Journal of the American Medical Informatics Association, 28(10), 2021, 2147–2154
doi: 10.1093/jamia/ocab124
Advance Access Publication Date: 1 August 2021
Research and Applications

Research and Applications

Application of natural language processing techniques to identify off-label drug usage from various online health communities

Brian Dreyfus[1], **Anuj Chaudhary**[2], **Parth Bhardwaj**[2], **and V. Karthikhaa Shree**[2]

[1]Epidemiology, Bristol Myers Squibb, Princeton, New Jersey, USA and [2]Mu Sigma, Bengaluru, India

Corresponding Author: Brian Dreyfus, MPH, Bristol Myers Squibb, Route 206 & Province Line Road, Princeton, NJ, USA; Brian.Dreyfus@BMS.com

Received 4 February 2021; Revised 17 May 2021; Editorial Decision 18 May 2021; Accepted 4 June 2021

ABSTRACT

Objective: Outcomes mentioned on online health communities (OHCs) by patients can serve as a source of evidence for off-label drug usage evaluation, but identifying these outcomes manually is tedious work. We have built a natural language processing model to identify off-label usage of drugs mentioned in these patient posts.
Materials and Methods: Single patient posts from 4 major OHCs were considered for this study. A text classification model was built to classify the posts as either relevant or not relevant based on patient experience. The relevant posts were passed through a spelling correction tool, CSpell, and then medications and indications from these posts were identified using cTAKES (clinical Text Analysis and Knowledge Extraction System), a named entity recognition tool. Drug and indication pairs were identified using a dependency parser. Finally, if the paired indication was not mentioned on the label of the drug approved by U.S. Food and Drug Administration, it was tagged as off-label use of that drug.
Results: Using this algorithm, we identified 289 off-label indications, achieving a recall of 76%.
Conclusions: The method designed in this study identifies and extracts the semantic relationship between drugs and indications from demotic posts in OHCs. The results demonstrate the feasibility of using natural language processing techniques in identifying off-label drug usage across online health forums for a variety of drugs. Understanding patients' off-label use of drugs may be able to help manufacturers innovate to better address patients' needs and assist doctors' prescribing decisions.

Key words: Online Health Communities, off-label, natural language processing techniques, openFDA application programming interface

INTRODUCTION

Off-label drug usage is the consumption of a drug for any indication not mentioned on its label, as approved by the U.S. Food and Drug Administration (FDA).[1,2] Research from a survey of office-based physicians found that 21% of prescriptions were for off-label use.[3] The availability of benefit-risk profiles of a medicine used off-label is limited or could be entirely unknown. Although physicians prescribing medications for off-label use may provide an opportunity for innovation, there are scant data on the safety profile of drugs used for unapproved indications.

Health authorities such as the FDA require the marketing authorization holder (MAH) to report on the off-label use of their medicines.[4] Traditionally, the monitoring of off-label use by an MAH is derived from the spontaneous reports that are received and stored in the corporate safety database. Research has shown that spontaneous reporting of adverse events represents no more than 5% to 10% of the real incidence of events, and off-label use reports have been observed with an even lower frequency.[5] The underreporting of off-label use adverse events, likely because of the unawareness about the possibility of reporting, as off-label use is permitted in the United States, and the fear of reporting medical errors limits the usefulness of these data sources in ascertaining the safety profile of a medicine for off-label indications.

In a chart review of patients treated with gabapentin, nearly 95% of the patients received the drug off-label, and the drug had minimal benefit.[6] A broad analysis of free text narratives in electronic medical records (EMRs) using natural language processing (NLP) had relatively good success at detecting off-label use.[7]

Social media websites, including online health communities (OHCs), Twitter, Facebook, and others, are potentially the largest source of data related to off-label use of medications. In addition to the sheer breadth of data available from social media, these data represent most directly the patient's voice. Existing research on OHC patient posts serve as source of evidence for off-label drug usage mentions.[8] User posts on Reddit related to opioid addiction were analyzed and alternative unapproved treatments were found to be promoted by the patients on other online communities.[9] For pharmaceutical companies that are the MAH of a medication, these posts present an opportunity to comprehensively evaluate the drug's benefit-risk profile. Because the posts in OHCs are in the form of health discussions, they contain requisite information such as the drug name and indication but clinical details regarding the case such as the medical history of the patient may not be available. The NLP method in this study is based on a patient taking a specific drug for a specific indication (s). Because the posts contain conversations and discussions, the main goal was to identify and extract the requisite information that contain the relevant patient experience. Pharmacovigilance departments currently do not routinely listen to the patient's voice via OHCs. The process of manually perusing through patient posts on online communities for off-label drug usage would be time-consuming. Also, to study these trends on a periodic basis is a tedious task.[10] Hence, to reduce the manual effort involved, we built a model to scan patient posts to online health forums, capture relevant posts with patient experience, and identify off-label drug usage. Though online health forums provide us with huge volumes of real-world data, not all of it can be consumed, owing to the noise in the data. A major part of the methodology is dedicated to identifying posts that capture patients experience associated with a specific drug name and an indication. This analysis represents the early stages of research being conducted on data sources beyond the traditional spontaneous reports to comprehensively ascertain patients' experiences with off-label use of medicines.

MATERIALS AND METHODS

A diagram of the overall methods, including data collection, relevance classification, spell-check, named entity recognition (NER), drug-disease association, and off-label identification, is presented in Figure 1.

