Modeling Electronic Health Records in Ensembles of Semantic Spaces for Adverse Drug Event Detection

Aron Henriksson
Department of Computer and Systems Sciences (DSV)
Stockholm University
Stockholm, Sweden
Email: aronhen@dsv.su.se

Jing Zhao
Department of Computer and Systems Sciences (DSV)
Stockholm University
Stockholm, Sweden
Email: jingzhao@dsv.su.se

Henrik Boström
Department of Computer and Systems Sciences (DSV)
Stockholm University
Stockholm, Sweden
Email: henrik.bostrom@dsv.su.se

Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University
Stockholm, Sweden
Email: hercules@dsv.su.se

Abstract*—Adverse drug events (ADEs) are heavily underreported in electronic health records (EHRs). Alerting systems that are able to detect potential ADEs on the basis of patient-specific EHR data would help to mitigate this problem. To that end, the use of machine learning has proven to be both efficient and effective; however, challenges remain in representing the heterogeneous EHR data, which moreover tends to be high-dimensional and exceedingly sparse, in a manner conducive to learning high-performing predictive models. Prior work has shown that distributional semantics – that is, natural language processing methods that, traditionally, model the meaning of words in semantic (vector) space on the basis of co-occurrence information – can be exploited to create effective representations of sequential EHR data of various kinds. When modeling data in semantic space, an important design decision concerns the size of the context window around an object of interest, which governs the scope of co-occurrence information that is taken into account and affects the composition of the resulting semantic space. Here, we report on experiments conducted on 27 clinical datasets, demonstrating that performance can be significantly improved by modeling EHR data in ensembles of semantic spaces, consisting of multiple semantic spaces built with different context window sizes. A follow-up investigation is conducted to study the impact on predictive performance as increasingly more semantic spaces are included in the ensemble, demonstrating that accuracy tends to improve with the number of semantic spaces, albeit not monotonically so. Finally, a number of different strategies for combining the semantic spaces are explored, demonstrating the advantage of early (feature) fusion over late (classifier) fusion. Semantic space ensembles allow multiple views of (sparse) data to be captured (densely) and thereby enable improved performance to be obtained on the task of detecting ADEs in EHRs.

I. INTRODUCTION

Electronic health records (EHRs) have emerged as a potentially valuable source for pharmacovigilance, which, due to the limitations of clinical trials in terms of duration and sample size, needs to be carried out throughout the life-cycle of a drug to inform decisions about its continued use in the treatment of patients. There are, in fact, many examples of drugs being taken off the market for newly discovered side effects [1], [2]. Adverse drug events (ADEs) – defined as undesired harms resulting from the use of a drug – are also the most common form of iatrogenic injury, causing approximately 3.7% of hospital admissions worldwide [3], making it a major public health concern.

A challenge for pharmacovigilance is that ADEs are heavily underreported [4], both in spontaneous reporting systems – wherein ADE case reports submitted voluntarily by patients and clinicians are collected – and in EHRs, wherein ADEs can be encoded by a limited set of diagnosis codes. To address the underreporting problem, alerting systems that can automatically detect ADEs in EHRs are potentially very valuable, and much research has been conducted to that end. Unfortunately, ADE detection systems are often based on general rules that have been found to be inaccurate [5]. The use of machine learning for detecting ADEs on the basis of patient-specific EHR data has emerged as a promising alternative [6]–[10]. There are, however, challenges involved in applying machine learning to EHR data, such as high dimensionality (there are many types of clinical events) and sparsity (patients, particularly within a given healthcare episode, are only exposed to a small subset of those events), that make it difficult to learn high-performing predictive models. To address these, we have previously proposed [11] a means of representing healthcare episodes using distributional semantics – that is, models that try to capture the meaning of words (or sequential items) based on co-occurrence information – to create dense and low-dimensional (vector) representations of heterogeneous types of clinical data: notes, drug codes, diagnosis codes and measurements. These can subsequently be extracted from the so-called semantic spaces and aggregated to create representations of healthcare episodes that have the following advantages: (1) mitigate the problem of sparsity in clinical data, (2) model and explicitly take into account similarities between clinical events, and (3) allow large amounts of unlabeled clinical data to be leveraged. The representation was shown to be more effective for predictive modeling than representations based on frequency distributions over words and said clinical events.

To create the proposed representation of healthcare episodes, a semantic space with a given set of hyperparameters was chosen. An important hyperparameter concerns the definition of *context*, i.e., the region in which co-occurrences are considered, typically a window of surrounding words (or items); this has been shown to affect the semantic properties that are modeled [12]–[14]. Here, this hyperparameter is exploited to create *ensembles of semantic spaces*. That is, instead of modeling the data in a single semantic space with a given (context) window size, the same data is modeled in multiple

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

semantic spaces with different window sizes. The features generated by the different semantic spaces are intended to capture the data more holistically and thereby enable improved predictive performance to be obtained on the task of detecting ADEs in healthcare episodes. As the extracted features are here provided to an ensemble-based learning algorithm – random forest – improvements can be obtained by allowing for the creation of either better-performing or more diverse base classifiers, as these are the main factors that contribute to the success of ensemble models [15].

