information

Article

Identifying Adverse Drug Reaction-Related Text from Social Media: A Multi-View Active Learning Approach with Various Document Representations

Jing Liu [1,2,*], Yue Wang [2], Lihua Huang [1], Chenghong Zhang [1] and Songzheng Zhao [3]

[1] School of Management, Fudan University, Shanghai 200433, China; lhhuang@fudan.edu.cn (L.H.); chzhang@fudan.edu.cn (C.Z.)
[2] School of Management Science and Engineering, Tianjin University of Finance and Economics, Tianjin 300222, China; wangyue@tjufe.edu.cn
[3] School of Management, Northwestern Polytechnical University, Xi'an 710072, China; zhaosongzheng@nwpu.edu.cn
* Correspondence: liujing@tjufe.edu.cn

Abstract: Adverse drug reactions (ADRs) are a huge public health issue. Identifying text that mentions ADRs from a large volume of social media data is important. However, we need to address two challenges for high-performing ADR-related text detection: the data imbalance problem and the requirement of simultaneously using data-driven information and handcrafted information. Therefore, we propose an approach named multi-view active learning using domain-specific and data-driven document representations (MVAL4D), endeavoring to enhance the predictive capability and alleviate the requirement of labeled data. Specifically, a new view-generation mechanism is proposed to generate multiple views by simultaneously exploiting various document representations obtained using handcrafted feature engineering and by performing deep learning methods. Moreover, different from previous active learning studies in which all instances are chosen using the same selection criterion, MVAL4D adopts different criteria (i.e., confidence and informativeness) to select potentially positive instances and potentially negative instances for manual annotation. The experimental results verify the effectiveness of MVAL4D. The proposed approach can be generalized to many other text classification tasks. Moreover, it can offer a solid foundation for the ADR mention extraction task, and improve the feasibility of monitoring drug safety using social media data.

Keywords: adverse drug reaction; multi-view active learning; selection strategy; document representation

check for updates

Citation: Liu, J.; Wang, Y.; Huang, L.; Zhang, C.; Zhao, S. Identifying Adverse Drug Reaction-Related Text from Social Media: A Multi-View Active Learning Approach with Various Document Representations. *Information* **2022**, *13*, 189. https://doi.org/10.3390/info13040189

Academic Editor: Norbert Fuhr

Received: 23 February 2022
Accepted: 6 April 2022
Published: 8 April 2022

1. Introduction

Adverse drug reactions (ADRs) are a huge public health issue, resulting in irreversible health consequences, millions of hospitalizations and deaths, and considerable financial losses [1,2]. For patients, pharmaceutical companies, and regulatory agencies, the timely and accurate identification of ADRs is critical. However, existing medication safety monitoring mechanisms suffer from several limitations. Pre-marketing clinical trials, for example, have homogeneous participants and are short in duration. One of the most crucial post-marketing monitoring channels, spontaneous reporting systems (SRS), is severely undervalued; it is estimated that up to 90 percent of cases are unreported [3].

Recently, social media has been a productive and additional data source for post-marketing medication safety monitoring. In the United States, 72 percent of the population uses social media actively (https://www.pewresearch.org/internet/fact-sheet/social-media/ (accessed on 1 March 2021)), and 43.55 percent of adults use medical-related platforms to seek health information [4]. However, mining ADRs from social media faces the following challenges. First, annotation is time-consuming and requires intensive domain knowledge. Second, only a small proportion of user-generated content mentions

ADRs (e.g., 7.2 percent reported in [5]). The information overload issue leads to a severely skewed data distribution that may have negative impacts on the text classification model's predictive capability. Third, there exist a variety of creative phrases, colloquial terms, and inevitable misspellings on social media. The traditional feature engineering-based model is generally incompetent to deal with these characteristics since it usually exploits shallow linguistic features, focuses on surface aspects of text, and fails to capture the semantic meaning.

This study endeavors to advance a cost-effective model to classify ADR-related text on social media among a massive volume of ADR-irrelevant text. To tackle the imbalance issue and cost annotation challenge in our context, existing studies have focused on post-remedy strategies after an imbalanced corpus has been produced, whereas we attempt to utilize active learning to reduce the degree of data imbalance when building the corpus. To capture the semantic information of text, we use deep learning-based document representation methods because they can extract hierarchical abstract features [6–8]. However, these methods ignore domain-specific knowledge because they are completely data-driven. In view of the fact that the traditional feature engineering-based method enables taking advantage of external knowledge and human ingenuity, we argue that the two types of features, being handcrafted based on external domain-specific knowledge bases and being completely data-driven, can complement each other.

Therefore, we propose a novel approach, named multi-view active learning using domain-specific and data-driven document representations (MVAL4D). Unlike existing selection strategies in active learning that measure all instances with a uniform criterion, MVAL4D adopts separate criteria, i.e., confidence and informativeness, to select instances that may belong to ADR-related text and ADR-irrelevant text for manual annotation. The novel selection strategy can aid in alleviating the imbalance issue. Concerning the simultaneous use of different strands of information, inspired by the work in [9], we propose a new view-generation mechanism. With one comprising shallow linguistic features and domain-specific knowledge-based features, and the others obtained by applying deep neural networks, various document representations serve different views in the multi-view active learning. This study is an extension of a preliminary version of the proposed approach [10] from the following aspects. On one hand, we have introduced an additional document representation derived by employing pre-trained BERT, a state-of-the-art approach for various natural language processing (NLP) tasks, to verify the scalable capability of our approach. Moreover, we have extended the preliminary version [10] by considering more view configurations in the first and second experiments to further verify the effectiveness of our approach. In addition, experiments are re-conducted, and results are re-analyzed due to the introduction of the BERT-derived document representation.

The remainder of this paper is organized as follows. In Section 2, we review prior work on identifying ADRs from social media and multi-view active learning. In Section 3, the framework of the novel multi-view active learning approach is elaborated. Specifically, we describe in detail the document representations obtained through different methods and provide an in-depth analysis of the proposed selection strategy in active learning. Section 4 details the experimental dataset and settings. We report the experimental results in Section 5, followed by the conclusions and future research directions of our work in Section 6.

2. Related Work

2.1. Identifying Adverse Drug Reactions from Social Media

Various data sources have been exploited for ADR mining, such as electronic health records [11], case reports [12], and biomedical literature [13]. It has gained increasing attention to automatically detect ADRs from social media. The exploited social platforms include both health-related forums (e.g., DailyStrength [14]) and microblogs (e.g., Twitter [8,14]). Sarker et al. [14] has found that two datasets sourced from DailyStrength and Twitter are compatible for multi-corpus training. Four consecutive shared tasks (i.e., PSB 2016 Social Media Mining Shared Task Workshop and the second, third, and fourth Social Media Mining for Health (SMM4H) Shared Tasks [15–17]) facilitate this line of research.

Existing works can be classified into four groups based on their study objectives: ADR-related text identification [2,6–8,14,18], ADR mention extraction [7,19–21], relation extraction [22,23], and concept normalization [24]. The first subtask aims to identify text mentioning ADRs. The extraction of ADR mentions is aimed at extracting structured ADR mentions; for example, extracting "dizzy" from the tweet "Feeling a little dizzy from the quetiapine I just popped." Concept relation extraction attempts to distinguish the type of relationship between a drug and an ADR mention. Concept normalization aims to map each ADR mention expressed in irregular language into terminology in the domain-specific knowledge base. For a survey of existing work at the early stages, please refer to the survey [1]. Considering that the first subtask is the focus of this paper, we only review studies regarding ADR-related text identification.

Early studies generally conducted handcrafted feature engineering and adopted traditional machine learning algorithms, e.g., the support vector machine (SVM) [2,14,18]. For example, Sarker et al. [14] has verified the effectiveness of several domain-specific features. Yang et al. [18] used the Latent Dirichlet Allocation (LDA) and partially supervised learning. Nowadays, the research tendency of the community is to abandon handcrafted feature engineering-based method for deep learning architectures [17]. Recent studies have resorted to deep learning methods, such as convolutional neural network (CNN) [6], bi-directional long short-term memory (Bi-LSTM) with the attention mechanism [6,25], and BERT [7]. In several studies, both handcrafted domain-specific features and distributed embedding features were simultaneously considered. For example, Wu et al. [6] combined word embedding, part-of-speech tag embedding, character-based representation, and other handcrafted features (e.g., sentiment scores and lexicon appearance). Dai and Wang [8] exploited word embedding, term frequency and inverse document frequency (tf-idf), negation features, and other domain knowledge features. Zhang et al. [2] extracted predicate–ADR pairs to derive holistic deep linguistic representations, which subsequently combined with shallow features.

