Journal of the American Medical Informatics Association, 0(0), 2019, 1–9
doi: 10.1093/jamia/ocz048
Research and Applications

AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

Research and Applications

Learning to detect and understand drug discontinuation events from clinical narratives

Feifan Liu,[1] Richeek Pradhan,[1] Emily Druhl,[2] Elaine Freund,[1] Weisong Liu,[3] Brian C. Sauer,[4] Fran Cunningham,[5] Adam J. Gordon,[6,7] Celena B. Peters,[4,6] and Hong Yu[2,3,8,9]

[1]Department of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts, USA, [2]Bedford VA Medical Center, Bedford, Massachusetts, USA, [3]Department of Computer Science, University of Massachusetts Lowell, Lowell, Massachusetts, USA, [4]Departments of Internal Medicine and Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA, [5]Department of Veterans Affairs Pharmacy Benefits Management Services, Hines, Illinois, USA, [6]Informatics, Decision-Enhancement, and Analytic Sciences Center (IDEAS 2.0), VA Salt Lake City Health Care System, Salt Lake City, Utah, USA, [7]Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA, [8]Department of Medicine, University of Massachusetts Medical School, Worcester, Massachusetts, USA and [9]Department of Computer Science, University of Massachusetts Amherst, Amherst, Massachusetts, USA

Corresponding Author: Hong Yu, FACMI, PhD, Department of Computer Science, University of Massachusetts Lowell, 220 Pawtucket St, Lowell, MA 01854, USA (hong_yu@uml.edu)

Received 3 October 2018; Revised 19 March 2019; Editorial Decision 25 March 2019; Accepted 26 March 2019

ABSTRACT

Objective: Identifying drug discontinuation (DDC) events and understanding their reasons are important for medication management and drug safety surveillance. Structured data resources are often incomplete and lack reason information. In this article, we assessed the ability of natural language processing (NLP) systems to unlock DDC information from clinical narratives automatically.

Materials and Methods: We collected 1867 de-identified providers' notes from the University of Massachusetts Medical School hospital electronic health record system. Then 2 human experts chart reviewed those clinical notes to annotate DDC events and their reasons. Using the annotated data, we developed and evaluated NLP systems to automatically identify drug discontinuations and reasons at the sentence level using a novel semantic enrichment-based vector representation (SEVR) method for enhanced feature representation.

Results: Our SEVR-based NLP system achieved the best performance of 0.785 (AUC-ROC) for detecting discontinuation events and 0.745 (AUC-ROC) for identifying reasons when testing this highly imbalanced data, outperforming 2 state-of-the-art non–SEVR-based models. Compared with a rule-based baseline system for discontinuation detection, our system improved the sensitivity significantly (57.75% vs 18.31%, absolute value) while retaining a high specificity of 99.25%, leading to a significant improvement in AUC-ROC by 32.83% (absolute value).

Conclusion: Experiments have shown that a high-performance NLP system can be developed to automatically identify DDCs and their reasons from providers' notes. The SEVR model effectively improved the system performance showing better generalization and robustness on unseen test data. Our work is an important step toward identifying reasons for drug discontinuation that will inform drug safety surveillance and pharmacovigilance.

Key words: natural language processing, drug surveillance, knowledge representation, supervised machine learning, electronic health records

Published by Oxford University Press on behalf of the American Medical Informatics Association 2019. This work is written by US Government employees and is in the public domain in the US.

INTRODUCTION

Drug discontinuation (DDC), the cessation of a drug treatment by either the clinician or the patient,[1] is a frequent event in patient care. Comprehensive information on DDC is critical to medication management (eg, avoiding medication errors that could adversely affect patient care)[2] and drug safety surveillance (eg, ensuring high-fidelity drug exposure measurement in population research).[3] Drugs can be discontinued in a myriad of circumstances, such as 1) naturally (eg, antibiotics are discontinued when pneumonia resolves); 2) by patients themselves (eg, when patients inappropriately stop taking the medications against the physician's instructions; or 3) withdrawn from the market by pharmaceutical companies and/or regulatory organizations. Existing studies typically group DDC reasons into 4 categories: 1) adverse events (AEs), 2) lack of treatment effectiveness, 3) patient preference/nonadherence, and 4) medication being no longer necessary.[4,5] The various categories contribute differently. Studies on causes of anti-tumor necrosis factor discontinuation, for instance, showed that AEs were responsible for treatment discontinuations in 48.7% of cases, and lack of treatment effectiveness represented the most significant single cause for anti-tumor necrosis factor treatment discontinuation (50%).[5] Another study of patients with human immunodeficiency viruses showed that the leading causes of discontinuation were intolerance/toxicity (58.5%) and poor adherence (24%).[6] Therefore, effectively extracting DDC events as well as understanding their reasons hold great potential for efficient pharmacovigilance, effective patient management, as well as improved patient safety and health care outcomes.

Existing pharmacovigilance methods often depend on structured pharmacy claims data or primary care prescribing data as sources of DDC information. However, those databases have limitations. For example, research shows substantial disagreement between nonadherence information inferred from pharmacy data and that reported by patients themselves.[7–9] Drug discontinuations due to patient nonadherence are rarely available in a structured database, and the lack of information regarding the DDC reason in the structured database makes it difficult to learn why a medication was discontinued. In contrast, clinical narratives, such as the provider notes, were identified as the "most reliable and readily accessible source" for detecting DDCs and determining reasons for them.[10] This is consistent with prior investigations which showed that a large portion of medical information is recorded only in narratives and not in the structured data.[11,12] However, manual chart reviews are time-consuming, labor-intensive, and prohibitively expensive.