II. BACKGROUND

The proposed method exploits the context window size hyperparameter of distributional semantic models to obtain improved predictive performance of ensemble classifiers for ADE detection. Key concepts are described in this section.

A. Context Window in Distributional Semantic Models

Distributional semantics is a computational approach to modeling the meaning of natural language that is based on the observation – and captured in the distributional hypothesis [16] – that words with similar meanings tend to appear in similar contexts. Initially motivated by the inability of the bag-of-words representation of documents to account for synonymy [17], which reduced the recall (or sensitivity) of information retrieval systems, models of distributional semantics have primarily been used to create (semantic) vector representations of words. These have proven useful in a wide array of natural language processing tasks [18]. In recent years, distributional semantics has been leveraged also in the biomedical [19] and clinical [20] domains.

Although different distributional semantic models may have different hyperparemeters, the definition of context is one that is common and key to all. The choice of context affects the properties of the semantic space [12]. An important distinction exists, for instance, between *syntagmatic* and *paradigmatic* relations, and which one is modeled depends on the context definition that is employed. The former typically holds between words that co-occur (e.g., {car, engine, road}) and is characterized by the size of the context region, while the latter holds between words that do not themselves co-occur but share neighbors (e.g., synonyms like {car, automobile}). Context is usually defined as a (sliding) window that is symmetric around the focus word. The size of the context window has been shown to play an important role in contrasting different semantic relations [13], and the optimal window size tends to be task-dependent [14]. For the tasks of extracting medical synonyms from large corpora [21] and recognizing named entities in clinical text [22], [23], it has been shown that combining semantic spaces with different hyperparameters, including window size, can lead to improved performance.

B. Ensemble Models

An explanation for the effectiveness of ensemble models can be traced back to the works of Marquis de Condorcet already in the 18th century, in which he formulated a theorem, known as Condorcet's jury theorem [24], which states that the error of the majority of a jury decreases with the number of jury members. This theorem holds under the assumption that each member is more likely to be correct than wrong, but also requires that the members make the errors independently. The latter means, for example, that nothing is gained from forming a jury whose members always agree; the overall error will be no lower than the error of each single member. The scenario can be translated directly into the framework of ensemble learning [15], where each model in the ensemble corresponds to a jury member. Besides the number of models in the ensemble, there are hence two components that affect the predictive performance: the performance of each individual model and to what extent the models vary in their predictions. The latter is often referred to as the diversity of the ensemble [25]. In a regression framework, i.e., when the task is numerical prediction, the (squared) error E of the ensemble is directly related to the average (squared) error A of the ensemble members, and their diversity D, i.e., the average (squared) deviation of each single prediction from the ensemble prediction, as shown by the following equation [26]:

$$E = A - D$$

The above states that the ensemble error can be no higher than the average model error, and the more diversity, the lower the ensemble error. It should, however, be noted that using this directly in search of an optimal ensemble is not straightforward, as there is normally a strong interplay between diversity and average model performance, e.g., perfect models will agree on all predictions. When it comes to classification accuracy, there is unfortunately no similar decomposition of ensemble performance into average model accuracy and diversity. Instead, many alternative diversity measures have been proposed in the literature [25]; however, their connection to ensemble performance has been shown to be questionable.

III. METHODS AND MATERIALS

This paper investigates the use of multiple semantic spaces, built with different context window sizes, for the creation of representations of healthcare episodes. These are intended to be exploited by a supervised learning algorithm when creating predictive models for detecting the presence or absence of a certain ADE in a healthcare episode. The main research question posed in this study is as follows: rather than modeling EHR data in a single semantic space, built with a given context window size, is it possible to obtain improved predictive performance by exploiting the fact that the size of the context window affects the properties of the semantic space? Modeling EHR in ensembles of semantic spaces, created by employing various context window sizes, would allow for the creation of additional features with some degree of diversity.

The investigation is carried out using 27 clinical datasets, each comprising negative and positive examples, in the form of healthcare episodes, with respect to a particular ADE. A series of follow-up experiments are then conducted to explore the contribution of the generated features and the impact on predictive performance as increasingly more semantic spaces are used. Finally, several late (or classifier) fusion approaches, wherein the predictions or class probabilities of separate classifiers are combined, are explored and compared to the proposed early (or feature) fusion approach, wherein the generated feature sets are simply concatenated prior to learning.

