








## Research and Applications

# Active neural networks to detect mentions of changes to medication treatment in social media

Davy Weissenbacher <sup>1</sup> Suyu Ge <sup>2</sup> Ari Klein <sup>1</sup> Karen O'Connor <sup>1</sup>  
Robert Gross <sup>1</sup> Sean Hennessy <sup>1</sup> and Graciela Gonzalez-Hernandez <sup>1</sup>

<sup>1</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, and <sup>2</sup>Department of Electronic Engineering, Tsinghua University, Beijing, China

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

Received 9 December 2020; Revised 13 April 2021; Editorial Decision 24 June 2021; Accepted 23 July 2021

### ABSTRACT

**Objective:** We address a first step toward using social media data to supplement current efforts in monitoring population-level medication nonadherence: detecting changes to medication treatment. Medication treatment changes, like changes to dosage or to frequency of intake, that are not overseen by physicians are, by that, non-adherence to medication. Despite the consequences, including worsening health conditions or death, 50% of patients are estimated to not take medications as indicated. Current methods to identify nonadherence have major limitations. Direct observation may be intrusive or expensive, and indirect observation through patient surveys relies heavily on patients' memory and candor. Using social media data in these studies may address these limitations.

**Methods:** We annotated 9830 tweets mentioning medications and trained a convolutional neural network (CNN) to find mentions of medication treatment changes, regardless of whether the change was recommended by a physician. We used active and transfer learning from 12 972 reviews we annotated from WebMD to address the class imbalance of our Twitter corpus. To validate our CNN and explore future directions, we annotated 1956 positive tweets as to whether they reflect nonadherence and categorized the reasons given.

**Results:** Our CNN achieved 0.50 F<sub>1</sub>-score on this new corpus. The manual analysis of positive tweets revealed that nonadherence is evident in a subset with 9 categories of reasons for nonadherence.

**Conclusion:** We showed that social media users publicly discuss medication treatment changes and may explain their reasons including when it constitutes nonadherence. This approach may be useful to supplement current efforts in adherence monitoring.

**Key words:** social media, pharmacovigilance, medication non-adherence, active learning, text classification

## INTRODUCTION

Medication nonadherence refers to when patients do not follow medication treatments as prescribed by their doctors. Nonadherence can be subdivided into 3 categories.<sup>1</sup> In primary nonadherence, patients do not fill their prescriptions or do not start their treatments. In nonpersistence, patients stop their treatments, intentionally or unintention-

ally, without being advised by a health professional to do so. In suboptimal execution, patients are taking their medications but not as recommended (eg, wrong dosage or frequency).

Medication nonadherence has long been recognized as a major contributor to health problems, with the first mention dating from Hippocrates.<sup>2–4</sup> In 2003, the WHO estimated that 50% of patients

in developed countries were failing to follow their medical treatment. In 2018, nonadherence led to an estimated 275 689 deaths at an annual cost of \$528 billion per year in the US alone.<sup>5</sup> Understanding its causes may help us design effective interventions to improve adherence.<sup>6,7</sup> According to Osterberg and Blaschke,<sup>8</sup> the real barriers to adherence lie in deficient interactions between patients, providers, and the healthcare system. For example, by recommending a complex treatment, a provider increases the probability that the patient will skip a medication; by maintaining high cost for a medication, a healthcare system increases the risk for the patient to not refill a prescription; and by showing poor knowledge of medication costs, a provider may prescribe expensive drugs even when more affordable alternatives are available. Furthermore, in clinical practice, patients rarely reveal their nonadherence. And even if they do, they may be reluctant to openly discuss with their healthcare providers the true reasons for altering their therapy against the providers' advice.<sup>9</sup>

Current methods to identify and understand nonadherence have major limitations. Direct observation may be intrusive or expensive, and indirect observation through patient surveys relies heavily on a patient's capacity to remember and report adherence to medication treatment. Our ultimate goal is to study for reasons for nonadherence using social media data at a large scale, as it is generally inexpensive, nonintrusive, and does not rely on a patient's memory of events in the distant past. To the best of our knowledge, the few studies on nonadherence using social media restrict their search to health forums dedicated to long-term or chronic conditions. This choice helps to process and interpret the data but greatly reduces the size of their corpora and, therefore, limits the types of reasons discovered. With 321 million active users per month in 2019,<sup>10</sup> we seek to assess whether Twitter could be a valuable source of data for nonadherence studies at scale.

In preliminary work,<sup>11</sup> we found that given the micro-blogging format of Twitter, many users report changes to their medication treatment in tweets separate from those reporting the reasons for those changes (ie, that the mention of a change in treatment does not always provide evidence of nonadherence). Thus, we approach the detection of nonadherence on Twitter in 2 stages: (1) the tweet-level detection of changes to medication treatment, and (2) the user-level analysis to determine whether the reasons for change could be understood as nonadherence. This study focuses on the first stage, aiming to automatically detect tweets in which users report changes to their treatment, regardless of whether the changes were recommended. Automatically detecting tweet-level reports of changes to medication treatment enables the large-scale use of Twitter data for studying nonadherence at the patient level.

The main contributions of our work are (1) the release of 2 corpora collected from social media, manually annotated with medication change, (2) a binary classifier based on neural networks to detect changes in treatment, and (3) a manual analysis of nonadherence reasons expressed in Twitter for a general set of drugs. Our annotation guidelines, our 2 corpora, and the codes of our classifier were made publicly available during the #SMM4H'21 shared-task, a natural language processing competition (available at: <https://healthlanguageprocessing.org/smm4h-2021/>).

## RELATED WORK

Prior work has focused on attempting to find nonadherence mentions and reasons for nonadherence from different data sources. For this purpose, researchers have mined the unstructured portion of

clinical notes<sup>12,13</sup> or messages from clinical portals.<sup>14</sup> However, these documents are protected, and they do not routinely contain self-reported nonadherence written by patients.

Automatically extracting reasons for nonadherence is harder from social media data. An intuitive approach to tackle this difficulty is to manually analyze a sample of posts, and this has been done with data from health forums, Facebook, and Twitter.<sup>15,16</sup> However, larger studies can benefit from natural language processing methods, to at least automatically filter relevant from irrelevant posts and reduce annotation burden and to analyze large amounts of data. Unsupervised methods are attractive for the latter, since they require few or no annotations to learn the task, relying on topic modeling<sup>17</sup> or interactive exploration of the data with search engines.<sup>18</sup> In general, unsupervised approaches resulted in high recall but low precision.