The imbalanced data issue is a major concern that must be resolved to enhance the performance of ADR-related text classification. The necessity of conducting further investigations of this issue has been highlighted. One of the predominant studies is [8], which conducted extensive experiments to investigate the effectiveness of several popular methods and proposed a novel approach based on word embedding and the synthetic minority oversampling technique. Other methods to deal with the imbalanced problem in the context included employing a weighted SVM [14] and applying a cost-sensitive strategy [26].

2.2. Multi-View Active Learning

Active learning aims to add specifically selected instances with ground-truth labels from an oracle (e.g., an annotator). Two aspects must be considered in multi-view active learning. First, how can multiple abundant and redundant views be derived? Second, how can valuable instances be selected for manual annotation in each iteration?

Prior research generated multiple views using different strategies. In the context of dealing with multimodal information, such as image classification and video recommendation, authors generally leveraged visual and textual features to generate two views naturally [27,28]. When universal resource locator (URL)-based features are available, they can be regarded as one view, in addition to a content feature-based view [29]. Chen et al. [30] adopted different parameter configurations of generative models for modeling the action view and time view. In multi-view semi-supervised learning, randomly partitioning a high-dimensional feature space into two subsets is an alternative method [31]. However, multimodal-based and URL-based view-generation mechanisms cannot be applied given the context that only textual data are available, which is the case with most text classification tasks.

As an important component of active learning, instance selection measures can be categorized into two groups (i.e., representativeness and informativeness) [32]. Representativeness-based active learning methods attempt to select instances that can represent the entire corpus. Diversity [28], density [30], and clustering analysis [29] are primary representativeness measures. In contrast, informativeness-based methods tend to select instances with a high degree of uncertainty, which is measured based on predictive error [27] or the degree of disagreement among different classifiers [30,33]. For example, co-testing [33] selects instances from those that received inconsistent pseudo labels in different views. To make active learning models more reliable and effective, recent literature has attempted to integrate multiple query criteria [28,30,32]. Cai et al. [27] adopted a self-defined strategy considering the uncertainty of an instance and its frequency of occurrence. Yan et al. [28] combined cross-media uncertainty and diversity. Chen et al. [30] proposed a novel integrated selection strategy considering both the degree of disagreement and regional density.

2.3. Research Gap

Based on the review of prior studies, we identified the following research gaps. First, most of existing studies have used a simple concatenation of shallow linguistic features, domain-specific knowledge-based features, and deep neural network-based features to identify ADR-related text. More effective feature fusion methods should be further investigated. Second, in the context of ADR-related text identification, prior research on addressing the imbalance issue tends to adopt post-remedy strategies after the imbalanced corpus has been derived. Few studies have focused on building a corpus that is as balanced as possible in the annotation process. The existing active learning studies have provided an insufficient ability to achieve this goal because all instances that wait for manual annotation are indiscriminately measured using the same criteria.

3. The Multi-View Active Learning Approach for ADR-Related Social Media Text Identification

3.1. Framework of the Proposed Approach

Prior studies have demonstrated the capability of partially supervised learning, which exploits unlabeled data with the original labeled data [18,22]. Active learning, one of the prominent research directions in the partially supervised learning field [34], enables human–machine collaboration. Specifically, "suitable" unlabeled instances are automatically selected by the machine and are annotated afterward by human experts. We attempt to reduce the degree of data imbalance with the aid of active learning. The framework of our proposed MVAL4D is shown in Figure 1. As depicted in Figure 1, we first convert collected data into several document representations, each of which is referred to as a view. Under each view, the original labeled data are then used to construct a balanced dataset for training a classifier. Following that, active learning is repeated iteratively. In each iteration, each classifier selects valuable instances (i.e., the most confident instances among candidate positive instances and the most informative instances among candidate negative instances), and experts manually annotate the union of these instances selected by all classifiers. Finally, for each view, to refine the classifier for the following iteration, we obtain an augmented dataset by adding a newly balanced dataset, which is derived based on instances chosen by classifiers of other views, to the current iteration's dataset. We explain the framework in detail in the following paragraphs.

3.1.1. Document Representation

Converting a document to a fixed-length vector is the goal of document representation learning. The traditional technique is the vector space model [35], which is capable of capturing word co-occurrence information. Other feature space generation methods include dimensionality reduction techniques (e.g., singular value decomposition and principal component analysis) [36] and the topic modeling method LDA [9,18]. For word vector generation, a neural network-based method word2vec [37] has been developed based on the distributional hypothesis, overcoming the drawbacks of traditional one-hot encoding, i.e., high dimensionality and semantic unawareness. The distributional hypothesis implies that words with similar linguistic contexts should derive close numeric representations [38]. Following word2vec, doc2vec [39] has been proposed to further learn distributed representations of sentences, paragraphs, and documents. In recent years, other deep learning-based methods (e.g., autoencoder, CNN, LSTM, and BERT) have achieved outstanding performance in a variety of NLP tasks [6,7,25,40]. In this study, considering the fact that one of important challenges motivating our research is the time-consuming and expert-intensive annotation process, and supervised learning methods generally require the construction of a large-scale labeled corpus, we prefer to investigate unsupervised learning approaches, including stacked autoencoder (SAE), doc2vec, and pre-trained BERT (see Section 3.2.1 for details), rather than supervised learning methods, such as CNN and LSTM.

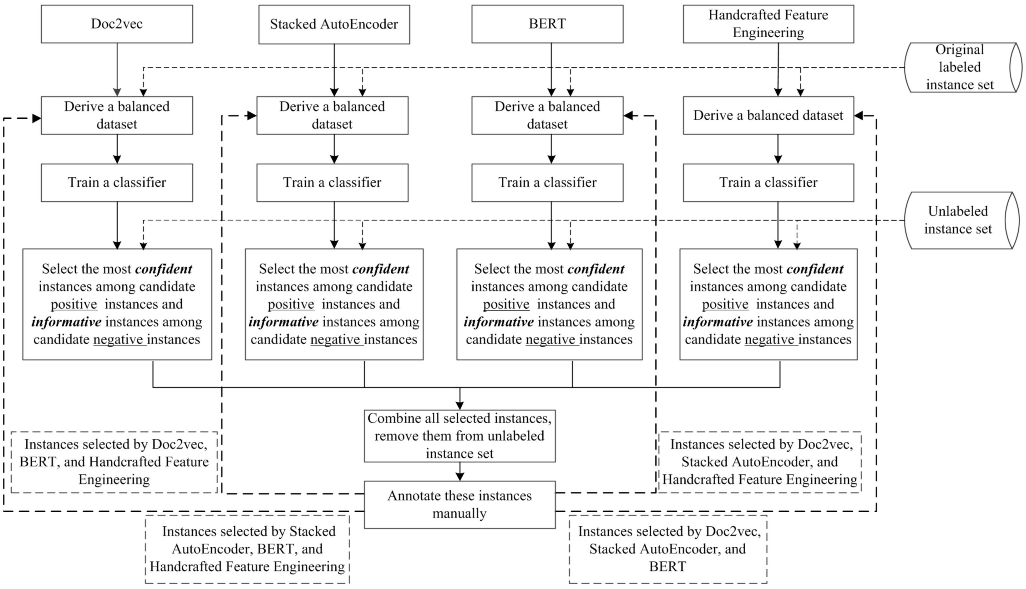

Doc2vec
Stacked AutoEncoder
BERT
Handcrafted Feature Engineering
Original labeled instance set
Derive a balanced dataset
Derive a balanced dataset
Derive a balanced dataset
Derive a balanced dataset
Train a classifier
Train a classifier
Train a classifier
Train a classifier
Unlabeled instance set
Select the most *confident* instances among candidate positive instances and *informative* instances among candidate negative instances
Select the most *confident* instances among candidate positive instances and *informative* instances among candidate negative instances
Select the most *confident* instances among candidate positive instances and *informative* instances among candidate negative instances
Select the most *confident* instances among candidate positive instances and *informative* instances among candidate negative instances
Combine all selected instances, remove them from unlabeled instance set
Instances selected by Doc2vec, BERT, and Handcrafted Feature Engineering
Instances selected by Doc2vec, Stacked AutoEncoder, and Handcrafted Feature Engineering
Annotate these instances manually
Instances selected by Stacked AutoEncoder, BERT, and Handcrafted Feature Engineering
Instances selected by Doc2vec, Stacked AutoEncoder, and BERT

Figure 1. Framework of the proposed approach.

Deep learning-based document representations are capable of effectively capturing task-independent semantic information; it is not intuitive to embed external domain knowledge into the deep neural networks. In view of the potential of handcrafted feature engineering to explore context-aware and task-dependent features, we argue that these two types of document representations can complement each other. Instead of simple concatenation that is widely used in prior studies, in this paper, we provide an alternative strategy (i.e., regarding various document representations as multiple views for multi-view active learning). In this way, domain-specific information and high-level abstract information contained in data can be simultaneously employed.