TABLE I. DESCRIPTION OF DATASETS

Dataset	Code Description	Episodes	Words (Lemmas)		Diagnoses (ICD-10)		Drugs (ATC)		Measurements	
			Types	Instances	Types	Instances	Types	Instances	Types	Instances
D64.2	Secondary sideroblastic anemia due to drugs and toxins	416	46125	2110354	536	6320	364	8960	304	60689
E27.3	Drug-induced adrenocortical insufficiency	34	9564	112789	143	248	157	662	138	3982
F11.0	Mental and behavioural disorders (MBDs) due to use of opioids: acute intoxication	76	12200	232203	180	367	159	687	157	3920
F11.2	MBDs due to use of opioids: dependence syndrome	308	30077	904496	486	1875	347	4329	260	23637
F13.0	MBDs due to use of sedatives or hypnotics: acute intoxication	120	14764	215626	232	390	204	1167	153	6178
F13.2	MBDs due to use of sedatives or hypnotics: dependence syndrome	76	12507	215321	220	484	195	922	167	4621
F15.0	MBDs due to use of other stimulants, including caffeine: acute intoxication	32	5849	39658	71	148	96	257	105	1427
F15.1	MBDs due to use of other stimulants, including caffeine: harmful use	46	9174	102697	122	259	142	573	137	4518
F15.2	MBDs due to use of other stimulants, including caffeine: dependence syndrome	256	25179	658428	394	1347	295	3439	209	22870
F19.0	MBDs due to multiple drug use: acute intoxication	122	15823	278873	237	475	214	1120	227	5519
F19.1	Other psychoactive substance abuse	74	12651	177644	186	373	186	985	152	4688
F19.2	MBDs due to multiple drug use: dependence syndrome	288	29291	799717	492	1259	326	3667	262	19653
F19.9	MBDs due to multiple drug use: unspecified mental and behavioural disorder	68	13144	177749	177	350	178	992	87	3743
G24.0	Drug-induced dystonia	28	10017	101769	76	132	136	599	113	3551
G62.0	Drug-induced polyneuropathy	20	4622	35997	41	71	93	219	56	1119
I95.2	Hypotension due to drugs	70	11528	145432	162	652	177	799	144	5252
L27.0	Generalized skin eruption due to drugs and medicaments	274	34504	1114979	556	1619	375	5324	273	28451
L27.1	Localized skin eruption due to drugs and medicaments	78	13477	234268	220	545	186	1260	128	6088
N14.1	Nephropathy induced by other drugs	28	9180	82075	105	387	128	335	99	2215
O35.5	Maternal care for (suspected) damage to fetus by drugs	128	10567	121849	278	882	223	1654	125	3894
T59.9	Toxic effect of unspecified gases, fumes and vapors	40	5803	47694	81	165	104	317	76	1467
T78.2	Adverse effects: anaphylactic shock, unspecified	102	13341	188250	208	602	200	1063	200	5384
T78.3	Adverse effects: angioneurotic oedema	266	22659	411014	393	1178	282	2454	208	9967
T78.4	Adverse effects: allergy, unspecified	1520	46575	1633049	926	4571	463	9567	370	39883
T80.8	Other complications following infusion, transfusion and therapeutic injection	732	39077	1655988	709	5323	425	9890	269	35283
T88.6	Anaphylactic shock due to correct drug or medicament properly administered	96	15137	227317	240	549	209	1290	185	6325
T88.7	Unspecified adverse effect of drug or medicament	564	42794	1436333	767	3303	467	7263	306	41793

A. Data Source

The 27 datasets were extracted from the Stockholm EPR Corpus [27], which contains around 700,000 patients' health records[1] over a two-year period (2009-2010) from Karolinska University Hospital in Stockholm, Sweden. The learning task is to detect healthcare episodes that involve a certain ADE, i.e., in which an ADE-specific ICD-10 diagnosis code has been assigned. As both inpatient and outpatients are included in this study, a healthcare episode is not necessarily defined according to admission and discharge; instead, it is defined based on the time interval between recorded activities for a patient. Here, a healthcare episode is delimited by at least three days of no registered activities. The healthcare episodes are described by four types of data: clinical notes, ICD-10 diagnosis codes, ATC drug codes and clinical measurements (here, represented as *types* of measurements, i.e., values are ignored). Only healthcare episodes that contained at least one of each of the four data types were retained. Each of the 27 datasets thus consists of healthcare episodes according to the above definition, where the positive examples have been assigned an ADE-related diagnosis code, i.e., have experienced a drug-induced disorder (e.g., *G24.0: Drug-induced dystonia*), and the negative examples are an equal number of randomly selected healthcare episodes from the EHR database in which that same code has not been assigned. The ADE-related diagnoses were selected on the basis of having been classified as indicating ADEs in a previous study [28] and being sufficiently frequent ($>$ 10 healthcare episodes) in the EHR database. The number of visits and characteristics of the datasets are described in Table I. In addition to the labeled datasets, the entire two years of data is used for building the semantic spaces. The notes are preprocessed by using Stagger [29] for tokenization and lemmatization of Swedish text and by removing all digits and punctuation. The notes contain approximately 3 M unique words (700 M instances), while there are 9046 diagnosis codes (51.6 M instances), 1272 drug codes (2.9 M instances) and 713 measurements (14.5 M instances).