Despite the challenges of annotation,<sup>19</sup> supervised methods gave the best results. Bigeard et al<sup>18</sup> achieved a 0.824 F1-score on health forums data in French with a Naive Bayes and hand-crafted features approach. Yin et al<sup>20</sup> reported comparable performance on health forums using a binary logistic regression model with word embeddings-based features and achieving a 0.882 F1-score. Both studies used 1000 or fewer annotated examples. In<sup>21</sup> the authors collected drug reviews written by users of the health forum WebMD (available at: <https://www.webmd.com/drugs/2/index>). They applied a binary classifier to detect sentences mentioning nonadherence, then a sequence labeler to extract the reasons in 4500 reviews. The performance of their classifier, a bidirectional Long Short-Term Memory neural network, when trained on 8000 examples and tested on 2000 examples gave a 0.828 F1-score. These performances are, however, aided by the fact that they were working on health forums where users focus their discussions on medical issues. In this context, phrases are unambiguous, and automatic systems can learn reliable linguistic patterns. A further limitation of these studies is that it is unclear whether they validated if the stated change in medication regimen was done with or without a doctor's approval. This is a defining characteristic of nonadherence that is particularly challenging to establish, even for human annotators.

Thus, in this study, our main interest is to detect medication treatment changes in Twitter. This is much less ambiguous and enables the deployment of automatic methods on a much larger scale. The volume of reports that can be collected on Twitter could provide a broader view of nonadherence behaviors (eg, stopping a treatment because of an adverse drug event experienced or feared). Moreover, in Twitter, nonadherent users could potentially be directly contacted and invited to participate in a study. In contrast, following up with the users would not be possible in WebMD since they post reviews anonymously. However, compared to forums, Twitter poses a new challenge for automatic detection. Whereas 55% of reviews mention a change in our WebMD corpus (see Section Corpora), in a separate study<sup>22</sup> we sought to ascertain the topics discussed by users of statins, annotating a corpus of 12 649 tweets that mention a statin, and found that only 1.9% of them (251 tweets) mentioned a change in medication treatment. Such sparsity of positive examples makes the collection of training examples difficult and degrades the performance of learning algorithms.<sup>23</sup>

We addressed the data sparsity issue by training our classifiers with transfer and active learning. In their review,<sup>24</sup> Haixiang et al observed that ensemble learning and resampling are the most popular techniques to train a classifier on class-imbalanced data. These techniques are intuitive and do not need to modify the training algorithm to be applied. In recent years, the combine used of active and

transfer learning has been investigated, in the context of class-imbalanced data, to reduce the size of training examples needed to optimize deep neural networks—the state-of-the-art of machine learning.<sup>25–27</sup> Inspired by the success of these 2 techniques, we explored their performance on our real and very imbalanced data set in this study. Note that resampling, ensemble, transfer, and active learning are not exclusive techniques and can be combined to further improve performance.<sup>28</sup>

## MATERIALS AND METHODS

Our first effort was to collect corpora suitable to train our classifiers with supervision. Knowing that mentions of medication changes in Twitter are rare and that training corpora for supervised methods usually contain several thousand positive examples, we judged the cost too high to create a balanced training corpus. We opted for an alternative solution. We took advantage of 2 training approaches to reduce the annotation effort, transfer and active learning.<sup>26</sup> We explain these approaches in detail in this section.

### Corpora

We detail 2 corpora collected for this study in this section and summarize their statistics in [Table 1](#).

**Twitter Corpus.** We collected 9830 tweets to train and evaluate the classifiers used for this study. We combined tweets mentioning drugs from 2 existing corpora: (1) the corpus released during the first track of the SMM4H'18 shared-task,<sup>29</sup> a natural language processing competition to identify drug mentions, and (2) the corpus collected by Golder et al.<sup>22</sup> To select the tweets of the first corpus, the authors collected 112 500 user timelines between 2014 and 2017. They applied 4 weak classifiers on the timelines to detect tweets mentioning medication names and manually annotated a subset of the tweets retrieved by at least 2 classifiers to increase the likelihood of selecting positive examples, or by only 1 classifier to select nonobvious negative examples. The classifiers were rule-based, lexicon-based, data-centric misspelling-based, and a neural network trained with weak supervision. From the first corpus, after manually excluding tweets not mentioning any medications or only mentioning dietary supplements or herbal remedies, we kept the 7457 tweets mentioning a drug product.

To select the tweets of the second corpus, Golder et al intermittently collected tweets from the Twitter Streaming Application Programming Interface (API) between 2013 and 2018. They collected all tweets mentioning the name of 8 statin medications and their variants. They randomly sampled 12 649 tweets and manually analyzed their contents. From the second corpus, we kept 2373 tweets. We selected these tweets because they were posted by users who were using statins themselves, knew personally the statin user that they reported on, or were healthcare professionals. We excluded all other tweets that Golder et al identified as nonhealth or informational/re-

search-related tweets, since they would have been negative examples for our task. Note that we had to annotate the 2373 tweets with medication changes, since Golder et al only annotated their corpus for nonadherence mentions, which is a subset of the tweets mentioning changes in medication treatments.

Annotation guidelines (<https://healthlanguageprocessing.org/smm4h-2021/task-3/>) were developed to help 2 annotators distinguish tweets that mention a change in the medication treatment and those that do not. One annotator labeled all tweets with “1” if the tweets mention a change in the medication treatment, “0” otherwise. Two annotators double-annotated 4931 of the 9830 tweets, and their interannotator agreement was 0.65 (Cohen's Kappa score). Disagreements were adjudicated by a third annotator. This interannotator agreement score reflects a moderate agreement between the annotators,<sup>30</sup> implying that they had to rely often on their common and medical knowledge to label the tweets. For example, in the tweet “I overdosed on Benadryl today,” our annotators interpreted differently from the language whether the person was taking an excess of the medication or not; in the tweet “I took some of Danny's antihistamine and I feel very high!” our annotators came to a different resolution over the ambiguity of a possible misuse of the medication prescribed to another person versus the use of an over-the-counter medication.

**WebMD Corpus.** Our second corpus consists of reviews written by anonymous users on WebMD. This website provides an opportunity for users to review drugs. A review is assigned to exactly 1 drug and is composed of 3 scores evaluating the satisfaction, effectiveness, and ease of use of the drug. The scores range from 1 to 5 stars, with 1 star being the lowest value. Users can also comment on their personal experience with the drug in a free text form. The comment is optional and limited to 2000 characters. In August 2018, we collected all of the available reviews from the WebMD website using in-house software, totaling 241 094 reviews (989.7 GB). We randomly sampled 12 972 reviews with a comment and 1 or 2 stars for satisfaction. We selected posts with low satisfaction scores since unsatisfied users are more likely to stop or to change medication. Two annotators labeled the reviews with “1” if a review mentions a change of medication, “0” otherwise. The interannotator agreement was also moderate with 0.74 Cohens' Kappa score;<sup>30</sup> disagreements were again adjudicated by a third annotator.