3.1.2. Selection Strategy in the Proposed Approach

In each iteration, we select the most confident instances among candidate positive (POS) instances to constitute the POS dataset and query the most informative instances among candidate negative (NEG) instances to constitute the NEG dataset. The confidence of an instance represents how likely the pseudo label of the instance is to be trusted, and the informativeness of an instance measures its degree of uncertainty. This novel selection strategy is detailed in Section 3.3.

3.1.3. Augmentation Strategy in the Proposed Approach

Different from prior studies (e.g., co-testing) in which different views share all newly annotated data, we adopt a co-training style strategy [41]. Specifically, to augment the training dataset for view v_i, we use the balanced outcome of instances selected under all other views, i.e., v_j $(j = 1, 2..N \;\&\; j \neq i)$, where N is the number of views. The augmentation strategy ensures multiple views' close cooperation and diversity, important for a high-performing ensemble model [42].

3.2. View-Generation Mechanism Using Various Document Representations

3.2.1. Stacked Autoencoder

There are two parts in an autoencoder network, i.e., encoder and decoder. An input vector I is transformed into a hidden representation R by an encoder. A decoder aims to map the hidden representation R back to I in the input space. When training a model using an autoencoder, the goal is to retain information encoded in the input data to the greatest extent by minimizing inconsistencies between the original input vector and the reconstructed vector (i.e., minimizing reconstruction error). Stacked autoencoder (SAE) contains multiple layers compared with a conventional autoencoder, and therefore has the potential to enhance the ability of representing nonlinearities and capturing abstract information. To train a SAE model, the greedy layer-wise training technique is used [43]. Specifically, each layer's resulting hidden representation serves as the input of its following layer. The final document representation prepared for our task is the obtained hidden representation at the last layer.

3.2.2. Doc2vec

To learn continuous and distributed vectors of variable-length sentences, paragraphs, and documents, doc2vec (also known as paragraph vector) [38] has been developed based on the idea of word2vec [37]. In this study, we employ the distributed memory model of paragraph vectors (PV-DM) since prior research has demonstrated the effectiveness of this structure for most tasks. The paragraph representations in PV-DM can be obtained by predicting the next word based on paragraph vectors and context word vectors. One advantage of the PV-DM model is that word order information is embodied by considering the context. In the training stage, for seen paragraphs, corresponding word and paragraph vectors are simultaneously trained. Word vectors are shared by paragraphs, whereas each paragraph is converted into a unique vector. In the inference phase, word vectors are fixed while the new paragraph vectors are trained until convergence.

3.2.3. BERT

BERT is proposed to jointly consider contexts from both directions in all layers, in order to pre-train deep bidirectional representations from massive unlabeled text [44]. The architecture of BERT is made up of multi-layer bidirectional transformers, in which self-attention is adopted to conduct parallel calculation of word pair relationships [45]. For downstream tasks, there are generally two steps involving in BERT: pre-training and fine-tuning. Pre-training aims to generate contextual token representations using unlabeled text. Pre-trained parameters are used to initialize the BERT model and then fine-tuned using labeled text in the fine-tuning phase. As mentioned in Section 3.1.1, in this study, we used the pre-trained BERT model without labeled text involvement. Two supervised tasks are carried out to guide the pre-training of a BERT model: masked language model (MLM) and next sentence prediction (NSP). The MLM task masks some tokens at random and aims to predict each token's original vocabulary ID by fusing its right and left contexts. The objective of the NSP task is to predict whether a text pair is “IsNext” or “NotNext”.

3.2.4. Handcrafted Feature Engineering

Feature engineering is a crucial process in the conventional machine learning community. It usually explores shallow linguistic information (e.g., n-grams) as the baseline.

Furthermore, experts with domain expertise often turn to external resources and domain-specific knowledge bases to generate additional features. We conduct feature engineering following [14]. In this paper, we provide a brief introduction of explored features using several domain-specific knowledge bases and linguistic resources, as shown in Table 1. For more detailed information, please refer to [14].

Table 1. Explored features in the handcrafted feature engineering.

Feature Type	Feature	Description	Knowledge Base/Tool
Shallow linguistic features	N-grams with tf-idf	Contiguous (n = 1, 2, 3) tokens in the text	/
Domain knowledge-based features	Medical semantic features	Concept IDs and semantic types that represent fine and broad categories of medical concepts	Unified Medical Language System (UMLS) [1]/MetaMap [2]
	The ADR lexicon match-based features	A flag of whether ADR mentions in a constructed lexicon are contained in the instance	COSTART [3], MedEffect [4], SIDER [5], and Consumer Health Vocabulary (CHV) [6]
		The number of ADR mentions	
	The negation features	Negated concepts	NegEx [7] incorporated in MetaMap
Other discriminative features	Synonym expansion features	Synonyms for each noun, verb, and adjective in an instance	WordNet [8]
	Change phrase features	less-good, more-good, more-bad, and less-bad	/
	Sentiword score feature	The overall sentiment score divided by the length of the instance	SentiWordNet [9]
	Topic-based features	Topic terms	Mallet [10]

[1]: https://www.nlm.nih.gov/research/umls/index.html, UMLS is a compendium which encompasses a large number of medical vocabularies (accessed on 15 March 2020). [2]: https://mmtx.nlm.nih.gov/ (accessed on 15 March 2020). [3]: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/index.html (accessed on 15 March 2020). [4]: https://www.canada.ca/en/health-canada/services/drugs-health-products/medeffect-canada.html (accessed on 15 March 2020). [5]: http://sideeffects.embl.de/ (accessed on 15 March 2020). [6]: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html (accessed on 15 March 2020). We use CHV since users on social media prefer to describe ADRs with colloquial language, rather than technical terms. [7]: https://code.google.com/p/negex/ (accessed on 15 March 2020). [8]: https://wordnet.princeton.edu/ (accessed on 15 March 2020). [9]: https://github.com/aesuli/SentiWordNet (accessed on 15 March 2020). [10]: http://mallet.cs.umass.edu/ (accessed on 15 March 2020).

3.2.5. Advantages of the Proposed View-Generation Mechanism

The view-generation mechanism in our work has two main advantages. First, it provides a bridge between the desire to simultaneously use different levels of information and the requirement to generate abundant and redundant views in multi-view active learning. Second, compared with existing view-generation methods, the mechanism is more feasible and generic for text classification tasks. For example, compared with methods using visual and textual features [27,28] and methods using content and URL descriptors [29], our work is more feasible because visual features and URL descriptors are unavailable in most social media-based text classification applications. Moreover, the proposed view-generation mechanism is generic across diverse applications. Unsupervised deep learning-based document representations are merely dependent on massive unlabeled data, which can be collected from social media. Domain-specific features are also available in most cases because feature engineering is a basic task for the deployment of traditional machine learning algorithms. Even in the worst circumstance in which domain-specific features cannot be extracted, we can use only shallow linguistic features (e.g., tf-idf), following the prior work [9].

3.3. Selection Strategy in the Proposed Approach

3.3.1. Motivation of the Selection Strategy

The objective of our proposed instance selection strategy is to reduce the degree of imbalance of newly labeled data and to build a corpus that is as balanced as possible. This objective can be achieved with the aid of deciding which instances to annotate in each iteration of active learning. To achieve this goal, an intuitive idea is to select positive instances and negative instances, which are predicted by the classifiers, for manual annotation. However, due to the generalization error of classifiers, false positives (FP, inferred as ADR-related but annotated as ADR-irrelevant) and false negatives (FN, inferred as ADR-irrelevant but annotated as ADR-related) are inevitable, as illustrated in Figure 2. In an ideal scenario where the number of FPs is equivalent to that of FNs, the effect of the FPs and FNs can be canceled out when constituting true positive instances (POStrue) and true negative instances (NEGtrue) (as depicted in Figure 2).

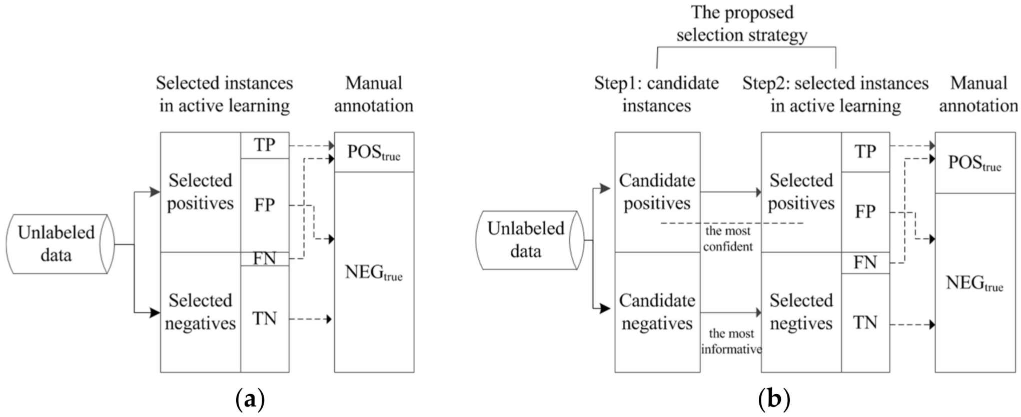


Figure 2. Motivation of the proposed selection strategy. (**a**) Active learning without our proposed selection strategy; (**b**) active learning using our proposed selection strategy.