Data

OHCs are forums where patients discuss, in the form of comments or posts, diseases and their impact, medical treatments, and drug composition. Social media forums, like blogs or social networks, including OHCs, generate huge amounts of data with a wide range of information driven by the contribution of the users sharing their experience, opinions, and acquired knowledge. Extensive exploration of OHC data represents the source of evidence for identifying off-label drug usage mentions for this study.

Taking into consideration the privacy policies and the complexities of the structure of the posts and comments on these OHCs, the following sources were selected: MedHelp, WebMD, Drugs.com, and HealthBoards.com. The analysis was conducted on a large and diverse drug list across therapeutic areas under 2 categories (positive and negative cohorts), with 6 drugs under each category. Positive cohort drugs were expected to have off-label drug usage, whereas the negative cohort drugs were not expected to have major off-label drug usage. After a review of the literature, the following drugs were selected as the positive cohort, as they were known to have off-label uses: pregabalin,[11] gabapentin,[11,12] lorazepam,[13] methylphenidate,[14] amitriptyline,[15] and quetiapine.[16] The negative cohort drugs considered were as follows: levonorgestrel, levofloxacin, clopidogrel, celecoxib, buprenorphine-naloxone, and natalizumab. These drugs were selected due to belonging to a therapeutic area or indication where off-label usage is difficult and unexpected based on literature review.

A total of 76 912 posts for the target drug list were scraped using Python code from the previously mentioned OHCs to form the base dataset for the analysis. The entire process can be divided into 4 major steps: (1) extract the posts from OHCs, (2) build a relevance classification model to target the posts with key information, (3) identify the medical entities like drug name and indication from the relevant posts, and (4) build a rule-based model to identify off-label drug usage using NLP techniques.

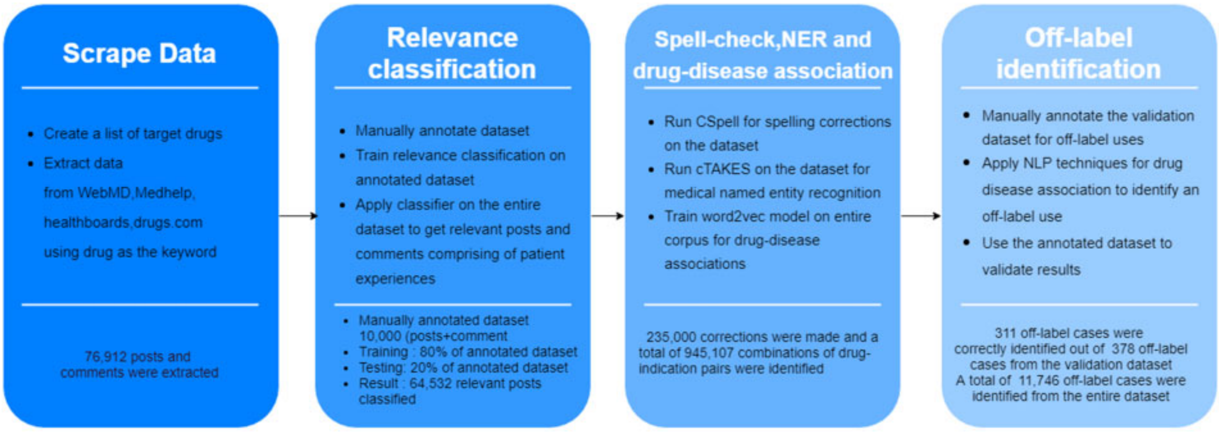
Scrape Data
Create a list of target drugs
Extract data from WebMD,Medhelp, healthboards,drugs.com using drug as the keyword
76,912 posts and comments were extracted
Relevance classification
Manually annotate dataset
Train relevance classification on annotated dataset
Apply classifier on the entire dataset to get relevant posts and comments comprising of patient experiences
Manually annotated dataset 10,000 (posts+comment
Training : 80% of annotated dataset
Testing: 20% of annotated dataset
Result : 64,532 relevant posts classified
Spell-check,NER and drug-disease association
Run CSpell for spelling corrections on the dataset
Run cTAKES on the dataset for medical named entity recognition
Train word2vec model on entire corpus for drug-disease associations
235,000 corrections were made and a total of 945,107 combinations of drug-indication pairs were identified
Off-label identification
Manually annotate the validation dataset for off-label uses
Apply NLP techniques for drug disease association to identify an off-label use
Use the annotated dataset to validate results
311 off-label cases were correctly identified out of 378 off-label cases from the validation dataset
A total of 11,746 off-label cases were identified from the entire dataset

Figure 1. Project process flow. cTAKES: clinical Text Analysis and Knowledge Extraction System; NER: named entity recognition; NLP: natural language processing.

