Journal of the American Medical Informatics Association, 26(12), 2019, 1618–1626
doi: 10.1093/jamia/ocz156
Advance Access Publication Date: 27 September 2019
Research and Applications

Research and Applications

Deep neural networks ensemble for detecting medication mentions in tweets

Davy Weissenbacher,[1] Abeed Sarker ,[1] Ari Klein ,[1] Karen O'Connor,[1] Arjun Magge,[2] and Graciela Gonzalez-Hernandez[1]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and [2]Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University, Tempe, Arizona, USA

Corresponding Author: Davy Weissenbacher, PhD, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, 480-492-0477, 404 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA; dweissen@pennmedicine.upenn.edu

Received 28 March 2019; Revised 26 July 2019; Editorial Decision 8 August 2019; Accepted 13 August 2019

ABSTRACT

Objective: Twitter posts are now recognized as an important source of patient-generated data, providing unique insights into population health. A fundamental step toward incorporating Twitter data in pharmacoepidemiologic research is to automatically recognize medication mentions in tweets. Given that lexical searches for medication names suffer from low recall due to misspellings or ambiguity with common words, we propose a more advanced method to recognize them.

Materials and Methods: We present *Kusuri*, an Ensemble Learning classifier able to identify tweets mentioning drug products and dietary supplements. *Kusuri* (薬, "medication" in Japanese) is composed of 2 modules: first, 4 different classifiers (lexicon based, spelling variant based, pattern based, and a weakly trained neural network) are applied in parallel to discover tweets potentially containing medication names; second, an ensemble of deep neural networks encoding morphological, semantic, and long-range dependencies of important words in the tweets makes the final decision.

Results: On a class-balanced (50-50) corpus of 15 005 tweets, *Kusuri* demonstrated performances close to human annotators with an F_1 score of 93.7%, the best score achieved thus far on this corpus. On a corpus made of all tweets posted by 112 Twitter users (98 959 tweets, with only 0.26% mentioning medications), *Kusuri* obtained an F_1 score of 78.8%. To the best of our knowledge, *Kusuri* is the first system to achieve this score on such an extremely imbalanced dataset.

Conclusions: The system identifies tweets mentioning drug names with performance high enough to ensure its usefulness, and is ready to be integrated in pharmacovigilance, toxicovigilance, or more generally, public health pipelines that depend on medication name mentions.

Key words: social media, pharmacovigilance, drug name detection, ensemble learning, text classification

INTRODUCTION

Twitter has been utilized as an important source of patient-generated data that can provide unique insights into population health.[1] Many of these studies involve retrieving tweets that mention drugs, for tasks such as syndromic surveillance,[2,3] pharmacovigilance,[4] and monitoring drug abuse.[5] A common approach is to search for tweets containing lexical matches of drug names occurring in a manually compiled dictionary. However, this approach has several limitations. Many tweets contain drugs that are misspelled or not referred to by name (eg, "it" or "antibiotic"). Even when a match is found, oftentimes the referent is not actually a

drug; for example, tweets that mention *Lyrica* are predominantly about the singer, Lyrica Anderson, and not about the antiepileptic drug. In this study, when using the lexical match approach on a corpus where names of drugs are naturally rare, we retrieved only 71% of the tweets that we manually identified as mentioning a drug, and more than 45% of the tweets retrieved were noise. Enhancing the utility of social media for public health research requires methods that are capable of improving the detection of posts that mention drugs.

The task of automatically detecting mentions of concepts in text is generally referred to as named entity recognition (NER).[6] State-of-the-art NER systems are based on machine learning (ML) and achieve performances close to humans when they are trained and evaluated on formal texts. However, they tend to perform relatively poor when they are trained and evaluated on social media.[7,8] Tweets are short messages, so they do not provide large contexts that NER systems can use to disambiguate concepts. Furthermore, the colloquial style of tweets—misspellings, elongations, abbreviations, neologisms, and nonstandard grammatical structures, cases, and punctuations—poses challenges for computing features in ML-based NER systems.[9] Although large sets of annotated corpora are available for training NER systems to detect general concepts on Twitter (eg, people, organizations), there is a need for the collection and annotation of additional data for automatic detection of more specialized concepts (eg, drugs and diseases).[10]

Over the last decade, researchers have competed to improve NER on tweets. Most challenges were organized for tweets written in not only English (eg, the Named Entity Recognition and Linking challenges series[11] or the Workshop on Noisy User-generated Text series)[12], but also other languages (eg, Conference sur l'Apprentissage Automatique 2017[13]). Specifically for drug detection in Twitter, we organized the Third Social Media Mining for Health Applications shared task in 2018.[14] The sizes of the corpora annotated during these challenges vary, from 4000 tweets[15] to 10 000 tweets.[13] ML methods have evolved over recent years, with a noticeable shift from support vector machine (SVM)– or conditional random field–based frameworks trained on carefully engineered features,[16] to deep neural networks that automatically discover relevant features from word embeddings. In the 2016 Workshop on Noisy User-generated Text, a large disparity between the performances obtained by the NERs on different types of entities was observed. The winners,[17] with an overall F_1 score of 52.4%, reported an F_1 score of 72.61% on Geo-Locations, the most frequent NEs in their corpus, but much lower scores on rare NEs, such as an F_1 score of 5.88% on TV Shows. *Kusuri* (薬, "medication" in Japanese; https://en.wiktionary.org/wiki/%E8%96%AC) detects names of drugs with sufficient performance even on a natural corpus in which drugs are very rarely mentioned.