Advances in natural language processing (NLP) have enabled a large-scale extraction of meaningful information from clinical narratives.[13] Several studies attempted to capture drug discontinuation signals from clinical notes in the context of patient nonadherence detection and medication reconciliation. For example, Turchin et al. used several heuristic rules to extract patient nonadherence information from physicians' notes of hypertensive patients.[14] Cimino et al. utilized the off-the-shelf NLP system MedLEE to identify medication concepts and detected drug discontinuation through the disappearance of the medication concept in patient's longitudinal notes.[15] Neither approach, however, detected the reasons for discontinuation. Morrison et al. applied the existing NLP tool TextMiner to identify AEs from clinical narratives as reasons for drug discontinuation; however, all the other reasons for discontinuation in that study depended on structured data. These shortfalls point out that developing NLP tools for a comprehensive analysis of DDC reasons is much needed.[4]

The objective of this study is to automatically identify and analyze DDC events from rich clinical narratives. We formulated the problem into a sentence-level classification task. Toward this end, we proposed a drug discontinuation taxonomy or schema, comprising of 4 high-level categories: "Adverse_Event," "Drug_Modification," "Non_Medical," and "Unknown." Examples for each category are shown in Table 1.

We applied different machine learning models and explored different representation methods. We developed a simple but novel representation method called semantic enrichment-based vector representation (SEVR), which shows promising performance for detecting drug discontinuation and reasons for discontinuation. The main contributions of this research are summarized as follows:

- a pioneer study to explore machine learning approaches for automatically identifying drug discontinuation events and reasons for discontinuation from electronic health record (EHR) narratives;
- an expert-annotated EHR corpus for drug discontinuation events and their reasons;
- a novel semantic enrichment-based vector representation;
- an evaluation of different approaches to the imbalanced data challenge.

MATERIALS AND METHODS

Data, annotation schema, annotation

The data used for this study came from the University of Massachusetts Memorial Health Care (UMMHC) EHR system. As cardiovascular disease and cancer are the largest contributors to the burden of chronic disease in the United States,[16] we sampled clinical notes from March 26, 2017 through January 8, 2015 with diagnosis codes related to cancer and cardiovascular conditions. With the approval of the institutional review board of the University of Massachusetts Medical School, we conducted an expert annotation of 1867 de-identified clinical notes through manual chart review, which focused on identifying drug discontinuation events and their reasons. The data contain 100 151 sentences, 1 237 103 words, and 2062 DDC events being annotated in 1978 sentences. For sentences containing multiple DDC events, we use the first 1 for sentence-level classification.

The overall statistics of DDC data are shown in Supplementary Table S1. The note length, sentence length, number of sentences, and number of contained DDCs vary significantly across all notes. Among 1867 notes, more than half of them (56.94%) don't have DDCs, and for those which do, 65% of them (526 out of 804) contains only 1 or 2 DDCs. On the sentence level, only ~2% of the sentences from 1867 notes contain DDCs. Therefore, although DDC events are common in patient care,[4,17,18] they are sparsely documented in clinical narratives, making NLP-based DDC detection a challenge.

For each DDC event, human experts also manually assigned a reason category using the schema we defined in Table 1. The overall inter-annotator agreement is 0.79 (kappa score) in this study. Among the 4 categories, Non_Medical DDC instances are the most common at 58%, and Adverse Event DDCs come in second at 26%, followed by Drug_ Modification and Unknown at 8% each.

NLP for discontinuation identification

We break down the task of DDC into 2 classification tasks:

- Binary detection: we classify an input sentence into 2 categories (DDC vs no DDC), referred to as "Discontinuation" hereafter.

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

Table 1. Reason categories for drug discontinuation

DDC reason category	Scope of coverage	Example
Adverse_Event	Adverse drug events leading to a DC, such as side effects, contraindication, toxicity, drug interactions, etc.	1. *Her Bactrim was discontinued and her rash is slowly resolving.* 2. *I have asked him to discontinue his Percocet as he is exceeding the Tylenol limitations per day.*
Drug_Modification	Non-AE related drug modification demanded by physician, such as drug switch due to ineffectiveness, or procedure accommodation, etc.	1. *He received a couple of doses in mid [**Date**] and treatment with Rituxan was stopped when he had obvious progression at that point.* 2. *She is about to undergo a herniorrhaphy, so I have written instructions for how to transition from Suboxone to morphine.*
Non_Medical	This category covers natural stop cases (eg, drug prescribed is finished or not needed) or patient self-discontinuation (eg, nonadherence or formulary change)	1. *The patient received 6 courses of cyclophosphamide, vincristine, and prednisone completed [**Date**].* 2. *She had refused Procrit shots of late.*
Unknown	No reason is mentioned	N/A

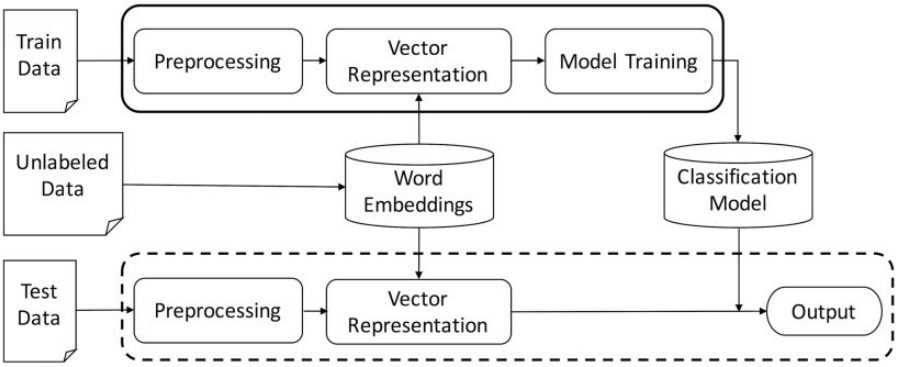

Figure 1. System Workflow. The classification model trained on the training data is applied on unseen test data.