[1]This research has been approved by the Regional Ethical Review Board in Stockholm, permission number 2012/834-31/5.

B. Modeling Clinical Data in Semantic Space

To create representations of healthcare episodes, the data first needs to be presented as a sequence. For each of the three structured data types, we extract all sequences of events that occur in the healthcare episodes of patients, ordered by time. These sequences are then processed one-by-one by the distributional semantics algorithm. For notes, we obtain sequences of words. The preprocessed notes – lemmatized, without digits and punctuation – are processed sentence-by-sentence.

word2vec is used to build the semantic spaces over the sequential data. This implements a recently developed model that has been inspired by research in deep learning and neural network-based language models [30]. It was chosen for its ability to produce high-quality vector representations of words, outperforming traditional context-counting based methods on a range of natural language processing tasks [31] and now considered state-of-the-art in distributional semantics. We employ the skip-gram architecture, which is better than the CBOW[2] alternative at capturing infrequent words (or sequential items). The algorithm constructs a vocabulary from the training data and learns vector representations of the words (or sequential items). It achieves this by training a neural network with a single hidden layer; given a set D of sequential items i and their contexts c, the objective function is to set the parameters Θ that maximize $p(c|i;\Theta)$ [32]:

$$\arg\max_{\Theta} \prod_{(i,c)\in D} p(c|i;\Theta)$$

Context is defined as an adjacent item within a (symmetric) window of a pre-specified size around the input item. The parameters that are learned in the hidden layer give us the semantic vectors. This distributional semantic model hence

[2]Continuous Bag of Words

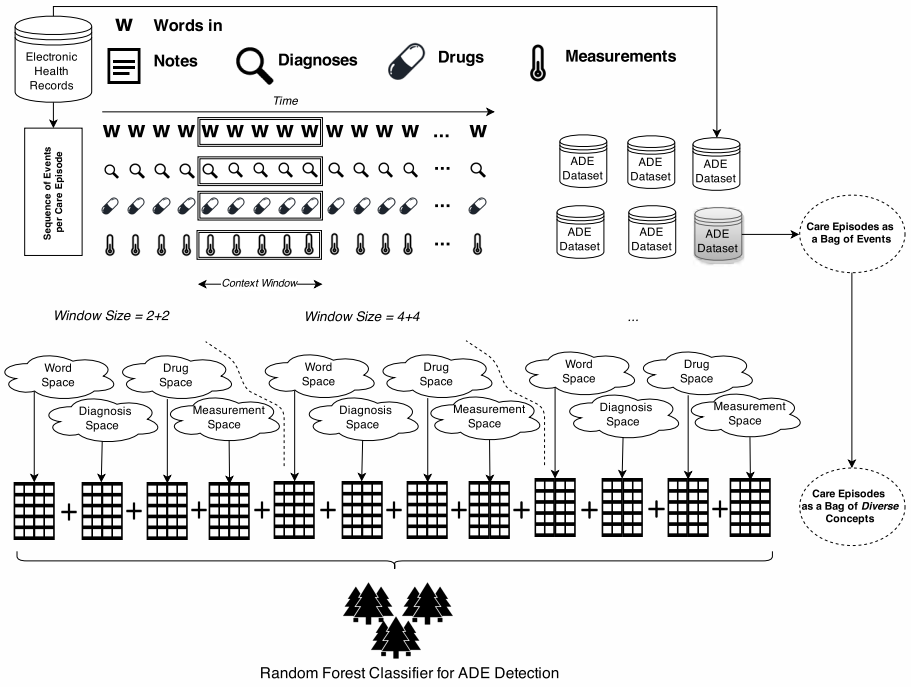

Fig. 1. Using ensembles of semantic spaces, built with different context window sizes, to model heterogeneous types of clinical data in semantic space and thereby create more holistic representations of healthcare episodes for adverse drug event detection

uses a supervised learning algorithm to learn semantic representations in a fully unsupervised way.

A semantic space is then created for each pre-specified context window size and set of input sequences. There is a set of input sequences for each data type: words, drug codes, diagnosis codes[3] and measurements. The semantic spaces are then used to create feature representations of healthcare episodes that are provided to the learning algorithm. This is achieved by simply summing the semantic vectors of the items in each healthcare episode, which is done separately for each semantic space and data type; these representations are then concatenated (Figure 1).