The primary objective of this study was to automatically detect tweets that mention drug products (prescription and over-the-counter) and dietary supplements. The Federal Drug Administration (FDA Glossary of Terms: https://www.fda.gov/drugs/informationondrugs/ucm079436.htm; Drug; Drug product) defines a drug product as the final form of a drug, containing the drug substance generally in association with other active or inactive ingredients. This study includes drug products that are referred to by their trademark names (eg, NyQuil), generic names (eg, acetaminophen), and class name (eg, antibiotic or seizure medication). We formulate this problem as a binary classification task. Formally, given a set of tweets T, our goal is to learn a function f such that f(t)=1 for all tweets t in T containing at least 1 phrase referring to a drug product/dietary supplement, f(t)=0 otherwise. A tweet is a positive example if it contains text referring to a drug (and not only "matching" a drug name), a negative example otherwise. For example, the tweet "I didn't know Lyrica had a twin" is a negative example because Lyrica refers to the singer, Lyrica Anderson, whereas the tweet "Lyrica experiences? I was on Gabapentin." is a positive example because, in this context, it mentions 2 antiepileptics. The use of sequence labeling to delimit drug name boundaries in the positive examples (named entity recognition) and their mapping to a standardized name (named entity identification) are outside the scope of this work.

The main contributions of this study are (1) a gold standard corpus of 15 005 annotated tweets, for training ML-based classifiers to automatically detect drug products mentioned on Twitter; (2) a binary classifier based on Ensemble Learning, which we call *Kusuri*; and (3) an evaluation of *Kusuri* on 98 959 tweets with the natural balance of 0.2% positive to 99.8% negative for the presence of medication names. We describe the corpora in the Materials and Methods as well as the details of our classifier followed by its evaluation in the Results.

Automatic drug name recognition has mostly been studied for extracting drug names from biomedical articles and medical documents, with several articles published[18] and challenges organized in the last decade.[19–21] Most works that have tackled the task of detecting drug names in Twitter have focused on building corpora. Sarker et al[22] created a large corpus of 260 000 tweets mentioning drugs. However, they restricted their search by strict matching to a preselected list of 250 drugs plus their lexical variants, and they did not annotate the generated corpus. In a similar study, Carbonell et al[10] explored the distributions of drug and disease names in Twitter as preliminary work for drug-drug interaction. While the authors searched for a larger set of drugs (all unambiguous drugs listed in DrugBank database), they did not annotate the corpus of 1.4 million tweets generated, and nor did they count false positives—ambiguous mentions—in their statistical analysis.

The first evaluation of automatic drug name recognition in Twitter that we are aware of was performed by Jimeno-Yepes et al,[23] on a corpus of 1300 tweets. Two off-the-shelf classifiers, MetaMap and the Stanford NER tagger, as well as an in-house classifier based on a conditional random fields with hand-crafted features, were evaluated. The latter obtained the best F_1 score of 65.8%. Aside from the aforementioned problem of selecting the tweets using a lexical match, other limitations to their study lie in additional choices made. To remove nonmedical tweets, they retained only tweets containing at least 2 medical concepts (eg, drug and disease). This ensured a good precision, but also artificially biased their corpus in 2 ways: by retaining only the tweets that mentioned the drugs in their dictionary and eliminating tweets that mention a drug alone (eg, "me and ZzzQuil are best friends"). In November 2018, we organized the Third Social Media Mining for Health Applications shared task (SMM4H),[14] with Task 1 of our challenge dedicated to the problem of the automatic recognition of drug names in Twitter. Eleven teams tested multiple approaches on the provided balanced corpus, which we selected using 4 classifiers (the first module of *Kusuri*). A wide range of deep learning–based classifiers were used by participants, as well as some feature-based classifiers and a few attempts with ensemble learning systems. The system THU_NGN by Wu et al,[24] an ensemble of hierarchical neural networks with multihead self-attention and integrating features modeling sentiments, was the top performer, with an F_1 score of 91.8%. This established a recent benchmark for the community for an artificially balanced corpus (with approximately the same number of positive and negative examples). Our evaluation data, described in the UPennHLP Twitter Pregnancy Corpus

subsection, includes both the artificially balanced corpus and, in addition, a corpus of all available tweets posted by selected Twitter users where the mentions of drug products were manually annotated. We refer to the later as a corpus with "natural" balance.

MATERIALS AND METHODS

We collected all publicly available tweets posted by 112 500 Twitter users (their timelines). To do so, we first used the Twitter streaming application programming interface to detect tweets mentioning keywords used when announcing a pregnancy. These keywords were manually defined. Then, we used a simple SVM classifier to confirm that the tweets were really announcing a pregnancy and discarded other tweets. The keywords and the SVM classifier are described in our previous work.[25] Once a tweet announcing a pregnancy was identified, we collected the timeline of the author of the tweet. We used the REST application programming interface provided by Twitter to download all tweets posted by this user within the Twitter-imposed limit of 3200 most recent tweets, and continued collection afterward. We did not remove bots or accounts managed by businesses and other entities, as they may also tweet about drugs. We intermittently collected posts from January 2014 to April 2017. After April 2017, we systematically collected posts until September 2017. Through this process, we collected a total of 421.5 million tweets. Using this dataset as a source allows us to avoid the bias of a drug-name keyword based collection. When a dataset is collected using a list of drug names, the resulting dataset will obviously contain only tweets mentioning the drugs occurring in the list: the you-find-what-you-are-looking-for bias (ie, confirmation) bias. This is evident when reported recall climbs to 98% or more. Our method, collecting all tweets posted by the users in our cohort, captured drugs mentioned by the users in the way that they naturally occur. Our dataset represents natural variants of drug names occurring in our collection, as expressed by Twitter users, and that would have been missed if not present in the list upfront.

All tweets were collected from public Twitter accounts, and a certificate of exemption was obtained from the Institutional Review Board of the University of Pennsylvania. All tweets used and released to the community were used and released without violating Twitter's terms and conditions.

UPennHLP Twitter Drug Corpus

Building a corpus of tweets containing drug names to train and evaluate a drug name classifier is a challenging task. Tweets mentioning drug names are extremely rare. We found that they only represent 0.26% of the tweets in the UPennHLP Twitter Pregnancy Corpus (see the following section), and are often ambiguous with common and proper nouns. If a naive lexicon matching method is used to create the corpus, it often matches a large number of tweets not containing any drug names.