- Multi-class classification: we classify an input sentence into 5 categories (4 reason categories in Table 1 and a "Non_DDC" category indicating no discontinuation event), referred to as "Reason" hereafter.

Figure 1 depicts our system workflow where the solid frame indicates the training process and the dashed frame represents the testing process.

Preprocessing

We performed a series of preprocessing steps on the raw clinical narratives: 1) we removed all the line breakers which typically cut sentences in the middle, 2) we conducted sentence and word tokenization and removed symbols and/or punctuations within a term, 3) morphological normalization was performed using the NLTK toolkit (http://www.nltk.org/), including word lemmatization and lowering cases. Finally, we filtered out sentences which contained less than 4 words.

Vector representation

Each input sentence is represented as a feature vector being fed into a machine learning classification algorithm. We explored 4 approaches for vector representation in this study.

First, we applied an n-gram based vector space model (NGram),[19] which has been successfully used in many information retrieval and text mining applications. Unigrams and bigrams were included as features, and term frequency–inverse document frequency (TDIDF) weighting was applied to reflect how discriminative a feature is within the current sentence.

We then explored word embedding (W2V) which represents a word as a fixed-length dense vector or embedding, so that semantically related words are close to each other in the embedding spaces. Compared with one-hot representation of NGram vectors, W2V overcomes the adverse effects of homonyms and the data sparsity issue in high-dimensional space. We used skip-gram[20] W2Vs to represent each word, and 3 aggregation approaches were tested at the sentence level: 1) maximum (pooling max value for each dimension), 2) average (averaging each dimension), and 3) inverse document frequency (IDF) weighted average (weighted averaging on each dimension based on IDF value of each word). As shown in Figure 1, we pretrained W2V on a large amount of unlabeled clinical narratives consisting of 180 000 clinical notes from UMMHC, and the dimension size for W2V was set at 200.

Next we implemented a word clustering approach[21] for a cluster-enriched vector representation (CEVR), where clusters are first generated from W2Vs using the K-means++[22] clustering method, and then NGram vectors are expanded by cluster IDFs of words appearing in a sentence.

Finally, we proposed the SEVR approach, whereby the semantic space of W2V was exploited to enrich the representativeness of original NGram vectors. The idea is that if 1 word serves as a useful feature representing 1 category, then other words with similar

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

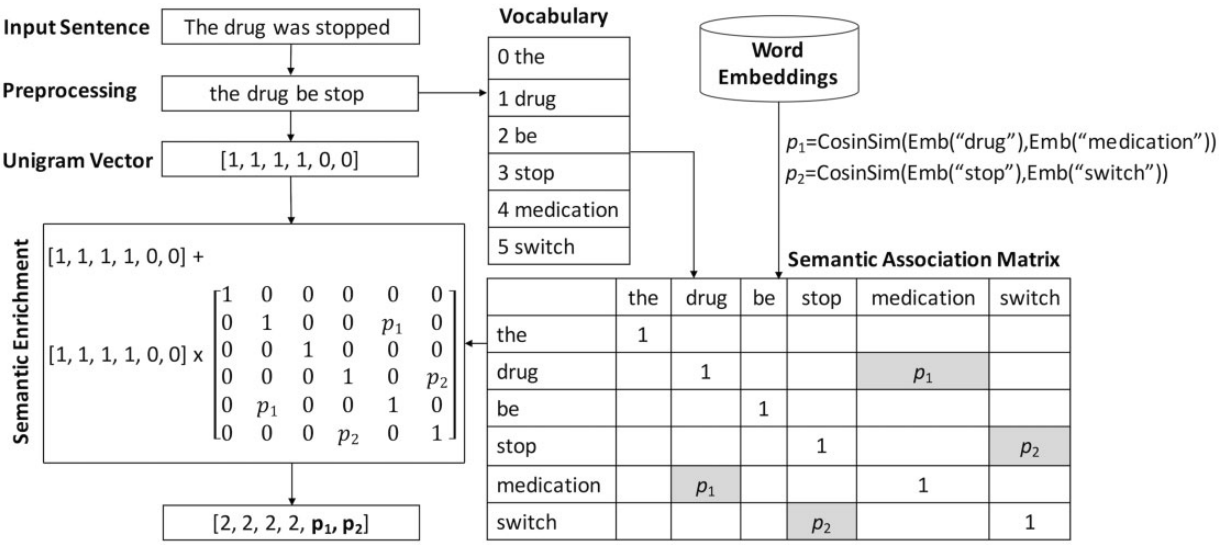

Figure 2. Illustrative example for semantic enrichment via word embeddings.

semantics should have the same representative power for that category. Thus, our hypothesis is that by inserting additional words with similar semantics into the original NGram vectors, the representativeness of the training example through SEVR will be richer and more beneficial for improved model training; this is especially true when limited training data are available or category distribution is imbalanced in multi-class classification. Compared with the CEVR approach, SEVR is expected to have better noise control by selecting only top semantically close terms in the embedding space. In addition, the weights of expanded elements are dynamically determined by the semantic similarity in SEVR, whereas they are static and treated equally in CEVR.