C. Experiments

A number of experiments are conducted to study the possibility of obtaining enhanced predictive performance on the task of detecting ADEs in healthcare episodes by utilizing ensembles of semantic spaces. These are described below.

Experiment 1: Using Ensembles of Semantic Spaces

Instead of modeling the data in a single semantic space, with a given set of hyperparameters, one may use a larger set of semantic spaces with different hyperparameters to create additional, diverse features. The features that are generated from each semantic space (and data type) are simply concatenated in the early fusion fashion and provided to the learning algorithm. In this first experiment, two feature sets are compared: (1) using a single semantic space, built with a symmetric context window of 12 items to the left and right of the focus item – the window size that, when used in isolation, yielded the best results – and (2) using an ensemble of semantic spaces, where the constituent semantic spaces are built with nine different window sizes: 2, 4, 6, 8, 10, 12, 14, 16, 18.

Experiment 2: Impact of Window Size on Variable Importance

To analyze the impact of using different window sizes when constructing the semantic spaces on the importance of the generated features, a variable importance analysis is conducted. Variable importance can be calculated for a given random forest model. Here, Gini importance is used, which is defined as the total decrease in node impurity (weighted by the probability of reaching that node), averaged over all the trees in the ensemble [33]. The variable importance scores are added up and averaged for each window size and ranked for each dataset.

Experiment 3: Inspecting Ensemble Classifiers

To investigate further the differences between the two sets of (random forest) classifiers from the first experiment – one with access to features generated by a single semantic space and one generated by an ensemble of semantic spaces – we calculate the average tree performance and compare it to ensemble performance. This provides some insight into the ensemble classifiers, in terms of tree quality and diversity. Then, to assess if the various semantic spaces contribute with diversity, we generate one classifier per semantic space and quantify their pairwise diversity using the Q-statistic [25], defined as:

$$Q = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}}$$

where, for a pair of classifiers, N_{00} denotes the number

[3]For diagnosis codes, 27 variants are created wherein the target ADE label code is excluded to avoid bias.

of cases where both are incorrect, N_{11} the number of cases where both are correct, N_{01} the number of cases where the first is incorrect but not the other, and finally N_{10} where the first is correct and the second incorrect. Q is between -1 and 1: classifiers that tend to recognize the same objects correctly will have positive values of Q, and those that commit errors on different objects – indicating diversity – will render Q negative.

Experiment 4: Adding Semantic Spaces to Ensemble

To study the impact on predictive performance as more semantic spaces are employed in the ensemble, we successively add semantic spaces according to the following strategies: *narrow to wide*, wherein semantic spaces built with increasingly wide window sizes are added from 2, 4, ..., 20, and *wide to narrow*, wherein semantic spaces built with increasingly narrow window sizes are added from 20, 18, ..., 2.

Experiment 5: Adopting Late Fusion Strategies

In the first experiment, the features from the semantic space ensemble were concatenated in an early fusion fashion. Here, we explore the following five late fusion strategies [34], where the constituents are random forest models. This experiment is designed like the previous experiment, where increasingly more semantic spaces are employed to generate the features.

- *S1 – Majority voting*: Classifies an instance based on the class that obtains the most votes by the constituent models. Mathematically, $class(x) = \arg\max_{c_i \in dom(y)} \sum_k g(y_k(x), c_i)$, where $y_k(x)$ is the prediction of the kth random forest and $g(y, c)$ is an indication function, defined as $g(y, c) = 1$ when $y = c$ or 0 when $y \neq c$.
- *S2 – Mean probability*: Classifies an instance according to the averaged probabilities that are obtained from the constituent models. Mathematically, $class(x) = \arg\max_{c_i \in dom(y)} (\sum_k \hat{P}_{M_k}(y = c_i|x)/K)$, where $\hat{P}_{M_k}(y = c|x)$ denotes the probability of class c given an instance x and K is the total number of constituent models.
- *S3 – Maximum probability*: Classifies an instance according to the constituent model that is most certain about its prediction, i.e., has the largest difference between the probabilities of two classes. Mathematically, $class(x) = \arg\max_{k \in 1...K} \left| \hat{P}_{M_k}(y = c_i|x) - \hat{P}_{M_k}(y = c_j|x) \right|$.
- *S4 – Weighted probability based on OOB accuracy*: Assigns a proportional weight to each constituent model according to its out-of-bag accuracy, and then multiplies the corresponding weight to each model's predicted class probabilities. Mathematically, $class(x) = \arg\max_{c_i \in dom(y)} \sum_k \big(w_k \times \hat{P}_{M_k}(y = c_i|x)\big)$, where w_k is the weight assigned to the kth constituent model, $\hat{P}_{oob_k}$, defined as $w_k = \hat{P}_{oob_k} / \sum_k \hat{P}_{oob_k}$.
- *S5 – Weighted predictions based on OOB accuracy*: Assigns a proportional weight to each constituent model according to its out-of-bag accuracy, then weights the votes and classifies an instance according to the class with the highest weighted votes. Mathematically, $class(x) = \arg\max_{c_i \in dom(y)} \sum_k \big(w_k \times g(y_k(x), c_i)\big)$.