Therefore, to build a gold-standard corpus, we had to rely on a more sophisticated method than simply lexicon matching. We created 4 simple classifiers to detect tweets mentioning drug names: one based on a lexicon matching, one on lexical variants matching, one on regular expressions, and a classifier trained with weak supervision. The 4 classifiers are described briefly below, and in detail in Supplementary Appendix A.

Lexicon-based drug classifier

The first classifier is built on top of a lexicon of drug names generated from the RxNorm Database (https://www.nlm.nih.gov/research/umls/rxnorm/overview.html, Accessed June 11, 2018). If a tweet contains a word or phrase occurring in the lexicon, the tweet is classified as a positive example without any further analysis. We chose RxNorm because it has a large coverage. It combines, in a unique database, 15 existing dictionaries, including DrugBank, a database often used in previous works on drug names detection.

Variant-based drug classifier

Names of drugs may have a complex morphology and, as a consequence, are often misspelled on Twitter. Lexicon-based approaches detect drugs mentioned in tweets only if the drug names are correctly spelled. The incapability to detect misspelled drug names results in low recall for the lexicon-based classifier. In an attempt to increase recall, we used a data-centric misspelling generation algorithm[26] to generate variants of drug names and used the variants to detect tweets mentioning misspelled drugs.

Weakly trained drug long short-term memory classifier

Our third classifier is a long short-term memory (LSTM) neural network that integrates an attention mechanism[27] and is trained on noisy training examples obtained through weak supervision. One annotator identified drug names tending to be unambiguous[28] in our timelines (eg, Benadryl or Xanax). We selected the ~126 500 tweets containing these unambiguous names in our timelines as positive examples. Then, given that drug names occur very rarely in tweets, we randomly selected an additional ~126 500 tweets from our timelines as negative examples and trained our LSTM on these examples. We chose this simple classifier to discover a large number of tweets that could potentially contain drug names, and integrated an attention mechanism to ensure that the neural network focuses on the words occurring recurrently in the context of drug mentions in tweets and discards irrelevant words.

Pattern-based drug classifier

Our last classifier implements a common method to detect general named entities, regular expressions (REs). REs describe precisely the linguistic contexts used in Twitter to speak about drugs. We manually crafted our REs by inspecting 9530 n-grams, which were the most frequent n-grams occurring before and after the most frequent unambiguous names of drugs in our ~126 500 tweets. We retained 81 patterns (eg, "prescibed me", "prescription filled for", "doctor switched"). When inspecting these n-grams, one annotator used his knowledge of the language to reject noisy patterns. Before including a pattern in the list, the annotator confirmed empirically by querying the pattern in a search engine indexing our dataset that the pattern actually retrieved tweets mentioning drugs in the first 100 tweets retrieved for the query. Davy Weissenbacher created the REs.

To obtain positive examples, we selected tweets retrieved by at least 2 classifiers, as they were most likely to mention drug names. To obtain negative examples, we selected tweets detected by only 1 classifier, given that if these tweets did not contain a drug name, they were nonobvious negative examples. Following this process, from our 421.5 million tweets, we created a corpus of 15 005 tweets, henceforth referred to as the UPennHLP Twitter Drug Corpus (Table 1). We removed from the corpus duplicated tweets, tweets not written in English, and tweets that were no longer on Twitter (eg, tweets deleted by the users) at the time of the collection. Two annotators annotated the corpus in its entirety, with a high interannotator agreement (IAA) measured as Cohen's kappa of .892. Our corpus was annotated by our 2 staff annotators, who have over 7 years combined experience annotating texts in the

Table 1. Statistics of the UPennHLP Twitter Drug and Pregnancy Corpora

	UPennHLP Twitter Drug Corpus	UPennHLP Twitter Pregnancy Corpus
Training set	9623 tweets (4975 +/4648 −)	69 272 tweets (181 +/69 091 −)
Testing set	5382 tweets (2852 +/2530 −)	29 687 tweets (77 +/29 610 −)
Users in training/ testing set	7584/4535	112/112
Users posting in training set and in testing set	1054 (these users posted 1713 tweets in testing set, 31.8% of testing set)	112 —

biomedical domain. One annotator holds a degree in biology and our senior annotator has a master of science degree in biomedical informatics. We randomly selected 9623 tweets for the training set (4975 positive and 4648 negative examples) and 5382 tweets for the testing set (2852 positive and 2530 negative examples). We publicly released the corpus and our guidelines to the research community for Task 1 of the SMM4H 2018 shared task.[14]

UPennHLP Twitter Pregnancy Corpus

A balanced corpus, such as the UPennHLP Twitter Drug Corpus, is a useful resource to study how people speak about drugs on social media. However, due to the mechanism of its construction, a balanced corpus does not represent the natural distribution of tweets mentioning drugs on Twitter. Consequently, any evaluation made on a balanced corpus will never be indicative of the performance expected from a drug name classifier used in practice. To further assess whether *Kusuri* could reliably be used in practice, we ran additional experiments on the corpus used for an epidemiologic study of birth defect outcomes.[29] For that, we collected 397 timelines of women that had announced their pregnancy on Twitter and that had tweeted during their pregnancy, and manually identified all tweets mentioning a drug during the period of pregnancy. It took an average of 2.5 hours to annotate each timeline. We ran our experiments on a subset of 112 timelines (98 959 tweets) from this corpus, referred to as UPennHLP Twitter Pregnancy Corpus in the remainder of the article (Table 1). The annotators manually verified that these timelines were owned by individuals by looking at their profiles and their posts and making a judgment call, and removed all timelines administered by bots or association/companies. Our senior annotator and the main author of this article annotated these timelines. An IAA of 0.88 (Cohen's kappa) was computed over 12 timelines, which were dual annotated by the senior annotator. We randomly selected 70% (69 272 tweets) of the *Pregnancy Corpus* for training and the remaining 30% (29 687 tweets) for testing. When splitting the corpus into a training set and a testing set, we kept the same percentage ratio, 70%/30%, of positive/negative examples in the training and in the testing sets (ie, 181/69 091 and 77/29 610 respectively).