Specifically, despite the size of the vocabulary, we exploited W2Vs trained on a larger context to create a semantic association matrix $M^{d\times d}$ which indicated the semantic association between words. For this purpose, 2 words are considered semantically associated if 1 appears in the top k similar terms of the other, and their cosine similarity in the W2Vs space are ≥ 0.5. The diagonal values of the association matrix were initialized as 1 and the rest as 0, and then the following equation (1) was used to update nondiagonal values for semantically associated word pairs.

$$m_{ij} = \begin{cases} cosineSim\left(W_i^{emb}, W_j^{emb}\right) & \text{if } w_i \text{ and } w_j \text{ are semantically associated} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Then we transformed the original sentence vector v^d into a semantic-enriched vector s^d through equation (2) before conventional TFIDF weighting and normalization were applied.

$$s^d = v^d + v^d \times M^{d\times d} \quad (2)$$

Figure 2 shows an illustrative example for semantic enrichment on top of NGram representation. Given the input sentence "The drug was stopped," the original term-frequency–based NGram vector would $v = [1, 1, 1, 1, 0, 0]$show that the last 2 words "medication" and "switch" in the vocabulary are not present in this sentence. This representation can potentially cause low generalization for testing or under-represented semantics for training, in cases where only 1 sentence or few sentences from the same category contain "medication" and "switch." However, as "drug" / "medication" and "stop" / "switch" are 2 semantically associated word pairs in the embedding space, their similarity scores p_1 *and* p_2 $(0 \leq p_1, p_2 \leq 1)$ were assigned in the corresponding cells in the semantic association matrix respectively as shown in Figure 2. The resulting semantic-enriched vector contains richer semantics representing its associated category with enlarged semantic signals from possibly under-represented words (eg, "medication" and "switch"). In other words, through enriched transformation, signals on the same semantic dimension from different training sentences are effectively reinforced.

Model training

For both discontinuation identification subtasks, we explored 1 of the state-of-the-art classification models, Support Vector Machines (SVMs), for model training. The SVM-based algorithm has shown reliable performance in many classification tasks,[23–25] and on certain tasks, it demonstrates better performance compared with complex deep learning models.[26] We also evaluated other classic machine learning algorithms for comparative analysis.

Cost-sensitive training and resampling for imbalanced data

Imbalanced data occur when label classes are disproportionally distributed. As shown in the Data section, discontinuation events are documented sparsely in clinical narratives, which results in highly imbalanced class distribution where negative events are dominated by both Discontinuation and Reason tasks. In this study, we evaluated 3 popular approaches to deal with imbalanced data: 1) *oversampling*, the process of randomly duplicating observations from the minority class (ie, discontinuation sentences) so that its influence can be enlarged during the training; 2) *downsampling*, which involves randomly removing observations from the majority class (ie, nondiscontinuation sentences) to prevent its signal from dominating the learning algorithm; and 3) *cost-sensitive training*, which aims to increase the penalty on classification errors from the minority class (ie, discontinuation class and discontinuation reason classes). This is implemented through class-weighting in this study.

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

Table 2. Comparison among different classification algorithms

		BaseLine	SVM	KNN	NN	RF	GB	ET
Discontinuation	Sn(%)	13.65	59.86	35.08	58.52	38.54	49.77	46.78
	Sp(%)	99.93	99.42	99.87	99.47	99.88	99.47	99.04
	AUC	0.568*	0.796	0.622*	0.79	0.692*	0.746*	0.729*
Reason	Sn(%)	–	51.92	42.73	52.21	43.81	47.54	45.11
	Sp(%)	–	91.24	87.06	91.03	87.81	88.77	87.77
	AUC	–	0.716	0.649*	0.716	0.658*	0.682*	0.664*

SVM: support vector machines; DT: decision tree; KNN: K nearest neighbors; NN: neural network; RF: random forest; GB: gradient boosting; ET: extremely randomized tree * indicates the statistical significance level of $P < .001$ for Student's t test on AUC measures (compared with SVM).

Evaluation metrics

We exploited standard metrics to evaluate the classification performance: Sensitivity (Sn), Specificity (Sp), and area under the curve (AUC) metrics. For detecting discontinuation (a binary classification), the results were reported with regard to positive DDC class; whereas, for reason classification, the results are presented with the macro-average across classes (ie, computing the metric independently for each class and then averaging). Compared with the micro-average metrics (ie, aggregating the contributions of all classes weighted by the number of instances from each class) macro-average avoids the dominating effects from the majority class when the data are imbalanced, as in our case.

RESULTS

We randomly selected 200 notes (10.71%) as held-out test data and the remaining 1668 notes as training data. The results reported a 10-fold cross-validation on training data and performance on test data.

Comparison of different classification algorithms

We evaluated several machine learning models, namely, SVMs with linear kernel, K nearest neighbors (KNN), multi-layer perceptron neural network (NN), random forest (RF), gradient boosting (GB), and extremely randomized trees (ET). For implementation, we used scikit-learn toolkit[27] where grid search was performed for parameter tuning. The selected parameters are listed in Supplementary Table S2. Each sentence is represented by TFIDF-weighted NGram (unigrams and bigrams) vector representations for this set of experiments. We also developed a rule-based baseline system for the binary Discontinuation task, which identifies DDC sentences based on whether they contain any inflectional variants of 4 terms ("discontinue," "stop," "switch," and "hold"). The 10-fold cross-validation results for the classification tasks on the training data are shown in Table 2.