Table 1. Classic text classification model results

Model	Vectorization Technique	Accuracy	Precision	Recall	F_1 Score
SVM	**Count Vectorizer (without Stopwords)[a]**	75.46%	81.88%	76.43%	79.06%
	Count Vectorizer[a]	77.28%	83.41%[d]	78.03%	80.63%
	TF-IDF (without Stopwords)[a]	76.52%	79.88%	81.88%	80.87%
	TF-IDF[a]	79.16%	81.26%	85.29%	83.23%
Naive Bayes	**Count Vectorizer (without Stopwords)[c]**	66.40%	69.30%	78.20%	73.48%
	Count Vectorizer[c]	69.90%	73.70%	77.10%	75.36%
	TF-IDF (without Stopwords)[c]	66.30%	65.20%	93.70%[d]	76.89%
	TF-IDF[c]	70.10%	68.10%	94.00%[d]	78.98%
RNN	**Count Vectorizer (without Stopwords)[c]**	76.00%	75.00%	75.00%	75.00%
	Count Vectorizer[c]	79.50%[d]	79.00%	78.00%	78.50%
	TF-IDF (without Stopwords)[c]	75.80%	68.00%	63.00%	65.40%
	LSTM+CNN[c]	81.00%[d]	79.00%	78.00%	78.50%
SGDC	**Count Vectorizer (without Stopwords)[a]**	75.16%	82.84%	74.42%	78.40%
	Count Vectorizer[a]	72.73%	87.15%[d]	64.51%	74.14%
	TF-IDF[a]	79.41%	81.48%	85.44%	83.41%[d]
	Count Vectorizer + Doc2Vec	68.73%	83.28%[d]	60.56%	70.13%
	TF-IDF + Doc2Vec	71.85%	84.92%[d]	65.12%	73.71%
Random Forest	**Count Vectorizer (without Stopwords)[a]**	76.98%	78.43%	85.54%	81.83%
	Count Vectorizer[a]	78.53%	78.41%	89.09%[d]	83.41%[d]
	TF-IDF (without Stopwords)[a]	77.65%	78.21%	87.49%	82.59%
	TF-IDF[a]	78.40%	78.73%	88.19%[d]	83.19%
	Count Vectorizer + TF-IDF[b]	78.28%	78.34%	88.69%[d]	83.19%
XGBoost	**Count Vectorizer (without Stopwords)[a]**	78.53%	81.87%	82.93%	82.40%
	Count Vectorizer[a]	80.35%[d]	83.09%[d]	84.83%	83.95%[d]
	TF-IDF (without Stopwords)[a]	78.31%	81.49%	83.08%	82.28%
	TF-IDF[a]	79.92%[d]	81.69%	86.19%	83.88%[d]
	Count Vectorizer + TF-IDF[b]	80.16%[d]	82.12%	85.99%	84.01%[d]

CNN: convolutional neural network; LSTM: long short-term memory; RNN: recurrent neural network; SGDC: stochastic gradient descent classifier; SVM: support vector machine; TF-IDF: term frequency–inverse document frequency.

[a] With 900 features.

[b] Total 900 features with 450 features for each vectorization technique.

[c] With 300 features.

[d] Top 5 results for each metric.

Relevance classification model

The posts or comments on OHCs are entered firsthand by the public.[17] These posts include knowledge about drugs, medical conditions (along with their personal experience, including general statements with no direct patient experience), and questions about the drug that do not have any information regarding consumption of the drug by patients. Comments with no information regarding the consumption of a drug did not contribute to our analysis. This analysis is focused on comments that include patient experience with a drug of interest and that mention an indication for use.[18] Comments from patients themselves, members of their family, or friends are referred to as a patient's personal experience. A patient's personal experience with drug consumption or any prescription of the drug by a physician is referred to as a relevant comment in this study.

To classify a comment as relevant or not relevant, a text classification model was built based on the previous definition. The conventional method of building a text classification model[19,20] was followed, in which the text is first converted to vectors using standard vectorization techniques (Count Vectorizer and term frequency–inverse document frequency [TF-IDF]) and also with combinations of vectorization techniques: Count Vectorizer and TF-IDF, Count Vectorizer and Doc2-Vec,[21] TF-IDF and Doc2Vec, and long short-term memory networks and convolutional neural networks. Second, these vectorization techniques were incorporated into the following models: support vector machines,[22–24] naive Bayes, recurrent neural network, stochastic gradient descent classifier, random forest,[22–24] and XGBoost.[24]

To train these models, 10 000 comments from the base dataset were randomly selected and manually annotated for relevance of the comment (mention of patient experience). The annotated dataset consisted of around 7000 relevant comments; the remaining 3000 comments were determined to be irrelevant. These 10 000 comments were further split into separate sets to train (80%) and test (20%) the text classification model.

The 6 vectorization techniques and their associated performance metrics are presented (by model) in Table 1. To identify the best model, we focused on F_1 scores. The highest F_1 score (84%; accuracy of 80%) was observed in the XGBoost model, with a vectorization ensemble of TF-IDF and Count Vectorizer. This model was further validated using k-fold cross-validation with k value as 10, which maintained consistency across folds. With the best-suited XGBoost model parameters, accuracy, precision, and recall were 83.67%, 85.76%, and 89.37%, respectively. This is the text classification model used to classify the base dataset into "relevant" and "not relevant" comments. The entire corpus of 76 912 posts was processed by this XGBoost model, yielding 64 000 posts tagged as relevant comments. This is the dataset that was examined for off-label drug usage.