Kusuri architecture

Kusuri, described in Figure 1, applies sequentially 2 modules to detect tweets mentioning drugs in the UPennHLP Twitter Pregnancy Corpus. This section describes each module. The success of deep learning classifiers in natural language processing lies on their ability to automatically discover relevant linguistic features from word embeddings[30]—an ability even more valuable when working on short and colloquial texts such as tweets. For this reason, we preferred to integrate in our modules deep learning classifiers over more traditional classifiers based on feature engineering.

Module 1: tweet prefilter

Kusuri applies our 4 classifiers—the lexicon-based, variant-based, pattern-based, and weakly trained LSTM classifiers—in parallel to discover tweets that potentially contain drug names. Among the tweets discovered, *Kusuri* selects the tweets classified by the lexicon classifier, by the variant-based classifier, and the tweets selected by both the pattern-based and the weakly trained classifiers. The tweets discovered by only 1 of the 2 last classifiers were too noisy and discarded. The tweets selected are then submitted to the Module 2, an ensemble of deep neural networks (DNN) that makes the final decision for the labels. The 4 classifiers act as filters, collecting only good candidates for the ensemble of DNNs, which was, in turn, optimized to recognize positive examples among them.

Module 2: ensemble of DNNs

As a single element for the ensemble of neural networks composing the second module of *Kusuri*, we designed a DNN following a standard architecture for classification of NEs. Described in Figure 2, our DNN starts by independently encoding each sequence of characters composing the tokens of a tweet through 3 layers sequentially connected: a recurrent layer, an attention layer, and a densely connected layer. All resulting vectors, encoding the morphological properties of the tokens, are then concatenated with their respective pretrained word embedding vectors, which encode the semantic properties of the tokens. The concatenated vectors are passed to a bidirectional-gated recurrent unit (GRU) layer to learn long-range dependencies between the words, followed by an attention layer that, as additional memory, helps the NN to focus on the most differentiating words for the classification. A final dense layer computes the probability for the tweet to contain a mention of a drug. All neural networks in our study were given pretrained word vectors as input. We chose the word vectors trained with the Glove algorithm on 2 billion tweets, available for download on the webpage of the project (https://nlp.stanford.edu/projects/glove/). Supplementary Appendix B describes in detail the preprocessing steps, the embeddings, and the parameters of our training. We experimented with early stopping to avoid overfitting when training our models. We kept 70% of our training corpus to train a model, and 30% for validation. We found that 8 iterations were sufficient to train the model before overfitting the training corpus. However, contrary to our expectation, the models trained with 8 iterations gave slightly lower performances on the test set of the UPennHLP Drug Corpus than models trained with 20 iterations, with F_1 scores of 92.7% and 93.1%, respectively. The reason for this is not clear, but it may be because our model continues to improve as more examples are provided in the training corpus, making the estimation of the best number of iterations inexact with early stopping. We report the results of our model trained with 20 iterations.

Owing to the stochastic nature of the initialization of the NN, the learning process may discover a local optimum and return a suboptimal model. To reduce the effect of local optimums, we resort to ensemble averaging. We independently learned 9 models using our DNN and computed the final decision, for a tweet to mention a medication name or not, by taking the mean of the probabilities

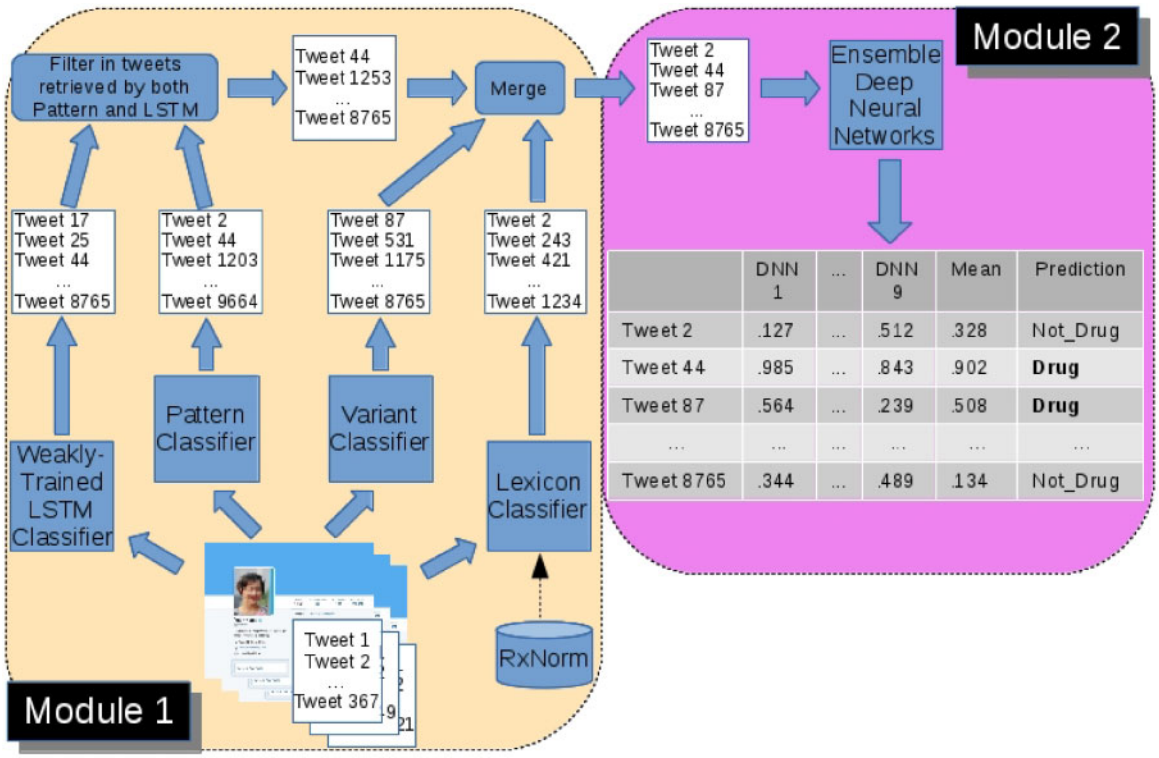

Figure 1. Architecture of *Kusuri,* an ensemble learning classifier for drug detection in Twitter. LSTM: long short-term memory.