The results show that SVM and NN models perform comparatively well, with SVM yielding the best AUC score of 0.796 for Discontinuation and both obtaining the best macro- average AUC score of 0.716 for Reason. Student's t tests for comparing performance from other models with SVM are calculated and shown in Table 2. The SVM model significantly outperformed the rule baseline system (0.796 vs 0.568 AUC) in 10-fold cross-validation, improving the sensitivity 4 folds (59.86% vs 13.65%) while maintaining the high specificity of 99.42%. We can see for the Discontinuation task that some algorithms achieved better specificity (99.88% and 99.87% for RF and KNN) but at the cost of much lower sensitivity (38.54% and 35.08%, respectively).

For the Reason task, KNN and RF achieved the lowest macro-average performance on all 3 metrics, whereas NN yielded the best sensitivity of 52.21% and SVM obtained the best specificity of 91.24. GB and ET, 2 representative ensemble approaches, performed in the middle range among different approaches in both tasks. We observed that both tasks yielded overall high specificity values, which is related to our data imbalance issue; ie, for each class except Non_DDC, the number of negative examples is much larger than the positive ones, so that the classifier tends to focus more on the majority negative class, leading to relatively high specificity (true negative rate).

Comparison of different vector representation methods

In this section, we focus on evaluating different vector representation methods described earlier. Whereas SVM achieved similar performance to NN, the training time for the latter is much slower than SVM, so we chose the SVM learning framework for subsequent experiments (C parameters are tuned through grid search toward optimal performance). The 5 representation variants are considered as follows:

- **NGram**: NGram vector space model with TFIDF weighting including unigrams and bigrams (optimal C of SVM is 50)
- **CEVR**: clustering-enriched NGram vector representation with TFIDF weighting. We empirically tuned the number of clusters and chose 500 based on 10-fold cross-validation performance (optimal C of SVM is 50)
- **W2V_avg**: word embedding aggregated through averaging over words in the sentence (optimal C of SVM is 1 for Discontinuation and 10 for Reason)
- **W2V_idfavg**: word embedding aggregated through EDF-weighted averaging over words in the sentence (optimal C of SVM is 0.5)
- **W2V_maxmin**: word embedding aggregated through maximum and minimum pooling respectively over words and then concatenated together (optimal C of SVM is 0.5 for Discontinuation and 1 for Reason)
- **SVER**: semantically enriched NGram vector representation via W2V. We experimented with different k values (top k similar words), and chose the optimal value 2 for this experiment (optimal C of SVM is 50)

Table 3 shows that SEVR, CEVR, and NGram performed much better than other counterparts for both tasks, where SEVR outperformed NGram on both Discontinuation (Sensitivity of 61.06% vs 59.86% and AUC of 0.803 vs 0.796) and Reason (Sensitivity of 52.35% vs 51.92% and AUC of 0.719 vs 0.716) at the statistical level of $P < .05$ and $P < .1$, respectively. Compared with CEVR,

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

Table 3. Comparison among different vector representations

		NGram	CEVR	W2V_avg	W2V_idfavg	W2V_maxmin	SEVR
Discontinuation	Sn(%)	59.86	59.97	14.4	19.02	25.47	61.06
	Sp(%)	99.42	99.44	99.93	98.67	99.74	99.43
	AUC	0.796**	0.776*	0.572*	0.588*	0.626*	0.803
Reason	Sn(%)	51.92	51.82	28.25	28.15	36.01	52.35
	Sp(%)	91.24	91.18	83.44	83.29	85.07	91.43
	AUC	0.716***	0.715	0.558*	0.557*	0.605*	0.719

*, ** and *** indicate the statistical significance level of $P < .001$, $P < .05$ and $P < .1$, respectively, for Student's t test on AUC metric (compared with SVM).

Table 4. Performance on test data

		Baseline	NGram	Cluster	SEVR
Discontinuation	Sn (%)	18.31	54.46	55.4	56.34
	Sp(%)	99.88	99.45	99.43	99.36
	AUC	0.591	0.770	0.774	0.779
Reason	Sn(%)	–	51.14	51.67	53.11
	Sp(%)	–	90.48	90.39	90.86
	AUC	–	0.708	0.710	0.720

SEVR also yields slight performance gain on both tasks but is not statistically significant ($P < .1$ for Discontinuation and $P = .11$ for Reason). Among 3 word-embedding representations, W2V_maxmin performed best with AUC score of 0.626 and 0.605 for Discontinuation and Reason, respectively, and W2V_idfavg and W2V_avg presented comparable performance.

Performance on test data

Based on the 10-fold cross-validation results on training data, we applied 3 best representation settings with the SVM classifiers to evaluate system performance on unseen test data—namely NGram, CEVR, and SEVR. The results in Table 4 suggest very good generalizability for all 3 representation methods, significantly outperforming the rule baseline on Discontinuation (AUC score of 0.779 vs 0.591). Among those 3, SEVR achieved the best performance on unseen test data for both Discontinuation (AUC score of 0.779) and Reason (macro-average AUC score of 0.72).