### Classification approaches

#### Baseline

We implemented a simple rule-based classifier as a baseline. It applies a set of regular expressions on a corpus and every post matched by 1 expression was labeled as mentioning a medication change. We designed 2 sets of regular expressions manually, 1 on the training set of the WebMD Corpus (RE\_WebMD), and 1 on the training set of the Twitter Corpus (RE\_Twitter). RE\_WebMD applies 40 regular expressions encoding generic patterns such as “side effect”, “(made|gave) me”, or

**Table 1.** Corpora statistics

	Counts per Corpus	
	WebMD	Twitter
Training Set	10 378 reviews (5746 +/4632 -)	5898 tweets (518 +/5380 -)
Validation Set	1297 reviews (741 +/556 -)	1572 tweets (138 +/1434 -)
Test Set	1297 reviews (735 +/562 -)	2360 tweets (208 +/2152 -)
<b>Total</b>	<b>12 972 reviews</b>	<b>9830 tweets</b>

“(ineffective|had no effect)”. RE\_Twitter applies 116 regular expressions, some borrowed from the RE\_WebMD but modified to include a placeholder indicating that a drug name should occur at a given position, for example “now on `_drug_name_`” or “took an extra `(w+\s){,2}_drug_name_`”. We manually wrote the regular expressions to recognize phrases repeated in the positive examples of each training set. We iteratively corrected these expressions to reject the negative examples they captured in the training set, until any further changes in the expressions degraded their overall performance on the validation sets. We evaluated this classifier on both test sets of the corpora.

### Convolutional Neural Network

We selected Convolutional Neural Networks (CNNs) to detect posts stating a medication change. We ran experiments involving active learning which requires training several networks during multiple iterations.<sup>31</sup> In this setting, CNNs present an advantage over more complex networks as they are fast to train. The common architecture of a CNN is a single convolutional layer of  $m$  filters producing  $m$  vectors. These vectors are combined into a unique vector by the following max-pooling layer. This vector is passed to a fully connected layer which computes the probability of the input to belong to a predefined class—in our case, for the post to mention a change in medication or not.

A trained CNN provides an efficient mechanism to compute the probability for an input post to mention a medication change. To be processed by the CNN, a post  $P$  is split into tokens. Tokens often coincide with words or punctuations, but they can be smaller or bigger elements like “New York,” composed of 2 tokens “New” and “York.” Each token of  $P$  is mapped to a vector of  $d$  real numbers, called a vector embedding. Vector embeddings are usually pretrained on large corpora to place nearby the vectors of tokens semantically or syntactically related in the  $d$ -dimensional space. All embeddings representing the tokens of  $P$  are concatenated into a unique matrix and presented to the CNN. The filters convolve over the matrix of embeddings to score each N-grams (ie,  $n$  consecutive tokens) occurring in the post. These scores express how much the N-Grams are related to a class. The following max-pooling layer acts as a threshold, keeping only the N-Grams most relevant for the classification.<sup>32</sup> The vector produced by the max-pooling layer is a dense representation of the post in a high-dimensional space and used by the last fully connected layer to predict the class of  $P$ . This last layer encodes the decision boundary, a hyperplane in this high-dimensional space that separates the representation of the posts mentioning change from posts that do not.<sup>33</sup> Figure 1 illustrates a decision boundary in a 2-dimension space. The representation of  $P$  is a point in this high-dimensional space and its position relative to the decision boundary indicates its class. The probability of  $P$  to belong to that class is simply the distance of the representation of  $P$  to the decision boundary, the closer  $P$  is from the decision boundary the closer its probability is to 0.5.

We used the common architecture for our CNNs. A CNN accepts a 400x100 matrix as input, where 100 is the number of tokens of a post to classify, and 400 is the dimension of the embeddings representing each token of the post; 100 is a fixed length, right padded for shorter posts, and right truncated for longer ones. We chose existing word embeddings, pretrained on 400 million tweets with word2vec.<sup>34</sup> Godin et al limited the size of the embeddings to 400 because larger embeddings were too complex to train efficiently. We assigned vectors randomly initialized for out-of-vocabulary tokens, fine-tuned during training. We used dropout on

the input and hidden layers of the CNN to avoid overfitting. We used a RELU activation function for all appropriate layers, except for the last layer, where we used a Softmax function. We implemented only 1 convolutional layer in our CNN with 400 filters, a kernel of size 3, and stride 1. We tuned all hyperparameters according to their performances on the validation sets, including the number of filters and the size of the kernel. We preprocessed our posts with a TweetTokenizer from the Natural Language Toolkit.<sup>35</sup> We removed user name handles from the tweets, reduced elongated words, and lowercased all posts. Our code is publicly available (at <https://healthlanguageprocessing.org/smm4h-2021/task-3/>) and provides details of all hyperparameters.

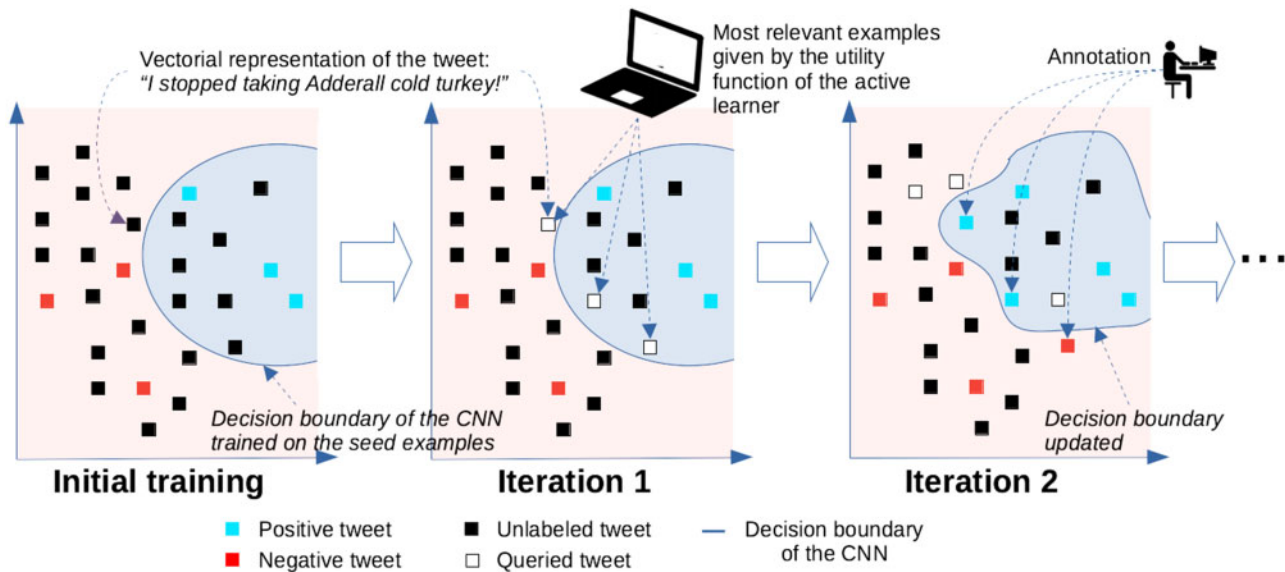
### Convolutional neural network with active and transfer learning

As a way to improve the training of our CNNs despite the strong imbalance between the tweets mentioning and not mentioning medication change in the Twitter Corpus, we used active and transfer learning. Our experiments involved a large number of hyperparameters chosen based on our prior experience. Their optimization through grid searches is left as future work.