computed by the models (because a soft voting algorithm[31] in our experiments did not improve over the simple averaging method, we kept the latter). When applied on the Pregnancy Corpus, all DNNs of the ensemble were trained on the Drug Corpus, and, at test time, all DNNs of the ensemble only have to classify the tweets in the Pregnancy Corpus filtered by the first module of *Kusuri*.

RESULTS

This section details the performances of *Kusuri* and its ensemble of DNNs during 2 series of experiments on the Drug Corpus and on the Pregnancy Corpus.

Drug detection in the UPennHLP Twitter Drug Corpus

We first ran a series of experiments to measure the performances of the ensemble of DNNs composing the second module of *Kusuri*. The detailed results are reported in Table 2. We compare the performance of our ensemble with 3 baseline classifiers.

The first baseline classifier is a combination of the lexicon- and variant-based classifiers. It labeled as positive examples all tweets of the test set of the Drug Corpus that contained a phrase found in our Lexicon or in our list of variants. This baseline classifier provides a good estimation of the performances to expect when a lexicon-based approach is used. We chose as a second baseline system a bidirectional-GRU classifier. Because this baseline system has a simpler architecture than our final DNN, their comparison allows us to estimate the benefits of the components we added in our system. The baseline system was trained on the same training data with the same hyperparameters, and took as input the same word embeddings, but it did not have information about the morphology of the words or the help of the attention mechanism. As a third strong baseline, we compare our system with the best system of the Task 1 of the SMM4H 2018 competition, the THU_NGN system.[24]

The results in Table 2 are interesting in several ways. The combined lexicon- and variant-based classifier has a high recall on the test data (88.5%), an unsurprising result considering the central role played by the lexicon and the variants during the construction of the Drug Corpus. This classifier is vulnerable to the frequent ambiguity of drug names, resulting in a low precision of 66.4%. The classifier has no knowledge of the context in which the name of a drug appears, and thus cannot disambiguate tweets mentioning *Lyrica* (antiepileptic vs Lyrica Anderson), *lozenge* (type of pills vs geometric shape), or *halls* (brand name vs misspelling for Hall's), for example. The fully supervised bidirectional-GRU confirms its ability to learn the features only from the word embeddings,[32] and achieves an F_1 score of 91.4%, a higher score than the IAA computed on this corpus. However, such systems can be improved as demonstrated by the better performances of the best DNN in the ensemble (4 in Table 2). The encoding of the token morphology and the attention layer of the best DNN improve the F_1 score by 1.7 points. Also, despite having a simpler architecture and attention mechanism, the best DNN system performs better than the ensemble of hierarchical NNs proposed by Wu et al.[24] The reason for this result is not clear, but it may be a suboptimal set of hyperparameters chosen by the authors or the difficulty to train such a complex network.

The highest performance is obtained by the ensemble DNNs, which shows an improvement of 0.6 points over the best model in the ensemble, with a final F_1 score of 93.7%. We confirmed the disagreement between the ensemble DNNs and the THU_NGN systems to be statistically significant with a McNemar test.[33] The null hypothesis was rejected with a significance level set to .001. We analyzed randomly selected labeling errors made by the ensemble DNNs (Table 3). We distinguished 8 nonexclusive categories of false positives. With 41 cases, most false positives were tweets discussing

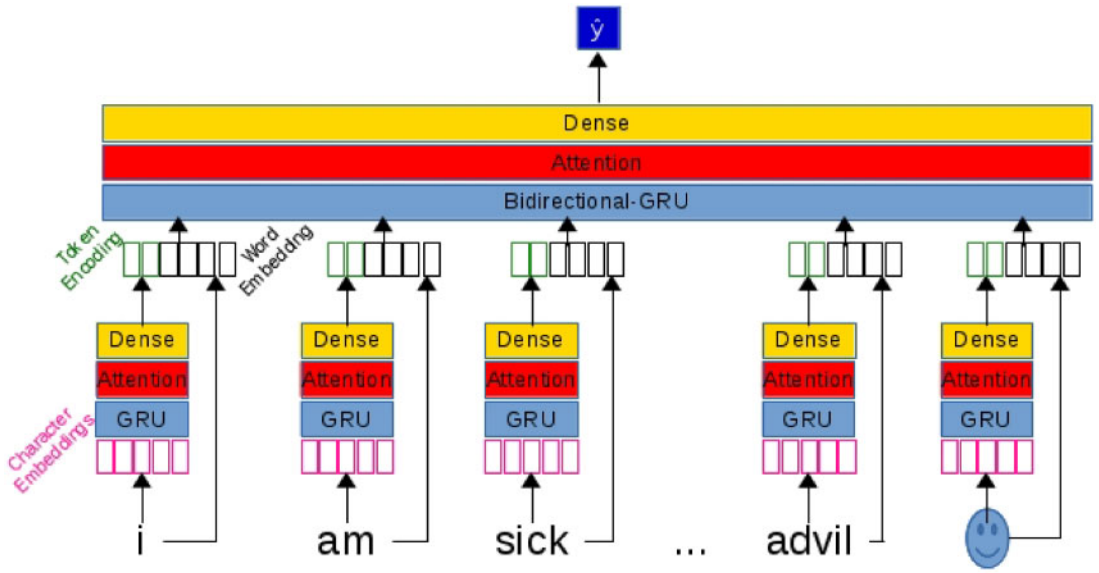

Figure 2. Deep neural network predicting ŷ, the probability for a tweet to mention a drug name. GRU: gated recurrent unit.