However, this is not the case for the ADR-related text identification task. The highly skewed data distribution in our context leads to low precision and high recall, meaning that the number of FPs is far larger than that of FNs and consequently results in the phenomenon that the number of POStrue instances is significantly smaller than that of NEGtrue instances, as presented in Figure 2. To alleviate the imbalance degree, it is important to increase the number of TPs, which means that potentially positive instances that are selected automatically by classifiers are true positives when manually annotated to the greatest extent. To achieve this goal, when selecting potentially positive instances in active learning, we attempt to select the most confident instances, resembling the strategy adopted in semi-supervised learning. Following traditional active learning methods, when selecting potentially negative instances, we attempt to select the most informative instances for an improved predictive performance. The illustrated example presented in Figure 2 denotes that the degree of data imbalance is significantly alleviated by adopting our proposed selection strategy.

Before introducing our proposed selection strategy in detail, we provide a brief description of the selection mechanism in co-testing [33], a well-known multi-view active learning approach. In co-testing-style methods, firstly, the model identifies contention instances in which there exists a certain degree of disagreement among different views regarding predicted labels. Subsequently, to determine the final chosen instances for manual annotation, co-testing adopts a uniform criterion for potential positive and negative instances. Inspired by these methods, our proposed selection strategy works in a similar way. However, unlike co-testing that uses a uniform criterion, our proposed selection strategy consists of two steps: finding candidate positive and negative instances, and determining selected positives and negatives for manual annotation by adopting different selection criteria.

3.3.2. Finding Candidate Positive and Negative Instances

1. Finding candidate positive instances

For candidate positive dataset generation, confident instances are required; therefore, we refer to disagreement-based semi-supervised learning, such as co-training [41], tri-training [46], and CoForest [47]. Specifically, we first train a classifier under each view. For the view i, we measure the confidence of an unlabeled instance $\boldsymbol{x}$ by performing the strategy of "majority teaches minority" with the aid of other classifiers [22].

$$\varphi(\boldsymbol{x}, i) = \frac{\max(m,\ N-1-m)}{N-1}, \tag{1}$$

where N represents the number of classifiers (that is, the number of views), and m is the number of other classifiers which classify the instance $\boldsymbol{x}$ as positive. For an instance, only when the following conditions are satisfied can it be considered as a candidate positive instance: $\varphi(\boldsymbol{x}, i) > \varphi$ where φ denotes a pre-defined threshold, and $m \geq \frac{N-1}{2}$ indicating that the other classifiers predict the instance $\boldsymbol{x}$ as ADR-related (the positive class) with the majority voting scheme.

2. Finding candidate negative instances

For candidate negative dataset generation under the view i, we query instances whose pseudo labels are negative predicted by h_i, mainly out of the following consideration. As illustrated in Figure 2, the number of FNs is fairly small. Moreover, FNs can contribute to POStrue, which is beneficial to alleviate the imbalance degree of a heavily skewed dataset.

3.3.3. Determining Selected Positives and Negatives for Manual Annotation

1. Determining selected positive instances

Among candidate positive instances, we increase the assurance of confidence by choosing the most confident instances using an additional confidence metric [9]. Specifically, for each view i, to measure the confidence of an unlabeled instance $\boldsymbol{x}$, we use the minimum value of confidence values computed under the other views. In this way, multiple views can complement each other.

$$C(\boldsymbol{x}, i) = argmin_{j=1,\,2,\ldots,N\ \&\ j\neq i}\left(-E_{\mathrm{h_j}} \times H(\boldsymbol{x}, j)\right), \tag{2}$$

where $E_{\mathrm{h_j}}$ is the error of the classifier h_j for view j, and $H(\boldsymbol{x}, j)$ represents the entropy function computed as follows:

$$H(\boldsymbol{x}, j) = -\sum_{c=1}^{2} P_{jc}(\boldsymbol{x}) log P_{jc}(\boldsymbol{x}). \tag{3}$$

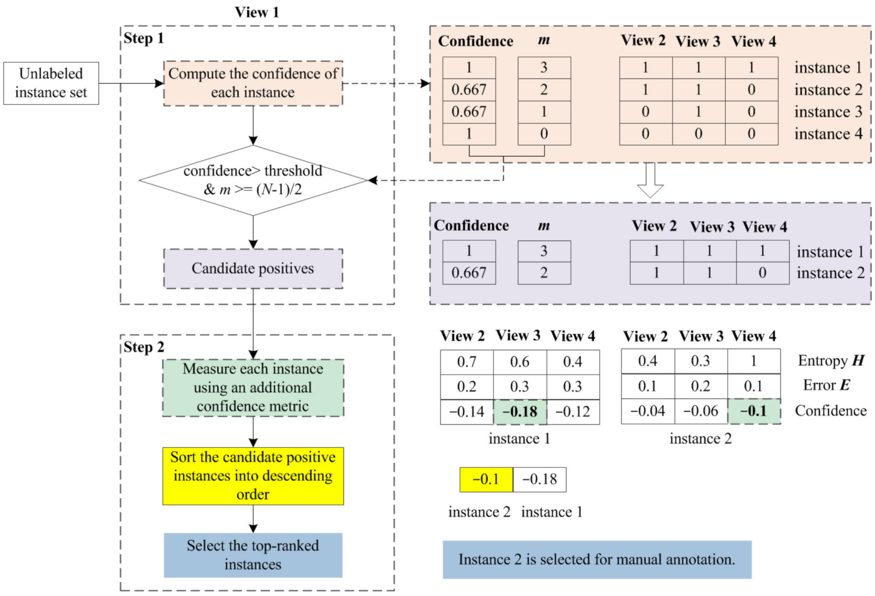
Candidate positive instances are sorted in descending order based on the confidence computed using Equation (2). Then, the top-ranked instances are regarded as potentially positive instances, which are selected for manual annotation. An illustrated example from the perspective of View 1 is described in Figure 3. For other views, the positive instance selection strategy works in the same way.

2. Determining selected negative instances

Given the candidate negative dataset, for each view i, to measure the confidence of an unlabeled instance $\boldsymbol{x}$, we use the maximum value of confidence values computed under the other views:

$$C'(\boldsymbol{x}, i) = argmax_{j=1,\,2,\ldots,N\ \&\ j\neq i}\left(-E_{\mathrm{h_j}} \times H(\boldsymbol{x}, j)\right). \tag{4}$$

A low confidence represents a high informativeness. Therefore, to select the most informative instances, candidate negative instances are sorted in ascending order based on the confidence computed using Equation (4). Then, the top-ranked instances are regarded as potentially negative instances, which are selected for manual annotation.


View 1
Step 1
Unlabeled instance set
Compute the confidence of each instance
confidence> threshold & *m* >= (*N*-1)/2
Candidate positives
Step 2
Measure each instance using an additional confidence metric
Sort the candidate positive instances into descending order
Select the top-ranked instances

Confidence	*m*	View 2	View 3	View 4	
1	3	1	1	1	instance 1
0.667	2	1	1	0	instance 2
0.667	1	0	1	0	instance 3
1	0	0	0	0	instance 4

Confidence	*m*	View 2	View 3	View 4	
1	3	1	1	1	instance 1
0.667	2	1	1	0	instance 2

View 2	View 3	View 4	View 2	View 3	View 4	
0.7	0.6	0.4	0.4	0.3	1	Entropy ***H***
0.2	0.3	0.3	0.1	0.2	0.1	Error ***E***
−0.14	**−0.18**	−0.12	−0.04	−0.06	**−0.1**	Confidence
instance 1			instance 2			

−0.1	−0.18
instance 2	instance 1

Instance 2 is selected for manual annotation.

Figure 3. Illustrated example of the proposed selection strategy for potentially positive instances.

3.4. Pseudocode for the Proposed Approach

Overall, the pseudocode of our proposed MVAL4D method is described in Algorithm 1. The proposed MVAL4D approach is scalable and generalized from the following perspectives. Regarding scalability, first, document representations derived by other methods can be naturally incorporated into MVAL4D by adding new views. Second, when selecting instances that belong to the majority class in active learning, the substitution of other selection measures (e.g., the representativeness measure) for the informativeness measure can be conducted. Regarding generalization, first, the proposed method can be generalized to other applications that face similar challenges. For example, the developed selection strategy can be applied to other tasks that implement active learning on imbalanced data. Second, the proposed view-generation mechanism may benefit other multi-view learning, such as multi-view semi-supervised learning.