Active learning denotes a family of supervised learning algorithms<sup>36</sup> for training a classifier with a limited number of annotated examples. In a standard supervised learning algorithm, all available annotated examples are used to train a classifier. In active learning, an artificial agent, the learner, is introduced to select, from a pool of unlabeled examples, which examples should be annotated for training a classifier—or an ensemble of classifiers. Whereas a passive learner randomly selects the examples, an active learner has an algorithm to select the most relevant examples. Intuitively, the active learner focuses on unlabeled examples likely to be incorrectly classified by its classifier (see Figure 1). Such examples may be located close to the decision boundary of the classifier, revealing the need to update the parameters of the classifier’s model (eg, changing the slope of a linear model) or they may be located in new areas in the features space, revealing the need to change the classifier’s model itself (eg, changing to a nonlinear decision boundary). An active learner learns a task with fewer examples than a passive learner since it queries only for examples resulting in useful changes of the classifier’s model and ignores redundant examples.

We first evaluated the benefits of active learning alone, on the WebMD corpus and on the Twitter corpus. We compared 3 standard learning strategies<sup>36</sup> to select the most relevant examples to annotate from the pool: random sampling as a baseline, uncertainty sampling, and disagreement sampling with a committee of 5 CNNs randomly initialized. Uncertainty sampling is the simplest utility function where the learner selects the examples for which its classifier is the less confident about the labels it assigned—that is, the probabilities of the predictions are close to 0.5. In the disagreement sampling, the learner compares the labels assigned by a committee of classifiers and selects the examples with which they disagree the most. For both utility functions, we chose the entropy and the vote entropy to measure the uncertainty of the classifiers). Our setting was identical for all experiments. We kept 10% of the training set as seed set and placed the remaining examples in a pool, with their labels hidden from the learner. The steps of the active learning algorithm are as follows:

- Our learner started its initial training on the seed set.
- The learner queried the labels for 200 new examples from the pool using its utility function. All examples of our corpus being pre-labeled, we simply released their labels to the learner.



**Figure 1:** Two iterations of the active learning algorithm. All Tweets in the blue zones delimited by the decision boundaries are labeled as positive by a CNN classifier, tweets in the red zone are labeled as negative. The decision boundaries defined by the models of the classifier are updated at each iteration to include newly labeled positive tweets. (Note: we reduced the dimensions of the vectors to two dimensions for representation purpose)

- The learner updated the model of its classifier with the 200 new examples and evaluated its performance on the validation set as well as on the test set and removed these 200 examples from the pool.
- The learner iterated steps b., c., and d. until a heuristic reached its stop condition or no example remained in the pool.

Figure 1 shows the step a. and 2 iterations of steps b., c., and d. The learner saved its classifier's models at each iteration c. and selected the model  $m$  at the iteration where it achieved its best performance on the validation set. We evaluated the learner on the test set using the model  $m$ . The model  $m$  may not be the model which obtained the best performance on the test set, but choosing  $m$  guarantees good performances for real applications. A limitation of this approach, however, is that to choose  $m$ , we need to know the iteration where  $m$  achieves its best performance on the validation set which requires all examples of the corpus to be annotated. Therefore, we defined a heuristic relying on patience to approximate this iteration and decide automatically when to stop the annotation process. For each classifier, we trained in parallel 5 models as described in Section Convolutional neural network with active and transfer learning. At each step d., if the mean of the F1-scores on the validation set was the highest recorded, the patience  $k$  was reset to 0 and a new iteration of steps b., c., and d. executed, querying 200 new examples to be annotated. If the mean of the F1-scores was not the highest recorded, the patience  $k$  was incremented by 1 and compared with the maximum patience threshold  $p$ . If  $k$  was inferior to  $p$ , a new iteration of steps b., c., and d. was executed; otherwise, the training stopped and the models which achieved the highest F1-scores recorded in the previous iterations were evaluated on the test sets. The patience  $p$  is a hyperparameter to define. We computed our heuristic on the mean F1-scores of 5 classifiers to reduce the fluctuations in the neural networks' performance.

**Transfer learning** is a heuristic to improve the performance of a supervised classifier.<sup>26</sup> With transfer learning, a classifier solves a new task by reusing in its inference the knowledge it acquired when solving a similar task. In the case of a neural network classifier, the knowledge is instantiated by the weights of the network.

We evaluated the benefits of transfer learning by implementing a passive learner using a single CNN. We started by training our classifier on all training examples of the WebMD corpus—the *source corpus*—to learn the linguistic patterns indicating a change in the medication treatment. We transferred its knowledge by continuing the training of the classifier with all training examples from the Twitter corpus—the *target corpus*. We evaluated our classifier on the test set of the Twitter corpus. To confirm that any increase of performance was due to the transfer and not simply due to the additional examples from the source corpus, we trained the classifier on all training examples of the source and target examples combined and randomly presented to the CNN, and evaluated the CNN on the test set of the Twitter corpus. We also measured the loss when our classifier did not have past knowledge—that is, when the CNN was only trained and evaluated on the target corpus. Considering the similarity between the source and the target corpora, we checked the performance of a CNN first trained on the training source corpus and, with no additional training step on the training set of the target corpus, evaluated on the test set of the target corpus.

**Transfer and Active Learning** were used in combination and evaluated with the same settings as for evaluating transfer learning, with 1 difference: we used an active learner with all examples annotated during the training phases of the classifier on the source and/or target corpora.

All classifiers and learners evaluated during our experiments are summarized in Table 2.

### Evaluation

We trained and evaluated our classifiers/learners 5 times and reported the means of their precision, recall, and F1 scores to reduce the differences caused by their stochastic optimizations. True Positives (TP) are posts that mention a change in treatment and are detected by a classifier. False Negatives (FN) are posts that mention a change in treatment but are not detected by a classifier. False Positives (FP) are posts that do not mention a change in treatment but are detected by a classifier. The Precision is the fraction of posts cor-

**Table 2.** Classifiers and learners summary

System	Description	Experiment
RegEx	Classifier using hand-crafted regular expressions	<i>RE_WebMD</i> —Expressions crafted on WebMD training set; evaluated on WebMD test set <i>RE_Twitter</i> —Expressions crafted on Twitter training set; evaluated on Twitter test set
Random	Passive learner using a single CNN and random sampling	<i>RA_WebMD</i> —Trained on WebMD training set; evaluated on WebMD test set <i>RA_Twitter</i> —Trained on Twitter training set; evaluated on Twitter test set <i>RA_Transfer</i> —Trained on WebMD training set, with/without update on Twitter training set; evaluated on Twitter test set <i>RA_W + T</i> —Trained on WebMD and Twitter training sets combined; evaluated on Twitter test set
Uncertainty	Active learner using a single CNN and uncertainty sampling	<i>UN_WebMD</i> —Trained on WebMD training set; evaluated on WebMD test set <i>UN_Twitter</i> —Trained on Twitter training set; evaluated on Twitter test set <i>UN_Transfer</i> —Trained on WebMD training set, with/without update on Twitter training set; evaluated on Twitter test set <i>UN_W + T</i> —Trained on WebMD and Twitter training sets combined; evaluated on Twitter test set
Committee	Active learner using a committee of 5 CNNs and disagreement sampling	<i>CO_WebMD</i> —Trained on WebMD training set; evaluated on WebMD test set <i>CO_Twitter</i> —Trained on Twitter training set; evaluated on Twitter test set <i>CO_Transfer</i> —Trained on WebMD training set, with/without update on Twitter training set; evaluated on Twitter test set <i>CO_W + T</i> —Trained on WebMD and Twitter training sets combined; evaluated on Twitter test set

rectly classified as positive among all posts classified as positive: TP/(TP+FP). The Recall is the fraction of posts mentioning a change successfully retrieved: TP/(TP+FN). The F1 score is the harmonic mean of the precision and recall; it summarizes the overall performance of the classifier.