NER using clinical text analysis and the knowledge extraction system

Spelling correction is one of the fundamental segments required for most of the NLP applications. The patient posts were written by lay-

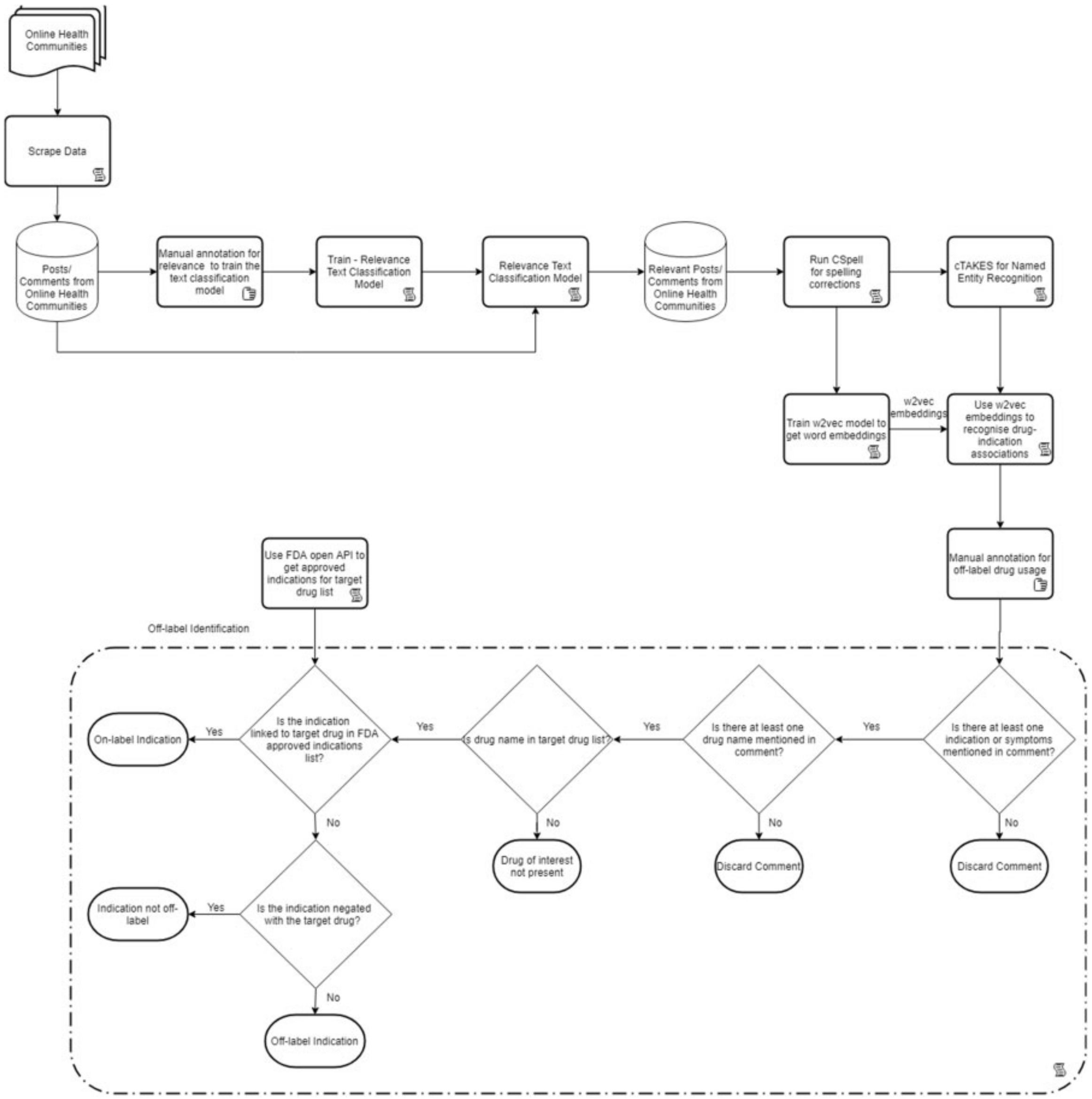
Online Health Communities
Scrape Data
Posts/ Comments from Online Health Communities
Manual annotation for relevance to train the text classification model
Train - Relevance Text Classification Model
Relevance Text Classification Model
Relevant Posts/ Comments from Online Health Communities
Run CSpell for spelling corrections
cTAKES for Named Entity Recognition
Train w2vec model to get word embeddings
w2vec embeddings
Use w2vec embeddings to recognise drug-indication associations
Manual annotation for off-label drug usage
Use FDA open API to get approved indications for target drug list
Off-label Identification
Is there at least one indication or symptoms mentioned in comment?
Yes
No
Discard Comment
Is there at least one drug name mentioned in comment?
Yes
No
Discard Comment
Is drug name in target drug list?
Yes
No
Drug of interest not present
Is the indication linked to target drug in FDA approved indications list?
Yes
On-label Indication
No
Is the indication negated with the target drug?
Yes
Indication not off-label
No
Off-label Indication

Figure 2. Detailed process flow for off-label identification. API: application programming interface; cTAKES: clinical Text Analysis and Knowledge Extraction System; FDA: U.S. Food and Drug Administration.

persons, so the health-related discussions in these posts sometimes contained irregular spellings along with orthographic errors. To standardize the language of the content and to prepare it for linguistic analysis, spelling streamlining was performed by incorporating the spell checker for consumer language (CSpell).[25] Using CSpell, more than 235 000 words, including misspelled medical entities, were identified in the relevant patient posts dataset. A total of 1730 relevant posts with the drug of interest were identified that would have been missed because of spelling errors in the posts.