Algorithm 1. MVAL4D

```
Input:
$L$: initial labeled instance set
$U$: unlabeled instance set
N: number of views
$L_i$: labeled instance set for the $i$th classifier ($i$ = 1,2, . . . ,$N$)
h: classification learning algorithm
$T$: maximum number of iterations
$L'_i$: set of instances that the $i$th classifier selects in each iteration ($i$ = 1,2, . . . ,$N$)
$K$: predefined total number of instances to label in each iteration
$\varphi$: confidence threshold
q: the number of selected negative instances divided by the number of selected positive instances
in each iteration ($0 \leq q \leq 1$)

Process:
1. Generate $N$ views using different methods (e.g., Doc2vec, average Word2vec, stacked
autoencoder, and handcrafted feature engineering)
2. Use the under-sampling strategy to address imbalance problem on the original labeled instance
set $L_i \leftarrow undersample\ (L)\ (i$ = 1,2,...,$N)$
3. Train a classifier for each view: $h_i \leftarrow h(L_i)\ (i = 1,\ 2, \ldots, N)$
4. For each view $i$ ($i$ = 1,2,..., $N$), select instances to constitute the $L_i{}'$:
```

Algorithm 1. ***Cont.***

```
4.1. To find potentially positive instances, select the $\frac{K}{N*(1+q)}$ most confident instances from $U$ to constitute $POS'_i$, as described in Section 3.3.2
4.2. For potentially negative instances, select the $q\left|POS'_i\right|$ most informative instances from $U$ to constitute $NEG'_i$, as described in Section 3.3.3
4.3. $L'_i \leftarrow POS'_i \cup NEG'_i$
5. Obtain the union of selected instance sets $L' = L'_1 \cup L'_2 \cup \ldots \cup L'_N$ and manually label these instances
6. For each view, combine $L_i$ with the union of selected instance sets derived by all other classifiers (with ground-truth labels) $L_i \leftarrow L_i \cup undersample\ (L'_i)$ ( $j$ = 1,2, . . . , $N$ & $j \neq i$)
7. Remove the union of selected instance sets $U = U - L'$
8. Repeat Steps 3 through 7 T times or until $U$ is $\varnothing$
Output: $F(x) = arg\ max_{y \in Y} \sum_{i=1}^{N} 1(y = h_i(x))$ % Majority voting scheme is adopted
```

4. Experimental Datasets and Settings

4.1. Experimental Dataset

To investigate the effectiveness of our proposed framework, we used an open-source dataset (http://diego.asu.edu/Publications/ADRClassify.html (accessed on 1 March 2020)) that consists of 10,822 instances collected from Twitter [14]. The released dataset does not directly contain actual tweets due to privacy, and therefore, we downloaded each corresponding tweet text based on the tweet ID and user ID. Due to the fact that several tweets have been removed, a total of 7060 instances were obtained, which include 6304 negative instances and 756 positive instances. The ratio of the ADR-related (positive) class to the ADR-irrelevant (negative) class is 1:8.34, indicating a significant imbalance degree. Moreover, approximately 2 million sentences were collected from a health-related forum, in order to support unsupervised document representation learning. We conducted data preprocessing, such as text tokenization, lemmatization, removing short sentences, removing hypertext mark-up language (HTML) tags using a custom regular expression, and converting text to lowercase. When exploring shallow linguistic features, we removed stop words and tokens whose frequencies were less than 3.

4.2. Evaluation Metrics

To evaluate the performance of the proposed MVAL4D, we used the average accuracy (AA) and the AUC of ROC (Receiver Operating Characteristics). The former evaluation metric is widely used for a classification task, and we introduced the latter one because it is suitable for the imbalanced dataset due to its invariance to the class distribution.

4.3. Experimental Procedure

A series of experiments were performed to evaluate the effectiveness of MVAL4D. Experiment A explored different document representation combinations to examine the complementary nature of various document representations. Experiment B assessed the effectiveness of the proposed selection strategy in terms of improving the predictive performance. Experiment C conducted sensitive analysis by using different initial rates of labeled instances. Experiment D further examined the effectiveness of MVAL4D compared to the supervised learning with different document representations and all labeled instances (i.e., without using active learning) and another baseline, i.e., a fined-tuned BERT model. For all experiments, we used 10-fold cross-validation, in order to reduce the influence of data variability.

4.4. Experimental Settings

We employed SVM as the classification algorithm following prior studies [8,14] and implemented it using LibSVM in WEKA Waikato Environment for Knowledge Analysis (WEKA) (http://www.cs.waikato.ac.nz/mL/weka/ (accessed on 15 March 2020)) package. The initial labeled rate is set at 40% for all experiments except Experiment C to guarantee the accuracy of the classifiers and to allow sufficient room for active learning. The q in Algorithm 1 is set to be 0.4 for all experiments. We used the genism (https://radimrehurek.com/gensim/ (accessed on 1 April 2020)) package to implement doc2vec. We adopted the released pre-trained model, specifically the BERT-base and uncased model, without performing fine-tuning. The hidden dimensionalities for each layer in the SAE model were 10,000, 5000, and 300, respectively. BERT was implemented using TensorFlow. We used a release source code (http://diego.asu.edu/Publications/ADRClassify.html (accessed on 15 March 2020)) to conduct the handcrafted feature engineering. The proposed MVAL4D approach was implemented in-house using WEKA.jar. We conducted parameter tuning to determine several parameters' values, which are listed in Table 2. Except when stated otherwise, we used default values of other parameters.

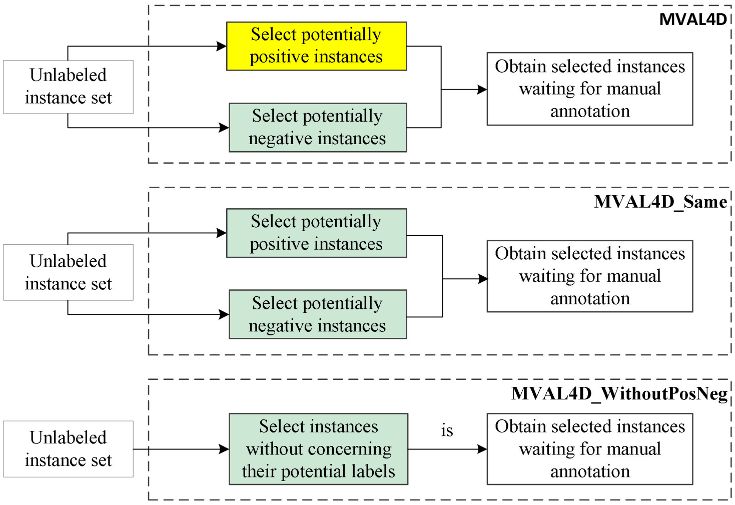
Table 2. Some parameters for various document representations.

Document Representation	# Features	C	Window	Min_Count	Max_Seq_Length
Doc2vec	128	16	5	60	/
BERT	768	4	/	/	35
Stacked autoencoder	300	16	/	/	/
Feature engineering	15,657	256	/	/	/

For these experiments, some abbreviations and their descriptions are listed in Table 3. Different selection strategies compared in the experiments are depicted in Figure 4.

Table 3. Abbreviations and their descriptions in the experiments.

Abbreviation	Description
D2V_BERT_SAE_FE	DRs fusing doc2vec, pre-trained BERT, stacked autoencoder, and feature engineering
D2V_FE	DRs fusing doc2vec and feature engineering
BERT_FE	DRs fusing pre-trained BERT and feature engineering
SAE_FE	DRs fusing stacked autoencoder and feature engineering
D2V_ BERT_SAE	DRs fusing doc2vec, pre-trained BERT, and stacked autoencoder
MVAL4D_Same	Separately selecting potentially positive and negative instances with the informativeness-based selection criterion
MVAL4D_ WithoutPosNeg	Using the informativeness-based criterion to select instances, without concerning their pseudo labels


MVAL4D
Unlabeled instance set
Select potentially positive instances
Select potentially negative instances
Obtain selected instances waiting for manual annotation
MVAL4D_Same
Unlabeled instance set
Select potentially positive instances
Select potentially negative instances
Obtain selected instances waiting for manual annotation
MVAL4D_WithoutPosNeg
Unlabeled instance set
Select instances without concerning their potential labels
is
Obtain selected instances waiting for manual annotation

Note: the figures in yellow denote that the confidence measure is used, whereas the figures in green represent that the informativeness measure is used.

Figure 4. Different selection strategies compared in the experiments.