### Validation

Using our classifier, we collected a large corpus of tweets mentioning a change of medication treatment and manually analyzed the tweets to determine whether users were nonadherent to their prescriptions and the reason why, if given. This was a preliminary investigation to evaluate the speed and accuracy of our classifier on a large set of tweets and to find if tweets that indicate a change of medication could lead to a finding of nonadherent behaviors and their reasons. The actual study of nonadherence at the population level using Twitter is left as future work since it will require a systematic collection of tweets for a chosen class of medications and the development of new models to detect nonadherence.

We collected 1 936 820 tweets from January 2019 to April 2020. We queried the stream of Twitter using the official application programming interface to retrieve tweets mentioning drug names, or their variants, from a predefined list of 1322 drugs. The 1322 drugs were randomly selected from the RxNorm database (<https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html>, accessed January 28, 2021). We applied our best classifier (see Section Benefits of transfer learning) on these 1.9 million tweets and

detected 5811 tweets with a probability to mention a change of medication equal to or higher than 0.95. From these 5811, we manually analyzed a subset of 3010 tweets, randomly selected. One annotator first confirmed the decision of our classifier—the tweet mentioned a medication treatment change—and then looked for the reasons of this change in the timeline of the user, up to 10 tweets posted before and after the tweet mentioning the change as previously done in.<sup>11</sup> If the tweet was a part of a discussion, the annotator also looked for the reasons into the discussion thread. We determined nonadherence if it was stated, or could be inferred, from the tweet that the user changed or stopped taking the medication without consulting their provider. For example, in the tweet, “took Prozac for a while, took myself off it didn’t like the side effects from it,” the user is clearly stating that they made the decision (“took myself off it”) rather than their doctor, thus it was labeled as nonadherence. Our senior annotator (KO) categorized the reasons for nonadherence and our experts in pharmacoepidemiology (SH and RG) validated a subset of them.

## RESULTS

### Automatic detection of medication change

We detail our results in Table 3. The drop of performance for all classifiers when applied on our Twitter corpus stands out. Whereas the learner with a committee of CNNs—the best learner on both corpora—achieved a high score on the WebMD corpus (82.8% F1-

score), its performance dropped from 82.8% to 50.4% F1-score on the Twitter corpus. This drop is likely due to the difference of genre between our 2 corpora.<sup>37</sup> When analyzing our data, we noticed that when users are reviewing drugs in WebMD, they only discuss their experiences and rarely diverge. WebMD reviews are also longer than tweets, providing more context. As shown by the high recall of RE\_WebMD (90.3%), we can express few generic and reliable patterns on this corpus to detect medication change (eg, the phrase “side effect” most often indicates a change. When users are posting on Twitter, they discuss other subjects than their medication experience. Such generic patterns become unreliable. To improve their precision, we needed to integrate more constraints to model the surrounding context in tweets, and consequently, to keep a high recall, multiplying close variation of the patterns. Such adaptation of the patterns remains challenging for both humans and classifiers.<sup>38</sup>

### Reducing annotation effort with active learning

In our evaluation in Section Automatic detection of medication change, we reported the performances of our classifiers with the best models on the validation sets. However, the best models on the validation sets were known because all examples in our corpora were annotated. In this section, we measure the reduction of the annotation effort possible with active learning. We report in Figure 2 the average performances on the test sets of our 2 corpora of our 3 learners when trained on various subsets of the training examples. We marked with black circles the performance of our classifiers when trained with active learning and stopped using our heuristic. The active learners reached their best performances plateau earlier than the baseline passive learner on both corpora, and active learning did not cause a significant drop in performance compared to learning on all examples available. On the WebMD corpus, with a patience  $k = 10$ , the UN\_WebMD was stopped after being trained on 37% of the training examples and achieved performance close to optimal on the test set with 82% F1-score. The CO\_WebMD was

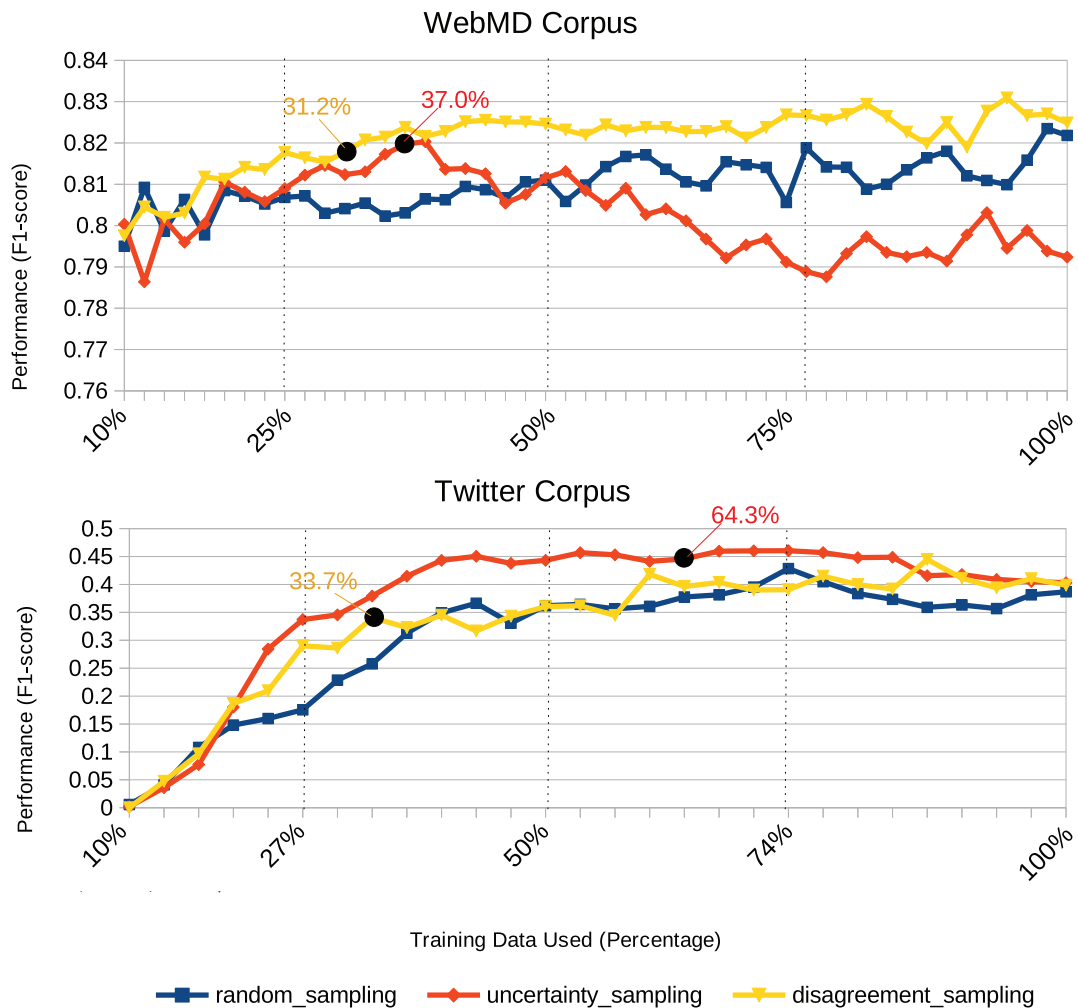
stopped too early, 7 iterations before reaching a good plateau of performance, and achieved 81.8% F1-score trained with only 31.2% of the training data available. Both scores are close to 82.2% F1-score, the score achieved by RA\_WebMD when trained on all training examples. On the Twitter corpus, we observed a similar pattern with a patience  $k = 3$ . UN\_Twitter was stopped late, after being trained with 64.3% of the training examples, but on the plateau of optimal performance. CO\_Twitter was again stopped too early, trained with 33.7% of the training examples. UN\_Twitter achieved 44.6% F1-score on the test set, a score higher than the score achieved by RA\_Twitter when trained on all examples with 38.7% F1-score.