Once the language of the content was streamlined, the NER functionality of the clinical Text Analysis and Knowledge Extraction System (cTAKES) was employed to identify the medical entities mentioned in the patient posts. Recognizing the key medical entities from the relevant patient posts classified by the text classification model is a crucial process for the off-label drug usage identification. cTAKES is an open-source NLP system used for information extraction from text.[26] By connecting to the Unified Medical Language System (UMLS) Metathesaurus database, the NER of cTAKES identifies medical entities grouped under procedures, disease disorders, signs and symptoms, and anatomy and drugs. The UMLS Metathesaurus is used as a complete knowledge source in the medical and healthcare field.[27] By utilizing the vast concepts and relationships contained by the UMLS, extraction of useful knowledge from patient posts in OHCs is possible.

In this process of identifying medical entities, a few indications identified were observed to be nonmedical entities. To filter out such entities, Word2Vec[28,29] embedding was implemented by calculating the cosine similarity score between the drug of interest and the indication mentioned in the comment. When the calculated similarity score returned a negative value, the identified entity was confirmed as nonmedical and removed from the final list.

Off-label identification

A diagram of the process for determining if a comment indicates a medication was used off-label is presented in Figure 2. The FDA-approved indications for the target drug list were extracted from the label of the

Table 2. Confusion matrix, including ambiguous comments: Dataset A

		Actual Value	
		Positive (1)	Negative (0)
Predicted Value	**Positive (1)**	TP = 301	FP = 462
	Negative (0)	FN = 77	TN = 1038

Metric results were as follows: accuracy = 71%; precision = 40%; recall = 80%; F_1 score = 52%. TN was used as indicated and correctly classified by the model; for an FN, the post was about off-label usage but was incorrectly classified by the model; for a TP, the post was about off-label usage and was correctly classified by the model; for an FP, the post was not about off-label usage and was incorrectly classified by the model.

FN: false negative; FP: false positive; TN: true negative; TP: true positive.

Table 3. Confusion matrix, excluding ambiguous comments: Dataset B

		Actual Value	
		Positive (1)	Negative (0)
Predicted Value	**Positive (1)**	TP = 301	FP = 149
	Negative (0)	FN = 77	TN = 581

Metric results were as follows: accuracy = 80%; precision = 67%; recall = 80%; F_1 score = 73%. For a TN, the post was not about off-label usage and was correctly classified by the model; for an FN, the post was about off-label usage but was incorrectly classified by the model; for a TP, the post was about off-label usage and was correctly classified by the model; for an FP, the post was not about off-label usage and was incorrectly classified by the model.

FN: false negative; FP: false positive; TN: true negative; TP: true positive.

drugs on the FDA's official website,[4] incorporating the openFDA application programming interface[30] to create a mapping file.

A sample of 5000 random comments was taken from the 64 000 relevant comments and were manually annotated as "off-label" or "not off-label." Comments that could not be classified as either "off-label" or "not off-label" (770 comments) were labeled as "ambiguous comments." Comments that did not contain the necessary drug and indication information required for off-label prediction (3122 comments) were removed from the manually annotated off-label dataset. The remaining 1878 comments, which included ambiguous comments, was labeled dataset A; a subset of dataset A, which excluded the 770 ambiguous comments, was labeled dataset B.

In order to understand the interdependency of words in a comment and procure the combinations of drug-indication pairs, ScispaCy[31] was used as the dependency parser. The result was sent to the NetworkX[32,33] tool, which was used to generate dependency trees. As dependency parsing works on a sentence level, the process of finding links between the drug and indication also works on a sentence level across all the sentences in the comment. In an event in which the drug and indication are not mentioned in the same sentence, the preceding and succeeding sentences are checked for an indication mention. If presence of an indication is detected, the sentence is checked for the presence of a word that is being used as a substitution for the drug (it, drug, medication). On detection of a substitute word, the word is replaced with the drug name and dependency is then checked. The resultant dependency between a pair of words from the dependence tree may be erroneous when the words are too far apart in a comment. Thus, a threshold of 9 jumps between the respective pair of words, using parameter tuning, was used. If the distance between the 2 words was within this threshold, the pair was retained for further analysis.

The drug-indication pairs at the comment level were compared with the approved indication list for the target drug created using openFDA application programming interface. If the pair was not in the approved indication list, the model flagged it as off-label for that drug, otherwise was flagged as not off-label. Further, to check and remove mentions of any negations between the drug and indication, NegSpacy,[34] a negation handling pipeline, was incorporated. Drug-indication pairs with negation involved were flagged as not off-label for that drug by the model. The result was in the form of an off-label flag (off-label or not off-label for the target drug) and confidence flag (confidence bucket of the off-label flag).

RESULTS

A total of 76 912 posts and comments were extracted from the 4 OHCs using 12 target drugs as keywords. Of that total, 64 000 comments were classified as relevant by the XGBoost relevance classifier. After these comments were passed through our NLP-based model, a total of 11 927 posts or comments were identified as off-label by the off-label identifier. For validation, our model was tested on a gold standard dataset of 5000 manually annotated comments, resulting in recall, accuracy, and precision scores of 80%, 71%, and 40%, respectively (Table 2). The model was also tested with a subset (dataset B) of the gold standard dataset to see the model performance without any ambiguous comments (Table 3). The off-label indications observed for our target drug list are shown in Table 4. Those indications identified most frequently (at least twice) for each drug in the positive cohort are as follows: lorazepam (sleep/insomnia, pain, and panic attack), amitriptyline (migraine, pain, and anxiety), gabapentin (pain, neuropathy, and migraine), methylphenidate (depression, anxiety, and sleep/insomnia), quetiapine (sleep/insomnia, anxiety, and mood control), and pregabalin (pain).