D. Experimental Setup

In all of the experiments, the random forest algorithm [35] is used to generate predictive models. The choice was made for its reputation of achieving high predictive performance, its ability to handle high-dimensional data, as well as the possibility of obtaining estimates of variable importance. The algorithm constructs an ensemble of decision trees, where each tree is built from a bootstrap replicate of the original instances, while a subset of all features is sampled at each node when building the tree – in both cases to increase diversity. The decision trees together vote for what class label to assign to an example; when the number of trees in the forest increases, the probability that a majority of trees makes an error decreases, given that the they perform better than random and that the errors are made independently. In this study, we use random forest with 500 trees and $\sqrt{n}$ features inspected at each node. In all experiments, models are evaluated using stratified 10-fold cross validation. The considered performance metrics are accuracy and area under the ROC curve (AUC). Accuracy corresponds to the percentage of correctly classified instances, while AUC estimates the probability that a model ranks a randomly chosen positive instance ahead of a negative one.

In this study, the Wilcoxon signed-rank test is employed when two competing models are compared. This test ranks the differences in performance of two models on each dataset, and compares the ranks of positive and negative differences. It was chosen for its robustness when comparing two classifiers [36]. When more than two competing models are compared, the Friedman test [36] is employed, where the null hypothesis is that the methods perform equally well.

IV. RESULTS

The results of the experiments are reported on in this section, in the same order as described in the *Experiments*.

A. Single versus Multiple Semantic Spaces

The main hypothesis providing motivation for this study – that by using multiple semantic spaces built with different window sizes (M) we can obtain enhanced performance compared to using only a single semantic space built with a single window size (S) – is tested in the first experiment. The results (Table II) on the 27 datasets confirm this, with the Wilcoxon signed-rank test rejecting the null hypothesis that the classifiers using the two feature sets perform equally well w.r.t. both accuracy and AUC ($p < 0.05$).

B. Variable Importance Analysis

The relative ranks, according to Gini importance, of features generated by semantic spaces of each used window size in the ensemble, averaged over datasets, are shown in Table III. The Friedman test rejects the null hypothesis that the different window sizes lead to the generation of equally important features, overall ($p < 0.001$) and specifically for

TABLE II. PREDICTIVE PERFORMANCE OF CLASSIFIERS GENERATED BY A SINGLE (S) AND MULTIPLE (M) SEMANTIC SPACES; NUMBERS IN BOLD INDICATE WINS ACCORDING TO UNROUNDED VALUES

		D64.2	E27.3	F11.0	F11.2	F13.0	F13.2	F15.0	F15.1	F15.2	F19.0	F19.1	F19.2	F19.9	G24.0	G62.0	I95.2	L27.0	L27.1	N14.1	O35.5	T59.9	T78.2	T78.3	T78.4	T80.8	T88.6	T88.7	Mean	p	
Accuracy	S	95.0	80.0	**92.9**	88.7	88.3	**89.2**	**92.5**	**90.0**	94.9	90.8	85.4	**90.2**	84.6	82.5	85.0	86.3	84.7	78.3	75.0	99.2	90.0	85.2	88.9	93.6	**94.7**	85.5	**83.3**	88.0	0.05	
	M	95.0	80.0	92.5	**89.3**	**90.8**	87.1	90.0	89.2	**95.4**	90.8	**88.8**	89.9	**85.8**	**85.0**	**90.0**	86.3	84.7	**79.6**	**77.5**	99.2	**92.5**	**87.2**	**89.3**	**93.8**	94.4	**86.5**	83.2	88.7	$>$	
AUC	S	0.97	**0.83**	**0.97**	0.95	0.95	0.93	0.88	**0.99**	0.98	0.96	0.95	**0.95**	0.95	**0.98**	0.90	0.93	0.91	0.79	0.82	1.00	1.00	0.92	0.94	0.98	0.98	0.91	0.89	0.93	0.05	
	M	**0.97**	0.82	0.96	**0.96**	**0.96**	**0.95**	0.88	0.98	**0.98**	0.96	**0.97**	0.95	**0.97**	0.95	0.90	**0.95**	**0.92**	**0.83**	**0.82**	1.00	1.00	0.92	**0.96**	**0.99**	**0.98**	**0.93**	**0.90**	0.94	$>$	

drug ($p < 0.05$) and text ($p < 0.001$) features. A narrow window size of 2 leads, on average, to the most useful features overall; however, for diagnoses and drugs, window sizes of 6 and 8, respectively, result in more useful features.