### Benefits of transfer learning

We achieved our best score on the Twitter corpus (50.4% F1-score) with CO\_Transfer, an active learner using a committee of CNNs and transfer learning. Because users express their change of medication with the same linguistic patterns in social media, our learner could learn the patterns from the WebMD corpus, a larger and more balanced corpus, and performed more efficiently the task on the Twitter corpus. The transfer allowed an increase of 10 F1-score points for the committee CO\_Transfer (50.4% F1-score) compared to the baseline learner RA\_Twitter (40.4% F1-score). A requirement, however, is to adjust the reliability of each pattern on the target corpus through additional training iterations on a small sample of examples from the target corpus. Without adapting its classifiers' models, there was no evidence our learner gained anything from the transfer as shown by the experience CO\_Transfer in Table 3. Note that the additional examples provided during the transfer learning are an important factor of improvement, as shown by the difference between RA\_Twitter (40.4% F1-score) and RA\_W+T (46.7% F1-score), but it is not the only factor. The order in which the training examples are presented to the classifier during the transfer with active learning is also a key factor for improvement and counts for 3.7 F1-score points (50.4–46.7).

**Table 3.** Binary classification of medication change in WebMD and Twitter posts. (Precision, recall and F1 scores are given in percentage; classifiers trained by an active learner are marked with italic fonts)

		Corpus: WebMD		
System		Precision	Recall	F1
	RE_WebMD—Regex	68.3	90.3	77.8
	RA_WebMD—Random	78.7	85.9	82.1
	<i>UN_WebMD—Uncertainty</i>	<b>81.4</b>	83.1	82.2
	<i>CO_WebMD—Committee</i>	80.2	85.7	<b>82.8</b>
		Corpus: Twitter		
System		Precision	Recall	F1
No Transfer	RE_Twitter—Regex	45.3	41.4	43.2
	RA_Twitter—Random	50.8	34.1	40.4
	RA_W+T—Random	46.8	47.1	46.7
	<i>UN_Twitter—Uncertainty</i>	39.3	<b>56.2</b>	45.5
	<i>UN_W+T—Uncertainty</i>	44.2	51.2	46.9
	<i>CO_Twitter—Committee</i>	51.7	34.7	41.3
	<i>CO_W+T—Committee</i>	52.2	37.8	43.7
Transfer with and without update on target corpus	RA_Transfer—Random	53.6/38.7	40.9/45.5	46.2/41.3
	<i>UN_Transfer—Uncertainty</i>	46.5/26.4	52.4/51.6	48.4/34.9
	<i>CO_Transfer—Committee</i>	<b>56.2/30.7</b>	45.8/52.5	<b>50.4/38.7</b>



**Figure 2:** Classifiers' performance on WebMD and Twitter corpora test sets with increasing training data. (Black circles indicate the percentage of training examples analyzed by the classifiers when achieving their best performance on validation sets given our stopping heuristic. We set patience  $k = 10$  on the WebMD corpus and patience  $k = 3$  on the Twitter corpus.)

### Analysis of medication nonadherence reported on Twitter

Our manual analysis of the subset of 3010 tweets, detected by our best classifier as very likely to mention a medication change, confirmed our preliminary results published in<sup>11</sup>: users post about their medication nonadherence on Twitter and were likely to explain their reasons in the tweets assessed (including previous and subsequent tweets). From these 3010 tweets, 1956 were True Positive (ie, tweets mentioning a medication change). Among these 1956 tweets, 19.2% (375/1956) were explicitly mentioning nonadherence with the reasons explained in the tweets themselves (68%, 255/375) or their contexts (9.1%, 34/375). In this example, “*i was taking my adderall less than prescribed to save money, [...]*”, the nonadherence reason was categorized as Access Issue due to cost. Out of 375, only 22.9% (86/375) did not have a reason indicated in their context.

Table 4 summarizes the classes of stated reasons for nonadherence. Adverse drug reaction was the most common reason mentioned for being nonadherent, with 19.2% (72/375) of our tweets. Misuse and abuse were the second reason with 17.8% (67/375). This reason could be over-represented in our sample because 57.9% (217/375) of the tweets mentioning nonadherence are mentioning mixed amphetamine salts Adderall, a drug abused for its stimulant

properties. Access issues to the drugs caused, for example, by their costs or problems with insurances and refill issues, were also a major concern to patients and counted for the third reason of nonadherence with 12.8% (48/375), before unintentional nonadherence of users, only 6.4% (24/375).

## DISCUSSION

In this study, our objective was to automatically detect tweets mentioning changes in medication treatment and manually analyze their context to assess whether the reasons of the changes were given and whether we could determine if they were advised by a physician or not.

Transfer and active learning appear to be efficient heuristics to help train classifiers on extremely imbalanced corpora. Compared to a traditional supervised learning approach, we were able to increase the performance of our learner by 9.1 points in F1-score by combining both heuristics. These positive results were obtained on 1 task and need to be confirmed. We intend to repeat our experiments on the adverse drug event detection and drug detection tasks described in<sup>23,39</sup>.