DISCUSSION

In this study, we built a text classification model to apply to patient posts from 4 OHCs (MedHelp, WebMD, Drugs.com, and HealthBoards.com) to determine the extent of off-label use of 12 medications (6 for which off-label use was likely, 6 for which off-label use was unlikely). NLP of social media is increasingly being used to address public health research questions.[35] We have designed a unique NLP algorithm that utilizes a combination of techniques to analyze OHCs for information about off-label use of medications.

Using our algorithm, relevant posts (relaying patient experience with a drug of interest and mentioning an indication) were screened for drug-indication pairs, which were then compared with the FDA label to determine if there was off-label use. The algorithm was validated using 5000 manually annotated comments (annotated and reviewed by authors), achieving an accuracy of 71%. Of 64 000 comments classified as relevant, a total of 11 927 (18.6%) posts or comments were identified as off-label. This is in good agreement with Radley et al,[3] who found that 21% of prescriptions written by office-based physicians were off-label. A somewhat higher percentage (28.1%) of pediatric patient visits resulted in an off-label use of a medication.[36]

Though our model showcases the results for identifying off-label drug usage, it accounts only for the patient-experienced sign and

Table 4. Model results

Drug Name	U.S. Food and Drug Administration–Approved Indication	Off-Label Indications Identified[a]
Positive cohort		
Lorazepam	Anxiety, depression	Sleep disorder/insomnia (59), pain (43), panic attacks (28), addiction (13), common cold (12), headaches (11), stroke (5), tachycardia (4), chills (3), fibromyalgia (2), hyperventilating (2), epilepsy (1),
Amitriptyline	Depression	Pain (31), sleep disorder/insomnia (30), migraine (15), headaches (8), anxiety (4), stroke (3), cystitis (2), neuropathy (2)
Gabapentin	Epilepsy, postherpetic neuralgia, seizures	Pain (52), sleep disorder/insomnia (17), depression (13), neuropathy (8), numbness (7), spasms (6), migraine (5), anxiety (4), fibromyalgia (4), swelling (3), addiction (3), heartburn (2), restless legs syndrome (2), panic attacks (1), menopause (1), irritable bowel syndrome (1), carcinoma (1), gastritis (1), angina (1)
Methylphenidate	Attention-deficit/hyperactivity disorder	Depression (25), anxiety (17), fatigue (9), insomnia (6), pain (6), stress disorder (3), cancer (3), seizures (2), autism (2), posttraumatic hyperactivity (1), nausea (1), tumors (1), appetite (1), apnea (1), diabetes (1), stroke (1), spasms (1)
Quetiapine	Bipolar disorder, bipolar I and bipolar II disorder, bipolar I disorder, bipolar I disorder manic, depression, manic, mental disorders, schizophrenia	Anxiety (49), Sleep disorder/insomnia (26), mood control (21), common cold (12), pain (12), panic attacks (8), anger management (7), addiction (7), paranoia (6), nauseous (5), agitation (5), fatigue (4), hallucinations (4), dyskinesia (3), migraine (2), stroke (2), dizziness (2), palpitations (2), maniac disorder (1), allergy (1), apnea (1), psychosis (1), headaches (1), chronic fatigue syndrome (1), panicky (1), hypothyroidism (1), hives (1)
Pregabalin	Spinal cord injury, diabetic peripheral neuropathy, fibromyalgia, neuropathic pain, partial onset seizures, postherpetic neuralgia	Pain (165), headaches (10), shingles (9), migraine (7), anxiety (7), influenza (6), neuralgia (5), depression (5), arthritis (4), cancer (3), paresthesia (2), spondylitis (1)
Negative cohort		
Celecoxib	Acute pain, dysmenorrhea, juvenile rheumatoid arthritis, osteoarthritis, pain, primary dysmenorrhea, rheumatoid arthritis	Tendonitis (4), cancer (2), swelling (2), spasms (2), headaches (1), hematuria (1), fibromyalgia (1), carcinoma (1), sleep disorder (1)
Buprenorphine	Overdose, pain, drug abuse	Anxiety (4), depression (3), fibromyalgia (1), tumors (1)
Natalizumab	Multiple sclerosis	Infection (2), asthma (1), neuropathy (1)
Clopidogrel	Antiplatelet agent	None identified
Levonorgestrel	Contraception	None identified
Levofloxacin	Broad spectrum antibiotic	None identified

[a]Off-label indications are listed in decreasing order of frequency.

symptom, regardless of it being a sign or symptom of the indication. In some cases, the entities identified as an off-label indication may be signs or symptoms of the approved indication for the drug.

Furthermore, the potential for off-label usage identification may be correlated to the classification of the drugs into positive and negative cohorts. The rationale behind using this method is to help with the validation of the algorithm. The model was expected to identify more off-label usage for drugs in the positive cohort and little to none in the negative cohort. This hypothesized outcome was successfully validated by the results. Reports from patients posted online provide a source of evidence for the use of medications for which there are no published efficacy data, but identification of these off-label uses by manual search is prohibitively time-consuming. Therefore, physicians are left to rely on their own clinical experience.[4] The automated search of outcomes related to off-label use has the potential to generate enough evidence to warrant the initiation of traditional clinical trials. Indeed, off-label use does present a possible safety concern. A recent study of off-label prescribing for children in 2 hospitals by Lee et al[37] found that the children who died had a higher number of off-label prescriptions than did the total cohort. The health authorities require marketing authorization holders to report off-label use of their medicines in periodic reports. This research marks a preliminary step in understanding the social media data from OHCs as a source of data source for off-label use. Further research will need to be conducted. The mining of OHCs could help provide a more complete picture of off-label use of medications.