To obtain a better understanding of how different models perform on each individual category for Reason, we present the performance per each reason category on test data in Supplementary Table S3. All 3 models performed the same on Drug_Modification, and both CEVR and SEVR achieved very close performance on Unknown. For the other 3 categories, SEVR yielded the best AUC scores of 0.805, 0.75, and 0.772, respectively. Among those categories, Drug_Modification and Unknown are shown to be relatively challenging to identify (AUC score of 0.553 and 0.719, respectively) due to much lower sensitivity on those 2 categories.

Performance of re-sampling and cost-sensitive training

We evaluated 3 strategies to overcome the highly imbalanced data based on SEVR representation. Specifically, we randomly sampled positive examples with repeat so that the total number of positive examples becomes m times the number of original positive examples for oversampling, and randomly sampled negative examples without repeat so that the total number of negative examples becomes n times the number of the positive examples for down-sampling. For cost-sensitive training, we experimented with different weight settings so that the class weight of minority class is set w times larger than the 1 of the majority class. Overall, the down-sampling method didn't work well, and we only reported the results for oversampling and cost-sensitive training.

The results in Supplementary Table S4 show that oversampling slightly improved the performance on both tasks, leading to the best AUC score of 0.785 for Discontinuation and 0.722 for Reason when oversampling parameter m equals 8 and 10, respectively.

Similarly, cost-sensitive training also yielded minor improvements on both tasks as shown in Supplementary Table S5, achieving the best AUC score of 0.78 on Discontinuation and 0.723 on Reason compared with the baseline of 0.779 and 0.72, respectively.

Based on these observations, we experimented with combining oversampling with cost-sensitive training, and the performance trends of AUC scores on 2 tasks are shown in Figure 3. The combination of cost-sensitive training with oversampling resulted in further performance gain on the Reason task, yielding the best AUC score of 0.745. Compared with the rule-based baseline for Discontinuation, the best system improved the sensitivity more than 3-fold (57.75% vs 18.31%) while maintaining the high specificity of 99.25%.

We then investigated the impact of oversampling and class-weighting on each reason category, respectively, as shown in Supplementary Table S6. We can see that through oversampling and class-weighting, the sensitivity of each reason category is significantly improved across the board (except for Drug_Modification), at the cost of a tiny number of specificity drops. The minor category Unknown benefits most from data imbalance handling strategies, improving sensitivity scores by 28.57% (56.25% vs 43.75%). We can see that Adverse_Event category achieves the best performance among all reason categories with the sensitivity of 62.96% and an AUC score of 0.814.

Error Analysis

The reason category of Drug_Modification imposes more challenges compared to the other categories. A possible reason is that the Drug_Modification category covers a wide spectrum of semantics consisting of various distinct modification concepts as well as possible overlap with the Adverse_Event category (eg, dosage change or drug switch due to AEs). For other challenging categories, such as Non_Medical and Unknown, each of them is likely to be expressed in different language variations leading to divergent semantics. By defining more refined categories, we could potentially improve the overall Reason performance, but this may require more annotated data.

In addition, overwhelming negative examples (the ratio of negative vs positive is 50.55 on the training data) may contain overlapping and ambiguous semantics among different reason categories. We summed up the confusion matrix during 10-fold cross-validation and calculated row-wise percentage as shown in Supplementary

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

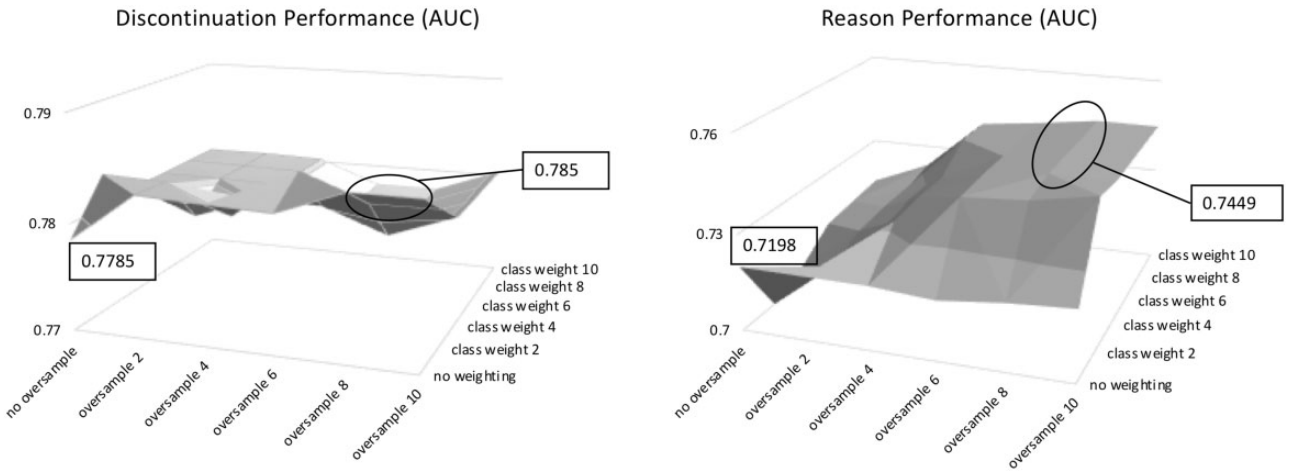

Figure 3. Performance trend on combined oversampling and cost-sensitive training.