**Table 4.** Reasons of nonadherence discovered in a sample of tweets mentioning drugs. Statements are unedited for spelling, punctuation, or format

Nonadherence Reason	Description/Example	PERCENT (COUNT)
Adverse Drug Reaction	Experienced/fear of adverse drug reaction <i>"I hate avapro after a few doses I got a sore throat that lead to nonstop coughing. I stopped taking it two days ago. [...]"</i>	19.2 (72)
Misuse/Abuse	Indication that the medication was being abused/misused <i>"[...] when I abused adderall now and then for a while [...]"</i>	17.8 (67)
Access issues	Unable to get medication (cost, insurance, refill issues, etc.) <i>"I stopped taking my Lamotrigine and took myself off quetiapine bc I no longer wanted to pay for them[...]"</i>	12.8 (48)
Beliefs	Various beliefs (not needed, being overmedicated, harmful, etc.) <i>"Yes big pharma is the reason I stopped taking my Xanax and other meds"</i>	11.2 (42)
Unintentional	Nonadherence seems unintentional (forgotten, error dosage, etc.) <i>"I didn't take my Adderall for a while because I lost my bottle and today was my first day back on it a [...]"</i>	6.4 (24)
Parent's decision	Parent took child off medication <i>"my dad's taking me off of my vyvanse prescription without my doctors advise and self-undiagnosing me even though he only took me off meds bc of my side effects from letting me have weekends off"</i>	2.4 (9)
Efficacy	Not effective or higher dosage needed <i>"When the pain med stopped working, I stopped going They gave me no other treatments. Was wasting my time!"</i>	2.3 (8)
Stigma	Patient felt a stigma being on medication <i>"The doctor treated me like a drug addict, [...] I discontinued the med because of the shame"</i>	1.3 (5)
Other	Did not fit into any of above categories <i>"[...] He has been neglecting caring for himself since finding out he couldn't get on the transplant list. He stopped taking his diuretic. [...]"</i>	3.7 (14)
Not available	Reason not found in tweet context <i>"I stopped taking ramipril on my own terms weeks ago."</i>	22.9 (86)
Total		100 (375)

We made 2 experimental choices to facilitate the validation of transfer and active learning for training a classifier on an imbalance corpus. We made these choices to greatly reduce the computation time during our experiments, knowing that we may have limited our performance. We chose a well-established neural network architecture for our classifiers and well-known utility functions for our active learners, whereas better alternatives already exist. Such an alternative could be a dense representation of the entire posts using ELMo or BERT neural networks.<sup>40</sup> With this representation, it also becomes possible to express new utility functions. When neural networks encode the posts, they encode their semantics and place similar posts close to each other in a multidimensional space and posts with different meanings far away. Such representation could help an active learner to explore unlabeled data by clustering posts expressing the same meaning despite lexical and syntactical variations.<sup>31,41</sup> We leave as future work the use of dense representation of the posts and the design of new utility functions exploiting this representation to improve the performance of our classifier.

In this study, we hypothesized that the linguistic patterns expressing medication change in social media are similar across different drug classes. We focused our efforts on developing a general learner by training it on corpora composed of posts mentioning any medication. To our knowledge, no prior work exists that targets medication change mentions in social media. However, in general, epidemiological studies in pharmacovigilance include only a particular class of drugs (rather than a random collection of them) to discover unknown reasons for nonadherence. The transfer learning

approach presented in this study can be used on a new corpus of tweets mentioning a specific drug class of interest by simply retraining our classifier on this new corpus as it was done in this study for the general collection of drugs. Future work in this direction will benchmark the performance of the classifier per medication class (for example, for statins or antihypertensives).

With the manual annotation of our corpora, we found that within tweets mentioning a drug name, 1.84% are explicit mentions of nonadherence (Given that in our Twitter corpus, 9.6% of the tweets mention a change in medication treatment, we can assume that 185,934 tweets in the 1.9 million tweets are mentioning a change in medication. Since we found that, among tweets mentioning a change in medication treatment, 19.2% are explicitly nonadherence, we conclude that 35,699 (185 934\*0.192) tweets are explicit mention of nonadherence, that is 1.84% (35 699/1 936 820) of the tweets in our initial 1.9 million tweets). This percentage, 1.84%, must be interpreted with regard to the size of the data generated on Twitter. The corpus of 1.9 million used for the validation of our approach was a sample of a larger database of 25 million tweets collected from August 2017 to April 2021. This database should therefore contain around 466 060 (25 329 350\*0.0184) explicit mentions of nonadherence. Yet, even 25 million is a relatively small number of tweets, since we collected tweets based on only 1322 drug names and their variants, and used the free standard Twitter streaming API, which returns a low percentage of the total number of tweets mentioning those drug names posted in real time. According to Morstatter et al,<sup>42</sup> the API returns only from 1% up to 43.5%

of the total tweets posted, depending on the traffic on Twitter. We believe the systematic use of Twitter data to be a promising complement to evidence-based efforts to understand medication nonadherence. Collecting tweets is cost-effective, does not represent a burden on the participants, is available in real time, and is abundant. Tweets are unmediated (not subject to researchers' biases), and they may provide additional contexts to understand the patients through their discussions, timelines, or even direct contact.

## CONCLUSION

In this study, we presented an ensemble of CNNs to detect tweets mentioning changes in medication treatment. Using transfer and active learning, we achieved 0.50 F1-score, a score high enough to collect a large number of tweets of interest and manually analyze their context to determine if users were nonadherent to their prescriptions. We conclude that Twitter users do state their nonadherence to medication treatments and are likely to explain their reasons in their timelines, suggesting that Twitter data, systematically collected and automatically analyzed, could supplement current efforts in identifying patient-stated reasons for nonadherence. A major challenge remains to fully automate the detection of nonadherence and their reasons for larger studies.

## 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 the National Library of Medicine.

## AUTHOR CONTRIBUTIONS

DW designed the experiments, preprocessed the data, implemented the active learning library, helped with the implementation of the regular expressions and the CNNs, computed the models, analyzed the results, and wrote the majority of the manuscript. SG implemented the CNNs and wrote its description. AK implemented the regular expressions. KO wrote the annotation guidelines, annotated the data, and computed the interannotator agreement. RG and SH helped define the guidelines and validate a subset of the annotations. GG-H secured funding and guided the overall study design. All authors proofread and edited the manuscript.