5. Results and Discussion

5.1. Experiment A: Effectiveness of Different View Configurations

We first analyze our experimental results in terms of views in multi-view active learning. We compared five view configurations, in which each document representation is regarded as a view. Four out of these five view configurations simultaneously use the data-driven and domain-specific document representations (with the exception of "D2V_ BERT_SAE"). Besides reporting results obtained by MVAL4D, we also provide results obtained by training supervised classifiers with equal numbers of total labeled instances with each view configuration and fusing these classifiers with majority voting (referred to as "MV_SL"). The experimental results are depicted in Table 4; the highest AA, AUC, and their improvements compared to "MV_SL" (referred to as "AA ↑" and "AUC ↑") are boldfaced. As presented in Table 4, exploring document representations based on doc2vec and handcrafted feature engineering (i.e., "D2V_FE") obtains the highest accuracy, AUC values, and AUC improvements (82.51%, 0.8823, and 1.71%, respectively). The results verify the complementary nature of the doc2vec-derived representation and handcrafted features. Moreover, the superiority of using "D2V_FE" over "D2V_BERT_SAE_FE" demonstrates the fact that it is not bound to deliver improved performance by introducing more sources of information. The reason may lie in the small size of the used training dataset. In addition, for each view configuration, employing active learning with our proposed MVAL4D approach achieves enhanced predictive capability compared with "MV_SL", demonstrating the effectiveness of the MVAL4D method. Moreover, the standard deviations of AA and AUC values of performing MVAL4D on the "D2V_ BERT_SAE_FE" are the lowest (0.010 and 0.018, respectively) among all view configurations.

Table 4. Performance of the proposed approach with different view configurations.

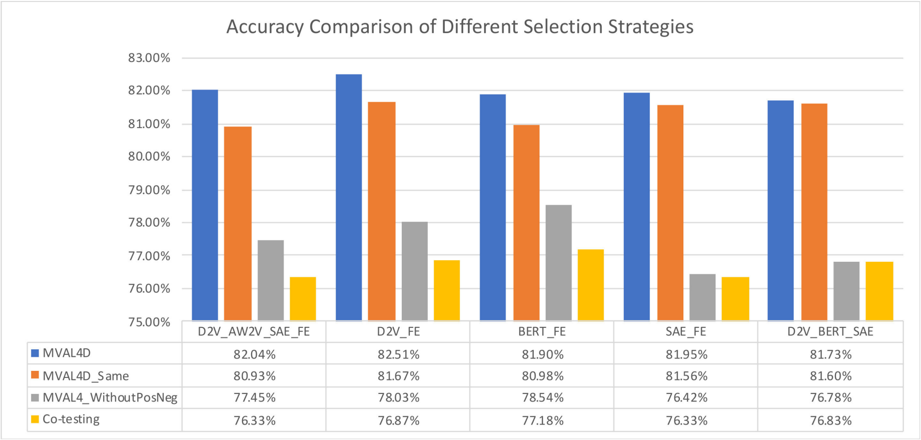
View Configurations	MVAL4D Recall		MV_SL F1_Score		AA↑	AUC↑
	AA	AUC	AA	AUC		
D2V_BERT_SAE_FE	82.04%	0.8816	76.57%	0.8705	7.14%	1.27%
D2V_FE	**82.51%**	**0.8823**	77.69%	0.8675	6.20%	**1.71%**
BERT_FE	81.90%	0.8734	77.12%	0.8697	6.19%	0.42%
SAE_FE	81.95%	0.8751	77.32%	0.8691	5.98%	0.70%
D2V_BERT_SAE	81.73%	0.8748	75.40%	0.8633	**8.40%**	1.33%

5.2. Experiment B: Effectiveness of Different Selection Strategies

For effectively performing active learning on imbalanced dataset, we tailor a selection strategy as described in Section 3.3. To evaluate the effectiveness of our proposed selection strategy, we compared it with two other strategies, i.e., "MVAL4D_Same" and "MVAL4D_WithoutPosNeg", as described in Section 4.4. Moreover, for comparison purpose, we implemented co-testing with the conservative query strategy. The results in Figures 5 and 6 suggest that our proposed selection approach can outperform other selection methods in terms of both accuracy and AUC.

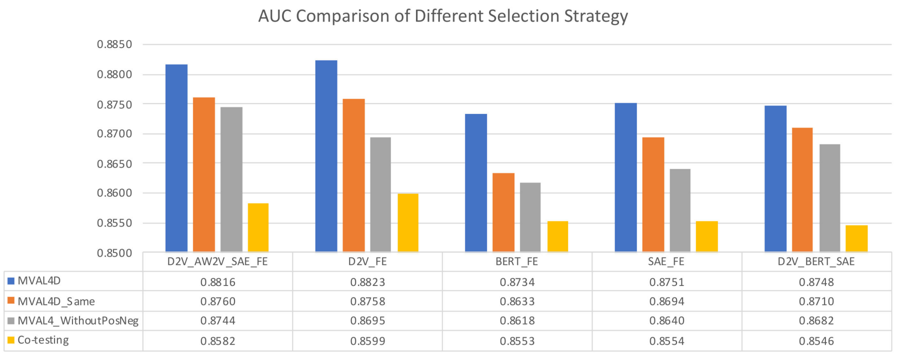
5.3. Experiment C: Performance Comparison with Different Numbers of Initial Labeled Instances

To validate the effectiveness of MVAL4D with different numbers of initial labeled instances, we used 20%, 40%, 60%, and 80% training data as the initial labeled data and the rest of the training data as unlabeled data. In terms of accuracy, as depicted in Figures 7 and 8, with the increasing numbers of initial labeled data, the accuracy values present a decreased trend, while the trend of AUC values is not obvious. Therefore, the perfect moment to start implementing MVAL4D is not easy to determine. In this case, we can use the model obtained using all document representations (i.e., "D2V_AW2V_SAE_FE") since it is the most robust across different numbers of initial labeled instances.



	D2V_AW2V_SAE_FE	D2V_FE	BERT_FE	SAE_FE	D2V_BERT_SAE
MVAL4D	82.04%	82.51%	81.90%	81.95%	81.73%
MVAL4D_Same	80.93%	81.67%	80.98%	81.56%	81.60%
MVAL4_WithoutPosNeg	77.45%	78.03%	78.54%	76.42%	76.78%
Co-testing	76.33%	76.87%	77.18%	76.33%	76.83%

Figure 5. Accuracy comparison of different selection strategies.



	D2V_AW2V_SAE_FE	D2V_FE	BERT_FE	SAE_FE	D2V_BERT_SAE
MVAL4D	0.8816	0.8823	0.8734	0.8751	0.8748
MVAL4D_Same	0.8760	0.8758	0.8633	0.8694	0.8710
MVAL4_WithoutPosNeg	0.8744	0.8695	0.8618	0.8640	0.8682
Co-testing	0.8582	0.8599	0.8553	0.8554	0.8546

Figure 6. AUC comparison of different selection strategies.

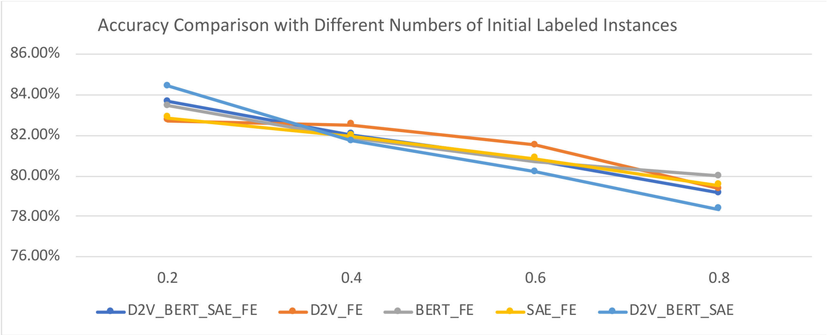


Figure 7. Accuracy comparison with different numbers of initial labeled instances.

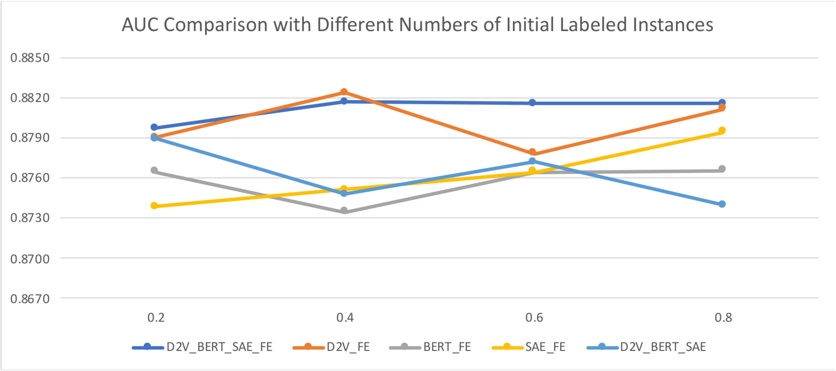


Figure 8. AUC comparison with different numbers of initial labeled instances.