Table S7. We can see the relatively poor sensitivity of all reason categories are all related to "Non_DDC"—especially for the Drug_Modification and Unknown categories. Causes of errors can be roughly categorized into the following groups:

1. Incomplete semantics. There are cases where larger context beyond the current sentence is needed. For example, "She wants to TAPER off of her paroxetine and wants to do this using liquid form. We will have her take a solution of paroxetine 5 mg/ml." By looking at only the second sentence, it is hard to determine whether it is a discontinuation event due to Drug_Modification or the doctor just started the drug.
2. Annotation error. For example, "He had severe pancytopenia" was labeled as a discontinuation event with the reason of Adverse_Event. But in the current context, there is no evidence showing this is an AE associated with a discontinued drug. The annotation errors are very rare (less than 1%) from what we have observed during the error analysis.
3. Mixed semantics. For example, "Diabetes was uncontrolled and patient is taken off metformin due to worsening renal function and started on Novolog and Lantus." This sentence contained both drug-stop and drug-start events which confused the system.
4. Implicit semantics. For example, "The patient was given 1 liter normal saline." The implication is that after "1 liter" the treatment was stopped naturally (Non_Medical), which lacks explicit discontinuation indicators.

DISCUSSION

Principal finding

To the best of our knowledge, this is the first study in applying machine learning-based NLP approaches to drug discontinuation detection and reason identification. We examined the performance of different algorithms and vector representation methods and found the automated system performs reasonably well. We also observed that each learning algorithm shows different advantages and disadvantages in different metrics, opening opportunities to integrate them in an intelligent way for future work.

The simple but effective SEVR approach was successfully applied on drug discontinuation detection and reason identification, showing marginally better performance on held-out unseen test data for both Discontinuation and Reason tasks compared with the traditional NGram and CEVR approaches. In addition, SEVR specifically achieved better sensitivity and AUC score for the AE reason category on Reason task, demonstrating its superior ability to pick up more AE signals due to enriched representations. Note that in equation (2) of the SEVR method we added the original n-gram frequency to the semantically enriched vector, which put more weights on the former. We performed empirical experiments regarding different enrichment mechanisms and found that the addition of original n-grams achieved better performance than directly using the enriched vector. That is possibly because the model prefers focusing more on the original n-grams but still benefits from the enriched semantics, which also makes it more robust to potentially introduced noises.

Intuitively, binary drug discontinuation detection would be considered an easy task and a simple key word matching may get good enough performance. However, we found that the rule-based baseline system suffers from a very low sensitivity of 18.31%, leading to an AUC score of 0.591 on the test data. It demonstrates that binary DDC detection from clinical narratives is actually a challenging task, and our learning model significantly improved the performance 4 folds in 10-fold cross-validation.

Resampling and Cost-Sensitive Training

Highly imbalanced data made both tasks very challenging. However, we observed that oversampling and cost-sensitive training are both helpful in mitigating this problem and combining them together can achieve further performance gain. Minor categories benefit more significantly from the imbalance handling strategies, such as Unknown. We also noticed that different reason categories prefer different sample rates and class weights, which indicates that optimizing the sample rates and class weights on individual reason categories would potentially improve the overall performance. For instance, the Adverse_Event reason category didn't seem to benefit from class- weighting as other reason categories did. In addition, jointly tuning the parameters for classifiers, sampling, and class-weighting on the training data—although computationally expensive—may further boost the system performance, which is worth exploring for future work.

Limitations

There are several limitations in this study. First, due to limited annotated data, we did not explore complicated deep learning models.

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

As we found in our experiments, using W2V itself can't beat the traditional TFIDF-weighted NGram representation. Therefore, 1 focus of our future work would be exploring data-scarce deep learning approaches (eg, learn a better sentence encoder, improve pre-trained W2Vs using unlabeled data and integrate them in a deep architecture such as convolutional neural networks,[28] or long-short term memory neural networks[29]). Second, this study assumes that the discontinuation event will be stated with reasons in the same sentence, and our analysis shows that larger context may be needed to determine the reasons. Third, in this study, our focus is recognizing high-level discontinuation reason categories. We plan to refine the category schema to accommodate more specific reason subcategories, and address challenges imposed by limited annotation samples as well as possible overlapping among subcategories.

CONCLUSION

We developed an NLP system to automatically detect discontinuation events and identify their associated reasons, where a semantic-enriched vector representation was proposed and evaluated. Our system has shown promising results, suggesting it could be applied on a large scale of EHR data for population-based observational studies. Therefore, this study has the potential to improve the efficacy of pharmacovigilance, enhance patient management, and reduce medical errors due to medication nonadherence. In addition, the rigorous comparative experiments shed light on this new clinical task and lay a solid foundation to motivate further advancement.

For future work, we will explore suitable deep learning algorithms to further improve the system's discriminative training and develop a more refined reason category schema to better profile discontinuation causes. We also plan to apply our system to EHR data from a different EHR system in order to evaluate its portability and generalizability.

FUNDING

This work was supported in part by VA Health Services Research & Development (HSR&D) in residence and National Institutes of Health (NIH) grants 3UG1DA040316-04S3 and R01HL125089. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

CONTRIBUTORS

FL designed and developed the algorithm and performed the experiments. FL and RP wrote the article. RP, ED, and EF performed chart review and created annotation guidelines. WL conducted the data collection and preprocessing. BS, FC, AG, and CP were involved in result analysis, discussion, and articleediting. HY oversaw the study, experiment design, and articlewriting.

SUPPLEMENTARY MATERIAL

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

Conflict of interest statement. None declared.