## DATA AVAILABILITY STATEMENT

The data underlying this article will be shared on reasonable request to the corresponding author.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

- De Geest S, Zullig LL, Dunbar-Jacob J, *et al*. ESPACOMP Medication Adherence Reporting Guideline (EMERGE). *Ann Intern Med* 2018; 169 (1): 30–5.
- Reddy M. Non-compliance in pharmacotherapy. *Indian J Psychol Med* 2012; 34 (2): 107–9.
- Vrijens B, De Geest S, Hughes DA, ABC Project Team, *et al*. A new taxonomy for describing and defining adherence to medications. *Br J Clin Pharmacol* 2012; 73 (5): 691–705.
- Hippocrates. Hippocrates, Prognostic. Regimen in Acute Diseases. The Sacred Disease. The Art. Breaths. Law. Decorum. Physician (Ch. 1). Dentition, vol. II, L. C. L. 148, Ed., Harvard University Press, 1923: DECORUM, xiii–xvi.
- Watanabe JH, McInnis T, Hirsch JD. Cost of prescription drug-related morbidity and mortality. *Ann Pharmacother* 2018; 52 (9): 829–37.
- Hugtenburg JG, Timmers L, Elders PJ, Vervloet M, van Dijk L. Definitions, variants, and causes of nonadherence with medication: a challenge for tailored interventions. *Patient Prefer Adherence* 2013; 7: 675–82.
- Marcum ZA, Sevick MA, Handler SM. Medication nonadherence: a diagnosable and treatable medical condition. *JAMA* 2013; 309 (20): 2105–6.
- Osterberg L, Blaschke T. Adherence to medication. *N Engl J Med* 2005; 353 (5): 487–97.
- Yin Lam W, Fresco P. Medication adherence measures: an overview. *Biomed Res Int* 2015; 2015: 1.
- Shaban H. Twitter reveals its daily active user numbers for the first time; 2019. *Washington Post*. <https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/> Accessed April 29, 2020.
- Onishi T, Weissenbacher D, Klein A, O'Connor K, Gonzalez-Hernandez G. Dealing with medication nonadherence expressions in Twitter. In: *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*; Association for Computational Linguistics; Brussels, Belgium; 2018: 32–3.
- Sohn S, Murphy SP, Masanz JJ, Koehler J-PA, Savova GK. Classification of medication status change in clinical narratives. *AMIA Annu Symp Proc* 2010; 2010: 762–6.
- Topaz M, Radhakrishnan K, Lei V, Zhou L. Mining clinicians' electronic documentation to identify heart failure patients with ineffective self-management: a pilot text-mining study. *Stud Health Technol Inform* 2016; 225: 856–7.
- Yin Z, Harrell M, Warner JL, Chen Q, Fabbri D, Malin BA. The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation. *J Am Med Inform Assoc* 2018; 25 (11): 1444–51.
- Mao JJ, Chung A, Benton A, *et al*. Online discussion of drug side effects and discontinuation among breast cancer survivors. *Pharmacoepidemiol Drug Saf* 2013; 22 (3): 256–62.
- Bhattacharya M, Snyder S, Malin M, *et al*. Using social media data in routine pharmacovigilance: a pilot study to identify safety signals and patient perspectives. *Pharm Med* 2017; 31 (3): 167–74.
- Abdellaoui R, Foulquié P, Texier N, Faviez C, Burgun A, Schück S. Detection of cases of noncompliance to drug treatment in patient forum posts: topic model approach. *J Med Internet Res* 2018; 20 (3): e85.
- Bigéard É, Thiessard F, Grabar N. Detecting drug non-compliance in internet fora using information retrieval and machine learning approaches. *Stud Health Technol Inform* 2019; 264: 30–4.
- Belz A, Ford E, Hoile R, Mullick A. Conceptualisation and annotation of drug nonadherence information for knowledge extraction from patient-generated texts. In: *Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text*; Association for Computational Linguistics; Hong Kong, China; 2019: 202–11.
- Yin, L. Song Z, Malin B. Reciprocity and its association with treatment adherence in an online breast cancer forum. In: *EEE 30th International Symposium on Computer-Based Medical Systems (CBMS); IEEE Computer Society; Thessaloniki, Greece*; 2017: 618–23.
- Xie J, Liu X, Zeng DD, Fang X. Understanding medication nonadherence from social media: a sentiment-enriched deep learning approach. *SSRN Electron J* 2017.
- Golder S, O'Connor K, Hennessy S, Gross R, Gonzalez-Hernandez G. Assessment of beliefs and attitudes about statins posted on Twitter. *JAMA Netw Open* 2020; 3 (6): e208953.
- Weissenbacher D, Sarker A, Klein A, O'Connor K, Magge A, Gonzalez-Hernandez G. Deep neural networks ensemble for detecting medication mentions in tweets. *J Am Med Inform Assoc* 2019; 26 (12): 1618–26.
- Haixiang G, Yijing L, Shang J, Mingyun G, Yuan Yue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst Appl* 2017; 73: 220–39.

25. Ein-Dor L, Halfon A, Gera A, *et al.* Active learning for BERT: an empirical study. In: *Proceedings of the 2020 Conference on EMNLP*; Association for Computational Linguistics; Online; 2020: 7949–62.
26. Kasai J, Qian K, Gurajada S, Li Y, Popa L. Low-resource deep entity resolution with transfer and active learning. In: *proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics; Florence, Italy; July 28–August 2; 2019: 5851–61.
27. Aggarwal U, Popescu A, Hudelot C. Active learning for imbalanced datasets. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*; Snowmass, USA; March 1–5; 2020: 1417–26.
28. Fernández AH, García SL, Galar M, Prati RC, Krawczyk B, Herrera F. *Learning from Imbalanced Data Sets*. Springer International Publishing, 2018.
29. 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*; Association for Computational Linguistics; Brussels, Belgium; October 31; 2018: 13–6.
30. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012; 22 (3): 276–82.
31. Lu J, Henchion M, Namee BM. Investigating the effectiveness of representations based on word-embeddings in active learning for labelling text datasets. *arXiv* 2019.
32. Jacovi A, Sar Shalom O, Goldberg Y. Analyzing and interpreting neural networks for NLP. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*; Association for Computational Linguistics; Brussels, Belgium; November 1; 2018: 56–65.
33. Samarasinghe S. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*. Auerbach Publications; 2006.
34. Godin F, Vandersmissen B, De Neve W, Van de Walle R. Multimedia Lab@ACL WNUT NER shared task: named entity recognition for twitter microposts using distributed word representations. In: *Proceedings of the Workshop on Noisy User-Generated Text*; Association for Computational Linguistics; Beijing, China; July 31; 2015: 146–53.
35. Bird S, Loper E, Klein E. *Natural Language Processing with Python*. Sebastopol: O'Reilly Media Inc; 2009.
36. Settles B. *Active Learning, Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publisher, 2012.
37. Poibeau T. *Extraction Automatique D'information: Du Texte Brut au Web Sémantique*. Hermes Science Publications; 2003.
38. Vanni L, Ducoffe M, Aguilar C, Precioso F, Mayaffre D. Textual Deconvolution Saliency (TDS): a deep tool box for linguistic analysis. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics; Melbourne, Australia; July 15–20; 2018: 548–57.
39. Weissenbacher D, Sarker A, Magge A, *et al.* Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019. In: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*; Association for Computational Linguistics; Florence, Italy; August 2; 2019: 21–30.
40. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*; Association for Computational Linguistics; Florence, Italy; August 1; 2019: 58–65.
41. Kholghi M, De Vine L, Sitbon L, Zuccon G. The benefits of word embeddings features for active learning in clinical information extraction. In: *Proceedings of Australasian Language Technology Association Workshop*; Melbourne, Australia; December 5–6; 2016: 25–34.
42. Morstatter F, Pfeffer J, Liu H, Carley KM. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*; AAAI press; Cambridge, USA; July 8–11; 2013: 400–8.