TABLE III. AVERAGE RANKS OF FEATURES GENERATED BY SEMANTIC SPACES WITH DIFFERENT CONTEXT WINDOW SIZES

Features	Context Window Size								p-value
	2	4	6	8	10	12	14	16	18	
Drugs	4.5	4.5	5.4	**4.2**	4.4	5.4	5.1	6.5	4.9	< 0.05
Measurements	**4.0**	5.1	5.6	5.3	5.2	5.6	5.1	5.0	4.0	0.1665
Diagnoses	5.2	5.0	**3.9**	5.9	5.7	5.1	5.3	4.0	4.8	0.0687
Notes	**2.3**	6.1	5.8	5.7	5.0	6.1	4.8	2.4	6.7	< 0.001
All	**2.5**	5.8	5.5	5.8	5.2	5.5	5.3	3.5	5.9	< 0.001

C. Ensemble Inspection

A closer inspection of the two classifiers that were compared in the first experiment reveals that, not only is ensemble performance higher when having access to features from multiple semantic spaces (M) compared to only a single semantic space (S), but also average tree performance (Table IV). The difference between ensemble performance and average tree performance is, however, larger for S than M.

TABLE IV. COMPARING AVERAGE TREE PERFORMANCE WITH ENSEMBLE PERFORMANCE

	Average Tree Performance		Ensemble Performance	
	Accuracy	AUC	Accuracy	AUC
S	72.74	0.531	87.95	0.932
M	74.11	0.551	88.65	0.939

When calculating pairwise diversity between random forest models built using a single semantic space, very high Q-statistic scores were observed, with all above 0.99, indicating that they tend to classify the same instances correctly and hence exhibit low inter-ensemble diversity.

D. Including More Semantic Spaces

The impact on predictive performance as increasingly more semantic spaces are used is depicted in Figure 2. The general trend is that performance increases as more semantic spaces are added; the increase is, however, not monotonic. Performance with the *wide to narrow* strategy is initially lower than with the *narrow to wide* strategy, yet evens out as soon as four or five semantic spaces have been added. The observed increase in performance beyond two semantic spaces is, for both strategies, smaller w.r.t. AUC in comparison to accuracy. The classifiers with multiple semantic spaces perform better than their best "constituent" model – that is, a classifier with access only to a single semantic space – in all but one case w.r.t. accuracy, and 12 out of 18 cases w.r.t. AUC.

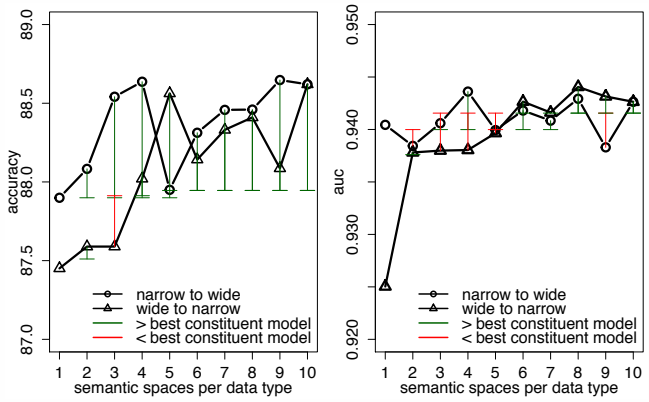

Fig. 2. Impact on predictive performance when adding features from increasingly more semantic spaces built with different context window sizes; the vertical lines indicate whether, and to what extent, the performance is better (green) or worse (red) than the best constituent (single-space) model

E. Early versus Late Fusion

When five late fusion strategies are compared to the early (feature) fusion approach, the predictive performance is almost invariably worse (Figure 3). The trend observed in the previous experiment, with higher performance with more semantic spaces, is not repeated with the *narrow to wide* strategy w.r.t. accuracy, with observed performance peaking after adding only three or four of the narrowest semantic spaces. With the exception of S3, there does seem to be such a trend w.r.t. AUC. With S1 and S4, AUC increases monotonically as more semantic spaces are added; however, the scores are much lower compared to the other strategies.

