medical Internet Research*, 15(9):e189, January.

Andreas M. Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1):59-68, January-February.

Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H. Gonzalez. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62:148–158, August.

Julia Kristeva. 1980. *Desire in Language: A Semiotic Approach to Literature and Art*. Columbia University Press, New York.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. In *Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden.

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):671–81, March.

Brendan O'Connor, Michael Krieger, and David Ahn. 2010. TweetMotif: Exploratory search and topic summarization for Twitter. In Proceedings of the *Fourth International AAAI Conference on Weblogs and Social Media*, pages 384-385, Washington, DC.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of the *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380-391, Atlanta, GA.

Zizi Papacharissi. 2012. Without you, I'm nothing: Performances of the self on Twitter. *International Journal of Communication*, 6:1989-2006,

Pranoti Pimpalkhute, Apurv Patki, Azadeh Nikfarjam, and Graciela Gonzalez. 2014. Phonetic spelling filter for keyword selection in drug mention mining from social media. In *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, pages 90–5.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Katrina M. Romagnoli, Scott D. Nelson, Lisa Hines, Philip Empey, Richard D Boyce, and Harry Hochheiser. 2017. Information needs for making clinical recommendations about potential drug-drug interactions: a synthesis of literature review and interviews. *BMC Medical Informatics and Decision Making*, 17(21).

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the *19th International World Wide Web Conference*, pages 851-860, Raleigh, NC.

Gwendolyn Seidman. 2013. Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Differences*, 54(3):402-407, February.

Laura Louise Struik and Neill Bruce Baskerville. 2014. The role of Facebook in Crush the Crave, a mobile- and social media-based smoking cessation intervention: Qualitative Framework Analysis of

posts. *Journal of Medical Internet Research*, 16(7):e170, July.

Valentin N. Voloshinov. 1973. *Marxism and the Philosophy of Language*. Seminar Press, New York.

Yiming Yang, and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, pages 412–420.

Li Zhou, Neil Dhopeshwarkar, Kimberly G Blumenthal, Foster R. Goss, Maxim Topaz, Sarah P. Slight, and David W. Bates. 2016. Drug allergies documented in electronic health records of a large healthcare system. *Allergy*, 71(9):1305–1313, September.