There are other ways to estimate off-label use of medications. Andrulyte and Bjerrum[38] identified off-label prescriptions through data mining of pharmacy servers. However, this method lacks an easy way to identify off-label use by indication. Surveys of hospital records offer another source of information about off-label prescribing,[39] though this method is limited to inpatient prescribing and can only be performed for a limited number of hospitals. Some hospitals have developed procedures such that doctors are required to complete a request form for off-label use of a medication.[40] Off-label prescription of medications can also be estimated from published studies.[41]

Wallach and Ross[42] suggest that the FDA be given authority to require pharmaceutical companies to conduct clinical trials investigating the safety and efficacy of off-label uses of their medications once those uses reach a certain sales or prescribing threshold. This additional research would help ensure that the safety and efficacy off-label uses are investigated in a coordinated manner. The mining of OHCs could provide one line of evidence regarding the degree of off-label use.

CONCLUSION

The primary goal of this study was to identify off-label drug usage from patient experience on OHCs using NLP techniques. In this communication, we have focused on the methodology and not on the metric values of the model.

Various factors, including (1) consideration of relevant patient posts with patient experience mentions, (2) streamlining the incorrect spellings in the content of the patient posts, and (3) extraction of drug names and indications mentioned in the patient posts were used for our primary analysis, and aided in identifying off-label drug usage. Because the approach we followed is a combination of machine learning and a rule-based model, along with the incorporation of various NLP techniques, we believe that our method is robust and can be confidently used for off-label drug usage identification from OHCs.

Information regarding off-label drug usage could possibly be used by drug manufacturers to guide expansion of research on their products and by authorities tasked with monitoring off-label usage. Understanding off-label use of medications could promote investment in the clinical trials needed to provide guidance regarding the safety and efficacy in new indications. However, owing to the novel nature of these data, it is premature to consider development toward new indications without additional research and evaluation with traditional methods. The scope of this work can be expanded in the future to patient age, dosage of the drug, and mode of administration of the drug, and to identify off-label usage of a wider range of drugs. Use of a consumer health vocabulary to map commonly used consumer terms with diseases or indications may lead to identification of more drug-disease associations. Extending the use of the current algorithm to real-time data sources could act as an early identification tool for off-label drug usage.

FUNDING

The work reported in this article was funded by Bristol Myers Squibb.

AUTHOR CONTRIBUTIONS

All authors contributed sufficiently and meaningfully to the conception and design of the project, and to drafting, editing, and revising the manuscript. All authors approved the final version for submission and agree to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

Writing and editorial support were provided by Rodney J. Moore, Ph.D. (Bristol Myers Squibb) under the direction of the authors.

DATA AVAILABILITY STATEMENT

No new data were generated in this research. The data underlying this article are available in MedHelp (https://www.medhelp.org/), WebMD (https://www.webmd.com/), Drugs.com (https://www.drugs.com/), and HealthBoards.com (https://www.healthboards.com/). The natural language processing programs used will be available on reasonable request to the corresponding author.

CONFLICT OF INTEREST STATEMENT

BD was an employee of Bristol Myers Squibb at the time of this research. AC, PB, and VKS were employees of Mu Sigma at the time of this research.

REFERENCES

1. Jansen RM, Off-label use of medication In, Beran R, ed. *Legal and Forensic Medicine*. Berlin, Germany: Springer; 2013:1601–15.
2. Wittich CM, Burkle CM, Lanier WL. Ten common questions (and their answers) about off-label drug use. *Mayo Clin Proc* 2012; 87 (10): 982–90.
3. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. *Arch Intern Med* 2006; 166 (9): 1021–6.
4. U.S. Food and Drug Administration. E2C(R2) periodic benefit-risk evaluation report (PBRER). July 2016. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e2cr2-periodic-benefit-risk-evaluation-report-pbrer. Accessed November 10, 2020.
5. Mascolo A, Scavone C, Bertini M, *et al.* Safety of anticancer agents used in children: a focus on their off-label use through data from the spontaneous reporting system. *Front Pharmacol* 2020; 11: 621.
6. Hamer AM, Haxby DG, McFarland BH, *et al.* Gabapentin use in a managed medicaid population. *J Manag Care Pharm* 2002; 8 (4): 266–71.
7. Jung K, LePendu P, Chen WS, *et al.* Automated detection of off-label drug use. *PLoS One* 2014; 9 (2): e89324.
8. Frost J, Okun S, Vaughan T, *et al.* Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. *J Med Internet Res* 2011; 13 (1): e6.
9. Chancellor S, Nitzburg G, Hu A, *et al.* Discovering alternative treatments for opioid use recovery using social media. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; 2019.
10. Billal B, Fonseca A, Sadat F, *et al.* Semi-supervised learning and social media text analysis towards multi-labeling categorization. In: *Proceedings of the IEEE International Conference on Big Data*; 2017: 1907–16.
11. Goodman CW, Brett AS. A clinical overview of off-label use of gabapentinoid drugs. *JAMA Intern Med* 2019; 179 (5): 695–701.
12. Peckham AM, Evoy KE, Ochs L, *et al.* Gabapentin for off-label use: evidence-based or cause for concern? *Subst Abuse* 2018; 12: 1178221818801311.
13. Ghiasi N, Bhansali RK, Marwaha R. Lorazepam. StatPearls. 2020. Updated March 31, 2020. https://www.ncbi.nlm.nih.gov/books/NBK532890/. Accessed November 16, 2020.
14. Van Zyl PM, Joubert G, Fechter L, *et al.* Methylphenidate use among students living in junior on-campus residences of the University of the Free State. *South African Family Pract* 2017; 59 (4): 123–7.
15. Thour A, Marwaha R. Amitriptyline. Treasure Island, FL: StatPearls Publishing; 2020. https://www.ncbi.nlm.nih.gov/books/NBK537225/. Accessed November 16, 2020.
16. Arts M, Petrykiv S, Fennema J, *et al.* Off-label prescriptions of quetiapine for sleep disturbances. *Eur Psychiatr* 2017; 41 (S1): S749–50.
17. van der Eijk M, Faber MJ, Aarts JWN, *et al.* Using online health communities to deliver patient-centered care to people with chronic conditions. *J Med Internet Res* 2013; 15 (6): e115.
18. Azadeh N, Ransohoff JD, Callahan A, *et al.* Profiling off-label prescriptions in cancer treatment using social health networks. *JAMIA Open* 2019; 2 (3): 301–5.
19. Habdank M, Rodehutskors N, Koch R. Relevancy assessment of tweets using supervised learning techniques: mining emergency related tweets for automated relevancy classification. In: *Proceedings of the International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*; 2017.
20. Mirończuk MM, Protasiewicz J. A recent overview of the state-of-the-art elements of text classification. *Expert Syst Appl* 2018; 106: 36–54.