Table 2. Precision, recall, and F1 scores for drug detection classifiers on the test set of the UPennHLP Twitter Drug Corpus

System	Precision	Recall	F_1 score
1. Lexicon + variant classifier	66.4	88.5	75.9
2. Supervised bidirectional-GRU	93.5	89.5	91.4
3. THU_NGN hierarchical-NNs	93.3	90.4	91.8
4. Best DNN model in the ensemble	93.7	92.5	93.1
5. Ensemble DNNs (module 2 of *Kusuri*)	95.1	92.5	93.7

DNN: deep neural network; GRU: gated recurrent unit; NN: neural network; THU_NGN.

medical topics without mentioning a drug. As medical tweets often describe symptoms or discuss medical concepts, their lexical fields are strongly associated with drug names (eg, cough, flu, doctor) and confuse the classifier. The causes of false negatives seem to mirror those for the false positives. With 36 cases, false negatives were mostly caused by the ambiguity of not only common English words (eg, airborne), but also dietary supplements and food products sometimes consumed for their medicinal properties (eg, clove, arnica, aloe). This could be a positive turn of events if nutritional supplements are to be included in a study. A second important cause was unseen, or rarely seen, drug names in our training corpus, with 25 cases.

The ensemble DNNs correctly detected 245 tweets that were incorrectly detected by the THU_NGN system. On the other hand, the THU_NGN system correctly detected 138 tweets that were incorrectly detected by our system. We manually analyzed 100 tweets that were randomly selected from the 245 tweets, but could not discern any evident patterns explaining the differences in performances between the 2 systems. Further linguistic analysis, such as the analysis proposed in Vanni et al,[30] may help uncover these patterns, but is beyond the scope of this study.

Drug detection in the UPennHLP Twitter Pregnancy Corpus

The results of the ensemble DNNs on the Drug Corpus are promising, but they were obtained based on ideal conditions. The training corpus is balanced, and most of the drugs found in the test set were present in the training set. These conditions are unlikely to be satisfied when the classifier is used on naturally occurring data. We ran a second series of experiments on the Pregnancy Corpus to get a more realistic evaluation of our classifiers. As for the previous experiments, we kept the lexicon- and variant-based classifier as well as an ensemble of bidirectional-GRU networks as baseline systems (1 and 2 in Table 4, respectively). Each network of the ensemble was trained on the Pregnancy Corpus with 5 iterations and a batch size of 64 examples, and a simple averaging was used to combine their results. Since the ensemble DNNs gave the best performances on the Drug Corpus (Table 2), we chose it as a third baseline system. This baseline applies the ensemble of DNNs without prefiltering the tweets using the first module of *Kusuri*. In system 3.a, we trained all DNNs of the ensemble on the training set of the Drug Corpus, with 20 iterations and a batch size of 2 examples. In system 3.b, we trained all DNNs of the ensemble on the training set of the Pregnancy Corpus, with 4 iterations and a batch size of 64 examples. These hyperparameters were the best parameters found after early stopping and a manual search through standard batch sizes of 2, 64, and 128. The last system evaluated was the "complete" *Kusuri* system, with both modules applied sequentially.

The results are reported in Table 4. What is striking about the figures in this table is the poor performances of the lexicon- and variant-based classifier and the ensemble DNNs classifier. The score of the former dropped from an F_1 score of 75.9% when applied on the Drug Corpus, to 62.2% when applied on the Pregnancy Corpus. As we used the lexicon to build the Drug Corpus, the drugs in the lexicon were overrepresented in this corpus, increasing the baseline's recall by 17.1 points. The ensemble DNNs classifier (3.a) did worse, with an F_1 score of only 18%. Trained on a balanced set of medically related tweets, the classifier was found too sensitive. It gives too much weight to words that are related to medication but used in other contexts such as *overdose*, *bi-polar*, or *isnt working*, resulting in a total of 549 false positives, in which only 77 tweets mentioned a drug in the test set. Surprisingly, the ensemble of bidirectional-GRU

Table 3. Categories of false positive and false negative made by the drug detection classifier on the test set of the UPennHLP Twitter Drug Corpus

Error category	Errors	Examples
False positive		
Medical topic	41	<user> you should see a dermatologist if you can. You may just need something to break you out of a cycle. I used a topical and took pills
		Lola may has a stye, or pink eye. Doc recommends warm compresses to see if it gets better today, but my eyes are itchy just looking at her.
Weighted words/ patterns	19	*i can take a* wax brazillian a g
		<user> i was robbed a foul when *i took a* three point shot and they got a few three pointers in. good game.
Ambiguous name	12	<user>I actually really like *Lyrica* & A1.
Food topic	11	This aerobically fermented product was tested & it's antibiotic residue free. also certified organic.
Insufficient context	7	<user> adding *Arnica* to my shopping list
Cosmetic topic	5	Doc prescribed me this dandruff shampoo, if it works, I'm definitely getting a sew in after I'm done using it
Unknown	2	Ice_Cream, Ice-Cream and More Ice-Cream. . .thats Ol i Want
Error annotation	3	–
False negative		
Ambiguous name	36	Trying Oil of Oregano & garlic for congestion for my sinus infection. [ambiguous dietary supplement]
		In the church the person close to me's sniffling & coughing. . . I need a bathe of bactine and some *Airborne*, right now [ambiguous English word]
Drug not/rarely seen	25	That's the *benzo* effects! [missing variant]
		Pennsylvania Appellate Court Revives 1, 000 *Prempro* Cases Against Pfizer [missing in lexicon]
		the *percocet-thief* plot makes Real World New Orleans look almost intriguing [preprocessing error]
Generic terms	18	Tossing and turning. I need ur *sleep aid*. Waiting patiently <user>
Nonmedical topic	11	<user>Meet Mr an Mrs Lexapro. . .... guarenteed fidelity.
Short tweets	3	arnica-ointment-7
Error annotation	7	–