5.4. Experiment D: Comparison between the Proposed Approach and Other Methods

When implementing MVAL4D, we use the view configuration of “D2V_FE” since it has yielded the highest performance. We compared our method with three types of baselines: majority voting-based ensembles that fuse multiple classifiers trained with different document representations, the supervised classifier using an individual document representation, and a fined-tuned BERT model. All of these baselines are obtained using all labeled instances in the training dataset, and the former two types are entitled with the suffix of “SL”, which is the abbreviation of “Supervised Learning”. The “FE_SL” represents the method in [14].

As shown in Figures 9 and 10, our proposed approach outperforms all baselines. Moreover, it is noteworthy that the improvements are obtained with fewer labeled instances in our method (5146 labeled instances) compared to 6354 labeled instances in baselines. This advantage reduces the annotation cost and enhance the feasibility of conducting drug safety surveillance from social media. Additionally, baselines belonging to the first type generally achieve improved performances over single document representation-based models in most cases. In addition, it is interesting to observe performances of “BERT_SL” and “BERT_FE_SL”. Specifically, using single BERT-based document representation yields the lowest accuracy and the second lowest AUC value, i.e., 75.64% and 0.8615, respectively. However, combining BERT and handcrafted features achieves a significant AUC improvement over “BERT_SL” and “FE_SL”, and delivers an AUC value of 0.8764, which is the second highest AUC value, only surpassed by our proposed method. This finding lends strong support to our intuition that handcrafted features and data-driven features can complement each other, and therefore, simultaneously leveraging them can contribute to the enhanced predictive capability.

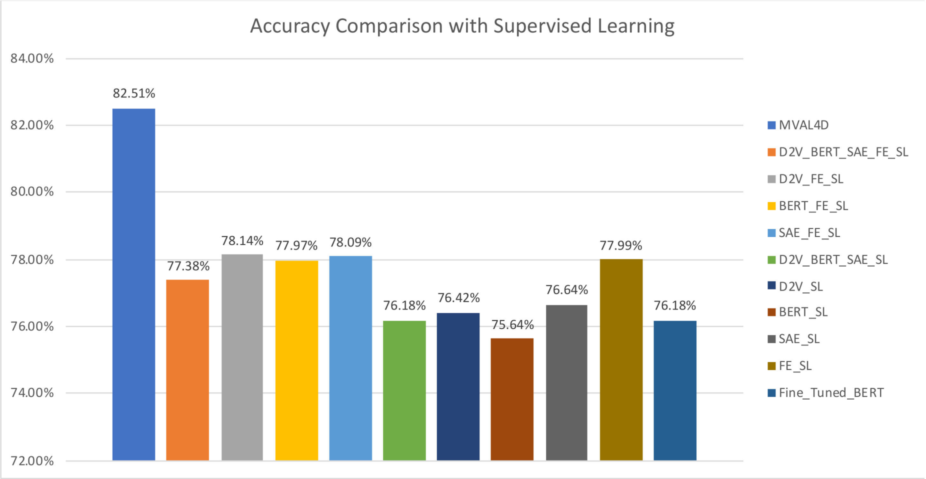


Figure 9. Accuracy comparison with supervised learning.

5.5. Discussions

First, we analyzed sentences that are misclassified by the feature-based method or data-driven methods (e.g., doc2vec) and those predicted correctly by our proposed approach. It was observed that MVAL4D can deal with diverse expressions and non-standard terms effectively. For example, the text “This night of no sleep is brought to you by Vyvanse.” was predicted as ADR-irrelevant class by the feature-based method, but can be identified correctly by MVAL4D. The reason may be that the inclusion of a deep learning-based method in MVAL4D enhances its capability of capturing semantic information. Moreover, the analysis demonstrated the limited capability of doc2vec in dealing with short text and distinguishing between ADRs and drug indications. With the inclusion of domain-specific knowledge, MVAL4D can alleviate this situation. For example, MVAL4D can correctly predict the text “Depression hurts, cymbalta can help”, which was misclassified by the doc2vec-based model. The abovementioned findings support our motivation of fusing data-driven semantic information and domain-specific information.

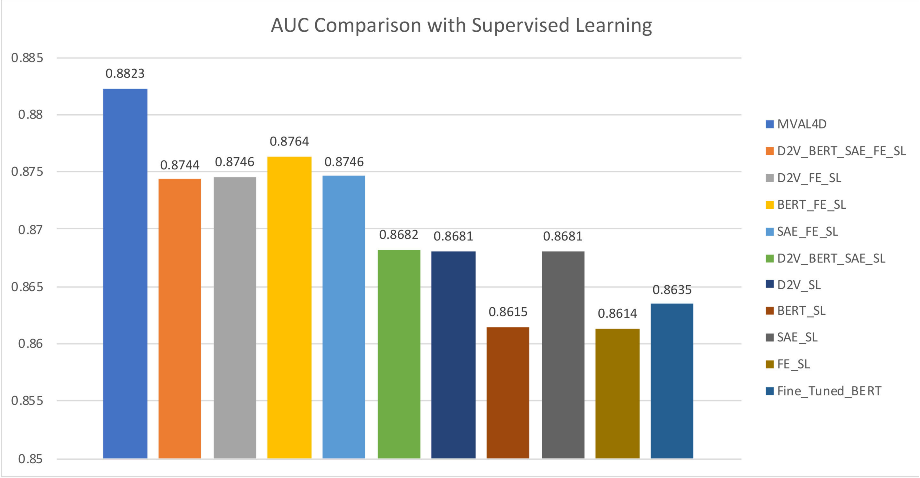


Figure 10. AUC comparison with supervised learning.

Additionally, we explored the false negatives and false positives generated by MVAL4D. The common conditions that MVAL4D fails to cope with are as follows: (i) text that is too short to provide adequate information, for example, "#restlesslegs #quetiapine"; (ii) explanatory and objective descriptions of ADRs without presenting personal subjective emotions, for example, "Slept 11 h last night on seroquel"; (iii) negative feelings on other aspects of a drug, for example, "I am run out of vyvanse so fast" and "this lozenge taste like shit"; (iv) text describing drug indications and negated ADRs, for example, "Taken more paracetamol to dull the aches."

6. Conclusions

In this study, we developed a multi-view active learning approach for recognizing ADR-related text from social media using various document representations. We have addressed several challenges. The first one is regarding the significant cost and difficulty of annotation. The second challenge is the requirement to simultaneously use data-driven information and domain-specific information. Specifically, we proposed a view-generation mechanism for performing multi-view active learning, and regarded each document representation as a view. In addition, we developed a novel selection strategy that separately uses informativeness-oriented and confidence-oriented measures to choose potentially negative instances and potentially positive instances. The experimental results show that our approach can achieve the enhanced predictive capability as compared with baselines.

This study has both methodological and practical contributions. Its methodological contributions are twofold. First, the view-generation mechanism proposed in our study can be generalized to address other text classification tasks. Second, the proposed selection strategy can guide other text classification tasks faced with the data imbalance issue. With regard to practical implications, our experimental results have demonstrated that the MVAL4D approach can obtain improved performance over existing approaches, even with fewer labeled instances. The effectiveness and efficiency of the ADR-related text detector model offer a solid foundation for the ADR mention extraction task, and improve the feasibility of monitoring drug safety using social media data. This study can provide valuable support for patients, regulatory authorities, pharmaceutical companies, and other stakeholders.

In the future, we plan to assess the generalizability of our proposed approach by applying it to other tasks. Moreover, we would like to further verify the scalability of the method. For example, we plan to incorporate features obtained by other feature extractors (e.g., graph neural network), and adopt other selection strategies in active learning. Moreover, we also plan to consider mutual information conveyed between different document representations.