21. Lau JH, Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*; 2016: 78–86.
22. Ru B, Warner-Hillard C, Ge Y, *et al.* Identifying serendipitous drug usages in patient forum data - a feasibility study. In: *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies*; 2017: 106–8.
23. Ru B, Li D, Yao L. Detecting serendipitous drug usage in social media with deep neural network models. In: *Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine*; 2018: 1083–90.
24. Wang S, Li Z, Wang Y, *et al.* Machine learning methods to predict social media disaster rumor refuters. *Int J Environ Res Public Health* 2019; 16 (8): 1452.
25. Lu CJ, Aronson AR, Shooshan SE, *et al.* Spell checker for consumer language (CSpell). *J Am Med Inform Assoc* 2019; 26 (3): 211–8.
26. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
27. Chen D, Zhang R, Liu K, *et al.* Knowledge discovery from posts in online health communities using unified medical language system. *Int J Environ Res Public Health* 2018; 15 (6): 1291.
28. Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. In: *NeurIPS Proceedings: Advances in Neural Information Processing Systems 26 (NIPS 2013)*; 2013.
29. Segura-Bedmar I, Suárez-Paniagua V, Martinez P. Exploring word embedding for drug name recognition. In: *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*; 2015: 64–72.
30. Kass-Hout TA, Xu Z, Mohebbi M, *et al.* OpenFDA: an innovative platform providing access to a wealth of FDA's publicly available data. *J Am Med Inform Assoc* 2016; 23 (3): 596–600.
31. Neumann M, King D, Beltagy I, *et al.* ScispaCy: fast and robust models for biomedical natural language processing. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*; 2019: 319–27.
32. Hagberg A, Swart PJ, Schult DA. Exploring Network Structure, Dynamics, and Function Using NetworkX. Report no. LA-UR-08-5495. Los Alamos, NM: Los Alamos National Laboratory; 2008.
33. Zhao K, Wang X, Cha S, *et al.* A multirelational social network analysis of an online health community for smoking cessation. *J Med Internet Res* 2016; 18 (8): e233.
34. Babanejad N, Agrawal A, An A, *et al.* A comprehensive analysis of preprocessing for word representation learning in affective tasks. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; 2020: 5799–810.
35. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumer-generated data. *Yearb Med Inform* 2019; 28 (1): 208–17.
36. Yackey K, Stukus K, Cohen D, *et al.* Off-label medication prescribing patterns in pediatrics: an update. *Hosp Pediatr* 2019; 9 (3): 186–93.
37. Lee J-H, Byon H-J, Choi S, *et al.* Safety and efficacy of off-label and unlicensed medicines in children. *J Korean Med Sci* 2018; 33 (37): e227.
38. Andrulyte M, Bjerrum OJ. Identifying off-label prescriptions through data mining in Danish community pharmacy servers: an exploratory study on desmopressin, diclofenac, fucidin, mirtazapine and quetiapine. *Basic Clin Pharmacol Toxicol* 2018; 123 (2): 155–60.
39. Landwehr C, Richardson J, Bint L, *et al.* Cross-sectional survey of off-label and unlicensed prescribing for inpatients at a paediatric teaching hospital in Western Australia. *PLoS One* 2019; 14 (1): e0210237.
40. Arocas Casañ V, Mateo Carmona J, García Molina O, *et al.* Off-label prescription of drugs at hospital. *Farm Hosp* 2016; 40: 63–71.
41. Allen HC, Garbe MC, Lees J, *et al.* Off-label medication use in children, more common than we think: a systematic review of the literature. *J Okla State Med Assoc* 2018; 111: 776–83.
42. Wallach JD, Ross JS. Gabapentin approvals, off-label use, and lessons for postmarketing evaluation efforts. *JAMA* 2018; 319 (8): 776–8.