networks (system 2.a) was capable of learning our task with very few positive examples. Despite the 5.11-point difference in F_1 score of *Kusuri* over system 2.a, a McNemar test shows that we cannot conclude this difference is significant. Additional experiments, with more positive examples, are needed to confirm *Kusuri*'s superiority. While lower than the ideal score of the ensemble DNNs classifier on the Drug Corpus, the F_1 score of *Kusuri* (78.8%) is comparable to the scores published for the best NERs when applied on the most frequent types of NEs in Twitter.[17] More importantly, we believe that this score is high enough to expect a positive impact of *Kusuri* when integrated in larger applications.

DISCUSSION

As stated before, our experiments show Kusuri outperforming the system 2.a (Table 4), a bidirectional GRU, but the difference is not statistically significant. Increasing the number of positive examples to the point at which the performance difference is statistically significant would require the collection and annotation of a much larger corpus that exhibits the natural balance (<1% of tweets positive for a drug mention). This could prove cost prohibitive.

In this study, we opted for an oversampling strategy and created an artificially balanced corpus to train our classifier, the UPenn Twitter Drug Corpus. However, while our ensemble of DNNs performs well on the Drug Corpus (an F_1 score of 93.7%), its performance drops considerably when we applied it to the Pregnancy Corpus (an F_1 score of 18%). Trained on our balanced corpus, this classifier was biased toward disambiguating examples easily recognizable by our basic filters and could not generalize well on other examples occurring in the Pregnancy Corpus, making the ensemble useless for real applications.

The solution we implemented with *Kusuri* is to prefilter the tweets of the timelines before applying our ensemble of DNNs. This solution increases performance by 5.1 points over the best baseline system (2.a in Table 4). However, our strategy is far from perfect, as it reduces the number of FPs with hard filters and, consequently, also limits the overall performance of *Kusuri* by removing 27% (21 tweets) of the few tweets mentioning drugs in the Pregnancy Corpus. We are currently replacing the hard filters with active learning to further train our ensemble of DNNs and reduce its oversensitivity to medical phrases in general tweets.

One may argue that, given that the selection of users in our cohort are women reporting a pregnancy, the drugs mentioned in our dataset are biased to a specific set of mediations, with a higher number of tweets mentioning drugs commonly used in pregnancy and a lower number of tweets mentioning drugs not recommended during this period. However, this limitation is alleviated to an extent due to the fact that our dataset includes tweets beyond those that were posted during pregnancy; we collected the full timelines available from the users, including a large number of tweets posted before and after pregnancy.

Finally, we designed our neural network around a standard representation of a sentence as a sequence of word embeddings learned on local windows of words. Better alternatives have recently been proposed[34] and could be integrated in our system to help drug name disambiguation. We could replace our word embeddings with ELMo or BERT, which learn each word embeddings within the whole context of a sentence,[35,36] or supplement our current sentence representation with sentence embeddings.[37]

Table 4. Precision, recall, F_1 scores, true positives, false positives, and false negatives for drug detection classifiers on the UPennHLP Twitter Pregnancy Corpus testing set

System	Precision	Recall	F_1 score	True positives/false positives/false negatives
1. Lexicon + variant classifier	55.0	71.43	62.15	55/45/22
2. Ensemble supervised bidirectional-GRUs				
a. Trained on UPennHLP Twitter Pregnancy Corpus	87.5	63.64	73.68	49/7/28
3. Ensemble DNNs (only module 2 [classifier] of *Kusuri*)				
a. Trained on UPennHLP Twitter Drug Corpus	10.15	80.52	18.02	62/549/15
b. Trained on UPennHLP Twitter Pregnancy Corpus	93.75	58.44	72.00	45/3/32
4. *Kusuri* (module 1 [filters] + module 2 [classifier])	94.55[a]	67.53	78.79[a]	52/3/25

DNN: deep neural network; GRU: gated recurrent unit; NN: neural network; THU_NGN.

CONCLUSION

In this article, we presented *Kusuri*, an ensemble learning classifier to identify tweets mentioning drug names. Given the unavailability of a labeled corpus to train our system, we created and annotated a balanced corpus of 15 005 tweets, the UPennHLP Twitter Drug Corpus. The ensemble of deep neural networks at the core of *Kusuri*'s decisions (Module 2) demonstrated performances close to human annotators without requiring engineered features on this corpus, with an F_1 score of 93.7%. However, because we built this corpus artificially, it did not represent the natural distribution of drug mentions in Twitter. We evaluated *Kusuri* on a second corpus, UPennHLP Twitter Pregnancy Corpus, made of all tweets posted by 112 Twitter users, a total of 98 959 annotated tweets, with only 258 tweets mentioning drugs. On this corpus, *Kusuri* obtained an F_1 score of 78.8%, a score comparable to the score obtained on the most frequent types of NEs by the best systems competing in well-established challenges, despite our corpus having only 0.26% positive instances in it. The code of Kusuri and the models used for these experiments are publicly available at https://bitbucket.org/pennhlp/kusuri/. The UPennHLP Twitter Drug Corpus is available at https://healthlanguageprocessing.org/kusuri. We will release the UPennHLP Twitter Pregnancy Corpus during the Fifth Social Media Mining for Health Applications shared task in 2020.

FUNDING

This work was supported by National Library of Medicine grant number R01LM011176 to GG-H. The content is solely the responsibility of the authors and does not necessarily represent the official view of National Library of Medicine.