REFERENCES

1. Medication Discontinuation. Wikipedia; 2017. https://en.wikipedia.org/w/index.php?title=Medication_discontinuation&oldid=850326026. Accessed December 30, 2017.
2. Gleason KM, Groszek JM, Sullivan C, Rooney D, Barnard C, Noskin GA. Reconciliation of discrepancies in medication histories and admission orders of newly hospitalized patients. *Am J Health-Syst Pharm* 2004; 61 (16): 1689–95.
3. Lee TA, Pickard AS. *Exposure Definition and Measurement*. Agency for Healthcare Research and Quality (US); 2013. https://www.ncbi.nlm.nih.gov/books/NBK126191/. Accessed August 28, 2017.
4. Morrison FJR, Zhang H, Skentzos S, Shubina M, Bentley-Lewis R, Turchin A. Reasons for discontinuation of lipid-lowering medications in patients with chronic kidney disease. *Cardiorenal Med* 2014; 4 (3–4): 225–33. PMID:25737687
5. Pan SMD, Dehler S, Ciurea A, Ziswiler H-R, Gabay C, Finckh A. Comparison of drug retention rates and causes of drug discontinuation between anti–tumor necrosis factor agents in rheumatoid arthritis. *Arthritis Care Res* 2009; 61 (5): 560–8.
6. Cicconi P, Cozzi-Lepri A, Castagna A, *et al*.; for the ICoNA Foundation Study Group. Insights into reasons for discontinuation according to year of starting first regimen of highly active antiretroviral therapy in a cohort of antiretroviral-naïve patients. *HIV Med* 2010; 11 (2): 104–13.
7. Hansen RA, Kim MM, Song L, Tu W, Wu J, Murray MD. Comparison of methods to assess medication adherence and classify nonadherence. *Ann Pharmacother* 2009; 43 (3): 413–22.
8. Cook CL, Wade WE, Martin BC, Perri M. Concordance among three self-reported measures of medication adherence and pharmacy refill records. *J Am Pharm Assoc* 2005; 45 (2): 151–9.
9. Wang PS, Benner JS, Glynn RJ, Winkelmayer WC, Mogun H, Avorn J. How well do patients report noncompliance with antihypertensive medications? A comparison of self-report versus filled prescriptions. *Pharmacoepidemiol Drug Saf* 2004; 13 (1): 11–9.
10. Mohamed IN, Helms PJ, Simpson CR, Milne RM, McLay JS. Using primary care prescribing databases for pharmacovigilance. *Br J Clin Pharmacol* 2011; 71 (2): 244–9.
11. Turchin A, Shubina M, Breydo E, Pendergrass ML, Einbinder JS. Comparison of information content of structured and narrative text data sources on the example of medication intensification. *J Am Med Inform Assoc* 2009; 16 (3): 362–70.
12. Kramer MH, Breydo E, Shubina M, Babcock K, Einbinder JS, Turchin A. Prevalence and factors affecting home blood pressure documentation in routine clinical care: a retrospective study. *BMC Health Serv Res* 2010; 10: 139.
13. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearb Med Inform* 2017; 26 (1): 214–27.
14. Turchin A, Wheeler HI, Labreche M, *et al*. Identification of documented medication non-adherence in physician notes. *AMIA Annu Symp Proc* 2008; 2008: 732–6.
15. Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. *Stud Health Technol Inf* 2007; 129: 679–83.
16. Koene RJ, Prizment AE, Blaes A, Konety SH. Shared risk factors in cardiovascular disease and cancer. *Circulation* 2016; 133 (11): 1104–14.
17. Galli JA, Pandya A, Vega-Olivo M, Dass C, Zhao H, Criner GJ. Pirfenidone and nintedanib for pulmonary fibrosis in clinical practice: tolerability and adverse drug reactions. *Respirology* 2017; 22 (6): 1171–8.
18. Cheung A. The CATIE trial: High rates of medication discontinuation in schizophrenic patients [Classics Series]. 2 Minute Med. 2013. https://www.2minutemedicine.com/the-catie-trial-high-rates-of-medication-discontinuation-in-schizophrenic-patients-classics-series/. Accessed January 1, 2018.
19. Salton G. A vector space model for information retrieval. *CACM* 1975; 18 (11): 613–20.

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019

20. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781; 2013. https://arxiv.org/abs/1301.3781. Accessed May 20, 2017.
21. Cha M, Gwon Y, Kung HT. Language modeling by clustering with word embeddings for text readability assessment. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York, NY: ACM; 2017: 2003–6.
22. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. In: *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*; 2007.
23. Yang Y, Liu X. A re-examination of text categorization methods. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM; 1999: 42–9.
24. Joachims T. Making large-scale SVM learning practical. In: *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press; 1999: 169–84.
25. Deilmai BR, Ahmad BB, Zabihi H. Comparison of two classification methods (MLC and SVM) to extract land use and land cover in Johor Malaysia. *IOP Conf Ser Earth Environ Sci* 2014; 20 (1): 012052.
26. Zhou C, Sun C, Liu Z, Lau FCM. A C-LSTM Neural Network for Text Classification. CoRR abs/1511.08630; 2015. https://arxiv.org/abs/1511.08630. Accessed January 30, 2017.
27. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–30.
28. Kim Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics; 2014: 1746–51.
29. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.

Downloaded from https://academic.oup.com/jamia/advance-article-abstract/doi/10.1093/jamia/ocz048/5481540 by University of Massachusetts Medical School user on 12 June 2019