Author Contributions: Conceptualization, J.L. and C.Z.; methodology, J.L. and Y.W.; software, Y.W.; validation, J.L.; formal analysis, J.L. and Y.W.; investigation, J.L.; resources, J.L. and Y.W.; data curation, Y.W.; writing—original draft preparation, J.L.; writing—review and editing, L.H. and C.Z.; visualization, Y.W.; supervision, L.H. and S.Z.; project administration, J.L., S.Z. and C.Z.; funding acquisition, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the National Natural Science Foundation of China (nos. 71701142, 71971170, and 71971067), the China Postdoctoral Science Foundation (no. 2018M640346), and the Tianjin Philosophy and Social Science Planning Project (no. TJKS19XSX-015).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset supporting this article can be obtained from http://diego.asu.edu/Publications/ADRClassify.html (accessed on 1 March 2020). The data are available in this article and should be cited as "Sarker, A. and G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, 2015. 53: p. 196–207".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; Gonzalez, G. Utilizing social media data for pharmacovigilance: A review. *J. Biomed. Inform.* **2015**, *54*, 202–212. [CrossRef] [PubMed]
2. Zhang, Y.; Cui, S.; Gao, H. Adverse drug reaction detection on social media with deep linguistic features. *J. Biomed. Inform.* **2020**, *106*, 103437. [CrossRef] [PubMed]
3. Hazell, L.; Shakir, S.A. Under-reporting of adverse drug reactions. *Drug Saf.* **2006**, *29*, 385–396. [CrossRef] [PubMed]
4. Amante, D.J.; Hogan, T.P.; Pagoto, S.L.; English, T.M.; Lapane, K.L. Access to care and use of the Internet to search for health information: Results from the US National Health Interview Survey. *J. Med. Internet Res.* **2015**, *17*, e106. [CrossRef] [PubMed]
5. Freifeld, C.C.; Brownstein, J.S.; Menone, C.M.; Bao, W.; Filice, R.; Kass-Hout, T.; Dasgupta, N. Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter. *Drug Saf.* **2014**, *37*, 343–350. [CrossRef] [PubMed]
6. Wu, C.; Wu, F.; Liu, J.; Wu, S.; Huang, Y.; Xie, X. Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention. In Proceedings of the Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–1 November 2018.
7. Fan, B.; Fan, W.; Smith, C.; Garner, H. Adverse drug event detection and extraction from open data: A deep learning approach. *Inf. Process. Manag.* **2020**, *57*, 102131. [CrossRef]
8. Dai, H.; Wang, C. Classifying adverse drug reactions from imbalanced twitter data. *Int. J. Med. Inform.* **2019**, *129*, 122–132. [CrossRef]
9. Kim, D.; Seo, D.; Cho, S.; Kang, P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Inf. Sci.* **2019**, *477*, 15–29. [CrossRef]
10. Liu, J.; Huang, L.; Zhang, C. An Active Learning Approach for Identifying Adverse Drug Reaction-Related Text from Social Media Using Various Document Representations. In *International Conference on Web Information Systems and Applications*; Springer: Berlin/Heidelberg, Germany, 2021.
11. Henriksson, A.; Kvist, M.; Dalianis, H.; Duneld, M. Identifying Adverse Drug Event Information in Clinical Notes with Distributional Semantic Representations of Context. *J. Biomed. Inform.* **2015**, *57*, 333–349. [CrossRef]
12. Gurulingappa, H.; Mateen-Rajput, A.; Toldo, L. Extraction of potential adverse drug events from medical case reports. *J. Biomed. Semant.* **2012**, *3*, 15. [CrossRef]
13. Van Mulligen, E.M.; Fourrier-Reglat, A.; Gurwitz, D.; Molokhia, M.; Nieto, A.; Trifiro, G.; Kors, J.A.; Furlong, L.I. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *J. Biomed. Inform.* **2012**, *45*, 879–884. [CrossRef]
14. Sarker, A.; Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.* **2015**, *53*, 196–207. [CrossRef]
15. Sarker, A.; Belousov, M.; Friedrichs, J.; Hakala, K.; Kiritchenko, S.; Mehryary, F.; Han, S.; Tran, T.; Rios, A.; Kavuluru, R.; et al. Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1274–1283. [CrossRef]
16. Weissenbacher, D.; Sarker, A.; Paul, M.; Gonzalez, 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 & Shared Task, Brussels, Belgium, 31 October–1 November 2018.

17. Weissenbacher, D.; Sarker, A.; Magge, A.; Daughton, A.; O'Connor, K.; Paul, M.; Gonzalez, G. Overview of the Fourth Social Media Mining for Health (#SMM4H) Shared Task at ACL 2019. In Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, Florence, Italy, 2 August 2019.
18. Yang, M.; Kiang, M.; Shang, W. Filtering big data from social media–Building an early warning system for adverse drug reactions. *J. Biomed. Inform.* **2015**, *54*, 230–240. [CrossRef]
19. Nikfarjam, A.; Sarker, A.; O'Connor, K.; Ginn, R.; Gonzalez, G. Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 671–681. [CrossRef]
20. Cocos, A.; Fiks, A.G.; Masino, A.J. Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 813–821. [CrossRef]
21. Tang, B.; Hu, J.; Wang, X.; Chen, Q. Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF. *Wirel. Commun. Mob. Comput.* **2018**, 2379208. [CrossRef]
22. Liu, J.; Wang, G.; Chen, G. Identifying Adverse Drug Events from Social Media using an Improved Semi-Supervised Method. *IEEE Intell. Syst.* **2019**, *34*, 66–74. [CrossRef]
23. Liu, X.; Chen, H. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. *J. Biomed. Inform.* **2015**, *58*, 268–279. [CrossRef]
24. Emadzadeh, E.; Sarker, A.; Nikfarjam, A.; Gonzalez, G. Hybrid Semantic Analysis for Mapping Adverse Drug Reaction Mentions in Tweets to Medical Terminology. *Am. Med. Inform. Assoc.* **2017**, *2017*, 679–688.
25. Chowdhury, S.; Zhang, C.; Yu, P.S. Multi-task pharmacovigilance mining from social media posts. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23 April 2018.
26. Patki, A.; Sarker, A.; Pimpalkhute, P.; Nikfarjam, A.; Ginn, R.; O'Connor, K.; Smith, K.; Gonzalez, G. Mining Adverse Drug Reaction Signals from Social Media: Going beyond Extraction. In *BioLink-SIG*; Oxford University Press: Oxford, UK, 2014.
27. Cai, J.J.; Tang, J.; Chen, Q.G.; Hu, Y.; Wang, X.; Huang, S.J. Multi-view active learning for video recommendation. In Proceedings of the IJCAI-19, Macao, China, 10–16 August 2019. Available online: https://www.ijcai.org/proceedings/2019/0284.pdf (accessed on 30 December 2019).
28. Yan, Y.; Nie, F.; Li, W.; Gao, C.; Yang, Y.; Xu, D. Image classification by cross-media active learning with privileged information. *IEEE Trans. Multimed.* **2016**, *18*, 2494–2502. [CrossRef]
29. Bhattacharjee, S.D.; Tolone, W.J.; Paranjape, V.S. Identifying malicious social media contents using multi-view context-aware active learning. *Future Gener. Comput. Syst.* **2019**, *100*, 365–379. [CrossRef]
30. Chen, L.; Fan, A.; Shi, H.; Chen, G. Search task success evaluation by exploiting multi-view active semi-supervised learning. *Inf. Process. Manag.* **2020**, *57*, 102180. [CrossRef]
31. Nigam, K.; Ghani, R. Analyzing the effectiveness and applicability of co-training. In Proceedings of the Ninth International Conference on Information and Knowledge Management, McLean, VA, USA, 6–11 November 2000.
32. Zhao, Y.; Shi, Z.; Zhang, J.; Chen, D.; Gu, L. A novel active learning framework for classification: Using weighted rank aggregation to achieve multiple query criteria. *Pattern Recognit.* **2019**, *93*, 581–602. [CrossRef]
33. Muslea, I.; Minton, S.; Knoblock, C.A. Active learning with multiple views. *J. Artif. Intell. Res.* **2006**, *27*, 203–233. [CrossRef]
34. Schwenker, F.; Trentin, E. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognit. Lett.* **2014**, *37*, 4–14. [CrossRef]
35. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1974**, *18*, 613–620. [CrossRef]
36. Li, Y.; Guo, H.; Zhang, Q.; Gu, M.; Yang, J. Imbalanced text sentiment classification using universal and domain-specific knowledge. *Knowl.-Based Syst.* **2018**, *160*, 1–15. [CrossRef]
37. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 31 May 2013.
38. Bengio, Y.; Ducharme, R.; Vincent, P. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
39. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
40. Lizarralde, I.; Mateos, C.; Zunino, A.; Majchrzak, T.A.; Grønli, T.-M. Discovering web services in social web service repositories using deep variational autoencoders. *Inf. Processing Manag.* **2020**, *57*, 102231. [CrossRef]
41. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, USA, 24 July 1998.
42. Windeatt, T.; Ardeshir, G. Decision Tree Simplification for Classifier Ensembles. *Int. J. Pattern Recognit. Artif. Intell.* **2004**, *18*, 749–776. [CrossRef]
43. Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*; Curran Associates: Vancouver, BC, Canada, 2007.
44. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. Proccedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

46. Zhou, Z.-H.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1529–1541. [CrossRef]
47. Li, M.; Zhou, Z.-H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *Syst. Man Cybern. Part A Syst. Hum. IEEE Trans.* **2007**, *37*, 1088–1098. [CrossRef]