AUTHOR CONTRIBUTIONS

DW designed the experiments, preprocessed the data and annotated a part of it, implemented *Kusuri* and computed the models, analyzed the prediction errors, and wrote the majority of the manuscript. AS integrated the variant-based drug classifier in *Kusuri*, wrote its description, and proofread the manuscript. AK edited the manuscript. KO annotated the data and computed the interannotator agreement. AM helped optimize the neural networks. GG-H supervised the overall study design and edited the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Sinnenberg, L Buttenheim, AM Padrez, K Mancheno, C Ungar, LMerchant RM. Twitter as a tool for health research: a systematic review. *Am J Public Health* 2017; 107 (1): e1–e8.
2. Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. *Artif Intell Med* 2014; 61 (3): 153–63.
3. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data. *J Med Internet Res* 2017; 19 (9): e315.
4. Magge A, Sarker A, Nikfarjam A, Gonzalez-Hernandez G. "Comment on: "deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts" *J Am Med Inform Assoc* 2019; 26 (6): 577–9.
5. Kazemi DM, Borsari B, Levine MJ, Dooley B. Systematic review of surveillance by social media platforms for illicit drug use. *J Public Health (Oxf)* 2017; 39 (4): 763–76.
6. Sekine S, Nobata C. Definition, dictionaries and tagger for extended named entity hierarchy. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04);* 2004: 1977–80.
7. Liu X, Zhang S, Wei F, Zhou M. Recognizing named entities in tweets. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*; 2011: 359–67. https://www.aclweb.org/anthology/P11-1037/
8. Sarker A, Belousov M, Friedrichs J, *et al*. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J Am Med Inform Assoc* 2018; 25 (10): 1274–83.
9. Ritter A, Clark S, Etzioni M, Etzioni O. Named entity recognition in tweets: an experimental study. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*; 2011: 1524–34.
10. Carbonell P, Mayer MA, Bravo À. Exploring brand-name drug mentions on twitter for pharmacovigilance. *Stud Health Technol Inform* 2015; 210: 55–9.
11. Rizzo G, Pereira B, Varga A, van Erp M, Basave AEC. Lessons learnt from the named entity recognition and linking (NEEL) challenge series. *Semant Web* 2017; 8 (5): 667–700.
12. Derczynski L, Nichols E, Erp MV, Limsopatham N. Results of the WNUT2017 shared task on novel and emerging entity recognition. In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*; 2017: 140–7.

13. Lopez C, Partalas I, Balikas G, *et al*. CAp 2017 challenge: twitter named entity recognition; *arXiv* 2017 Jul 24 [E-pub ahead of print].
14. Weissenbacher D, Sarker A, Paul MJ, Gonzalez-Hernandez G. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In: *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*; 2018: 13–16.
15. Strauss B, Toma BE, Ritter A, Marneffe M-C, Xu DW. Results of the wnut16 named entity recognition shared task. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*; 2016: 138–44.
16. Sileo D, Pradel C, Muller P, De Cruys TV. Synapse at CAp 2017 NER challenge: Fasttext CRF. *arXiv* 2017 Sept 14 [E-pub ahead of print].
17. Limsopatham N, Collier N. Bidirectional LSTM for named entity recognition in twitter messages. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text*; 2016: 145–52.
18. Liu S, Tang B, Chen Q, Wan X. Drug name recognition: approaches and resources. *Information* 2015; 6 (4): 790–810.
19. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010; 17 (5): 514–8.
20. Segura-Bedmar I, Martínez P, Herrero-Za M. Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In: *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*; 2013: 341–50.
21. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A. CHEMDNER: the drugs and chemical names extraction challenge. *J Cheminform* 2015; 7: S1.
22. Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from twitter chatter: language models and their utilities. *Data Brief* 2017; 10: 122–31.
23. Jimeno-Yepes A, MacKinlay A, Han B, Chen Q. Identifying diseases, drugs, and symptoms in twitter. *Stud Health Technol Inform* 2015; 216: 643–7.
24. Wu C, Wu F, Wu J, *et al*. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In: *Proceedings of the 2018 EMNLP Workshop SMM4H: the 3rd Social Media Mining for Health Applications Workshop and Shared Task*; 2018: 34–7.
25. Sarker A, Chandrashekar P, Magge A, Cai H, Klein A, Gonzalez G. Discovering cohorts of pregnant women from social media for safety surveillance and analysis. *J Med Internet Res* 2017; 19 (10): e361.
26. Sarker A, Gonzalez-Hernandez G. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *J Biomed Inform* 2018; 88: 98–107.
27. Shen S-S, Lee HY. Neural attention models for sequence classification: analysis and application to key term extraction and dialogue act detection. In: *Proceedings of INTERSPEECH'16*; 2016: 2716–20.
28. Grave E. Weakly supervised named entity classification. Workshop on Automated Knowledge Base Construction (AKBC); December 13, 2014; Montreal, Canada.
29. Golder SP, Chiuve S, Weissenbacher D, *et al*. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. *Drug Saf* 2019; 42: 389–400.
30. Vanni L, Ducoffe M, Mayaffre D, *et al*. Text deconvolution saliency (TDS): a deep tool box for linguistic analysis. In: *Proceedings of ACL'18, 56th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2018.
31. Raschka S, Mirjalili V. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow*. 2nd ed. Birmingham, UK: Packt Publishing Ltd; 2017.
32. Chalapathy R, Borzeshi E, Piccardi M. An investigation of recurrent neural architectures for drug name recognition. In: proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI); 2016: 1–5.
33. Dietterich TG. Approximate Statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998; 10 (7): 1895–923.
34. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multitask benchmark and analysis platform for natural language understanding. *BlackboxNLP@EMNLP*; 2018.
35. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* 2019 May 24 [E-pub ahead of print].
36. Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*; 2018: 227–37.
37. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; 2017: 670–80.