V. DISCUSSION

The hypothesis that improved predictive performance, in terms of accuracy and AUC, on the task of detecting ADEs in healthcare episodes could be obtained by exploiting multiple (M) semantic spaces, built with different window sizes, as opposed to a single (S) semantic space, was confirmed. Previous work has shown that using distributional semantics to model EHR data for ADE detection is more effective than common alternatives and facilitates the exploitation of heterogeneous data types that complement each other [11]; here, the predictive performance was improved further by modeling the data in ensembles of semantic spaces. This effectively

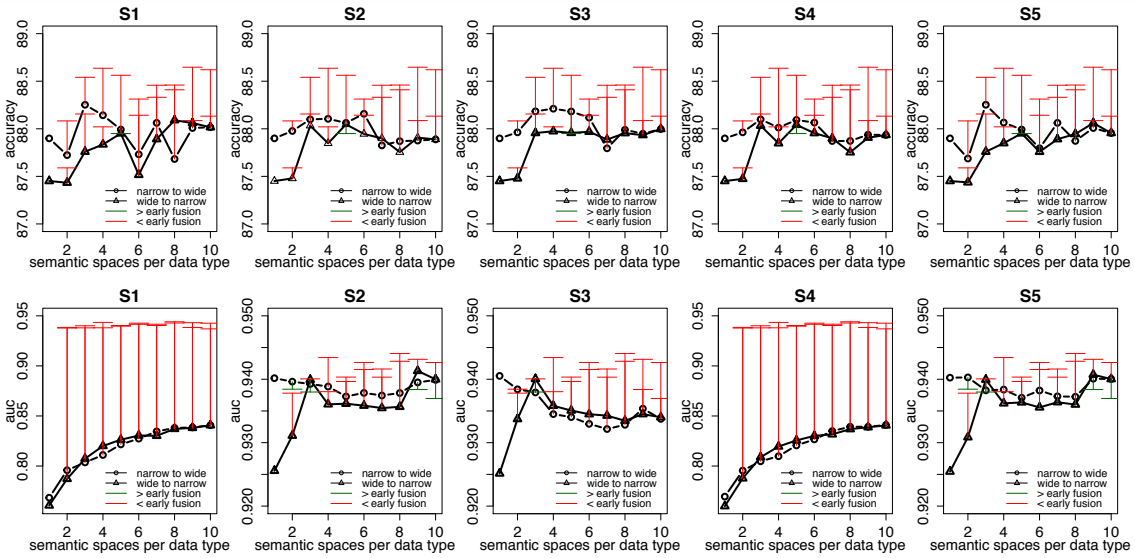

Fig. 3. Late fusion strategies, combining ensemble classifiers, compared to early fusion, fusing features from the semantic spaces built with different window sizes; the vertical lines indicate whether, and to what extent, the performance is better (green) or worse (red) than the corresponding early fusion model

makes better use of the large amounts of unlabeled data that indeed tend to be readily available. As mentioned in the background section, ensemble performance cannot readily be decomposed into constituent model performance and diversity, making it difficult to explain the improved performance. While the inspection of the two ensemble classifiers showed that M has both a higher average tree performance and a higher ensemble performance, it also revealed that the difference between ensemble performance and average tree performance was larger in S than in M, possibly indicating more diversity – and unexpectedly so – in the former. In any case, while it is difficult to attribute the improved performance of M to increased diversity, it can to a large extent be explained by the increase in average tree performance. Moreover, when generating features with increasingly more semantic spaces, performance generally improved, although with most of the gains obtained after utilizing only a few semantic spaces. The observed results, however, seem to indicate some sensitivity w.r.t. which semantic spaces are included in the ensemble.

One question concerned how best to utilize the multiple semantic spaces in the endeavor of obtaining improved predictive performance: should the generated features simply be concatenated in the early (feature) fusion fashion, or should they be exploited by separate random forest models and later be combined in the late (classifier) fusion approach? The results are unequivocal, with the early fusion strategy outperforming all investigated late fusion strategies. This is consistent with a study in which early and late fusion strategies were compared, coming out in favor of the former when using an ensemble of decision trees [37]. The calculation of inter-ensemble diversity provides some insight into the relative failure of the late fusion strategies, exhibiting extremely little diversity between the constituent models. This is partly also reflected in the average predictive performance scores of the constituent models, which are very similar (but not reported in the paper). It is thus clear that the window size in distributional semantic models is not merely a hyperparameter that needs to be tuned; however, when multiple semantic spaces with different sizes are utilized conjointly, and features are provided to an ensemble learning algorithm, a significant improvement can be observed. A possible partial explanation for the relative success of the early fusion approach vis-à-vis the late fusion strategies can be the former's ability to exploit variable interactions, i.e., when some combination of features from different window sizes interact.

When analyzing the importance of features generated by different semantic spaces, it was shown that narrow window sizes generally led to more useful features. This is somewhat surprising, both given that the recommended window size for word2vec's skip-gram model is 10 and that the best observed single semantic space classifier was obtained with a window size of 12. It was moreover found that window size has a significant impact on relative variable importance for text and drug features, but not for diagnoses and measurements, even if the p-values are all rather low. One potential problem that applies to all structured EHR data is that the timestamps are not always reliable and many events are given identical timestamps, which makes their ordering somewhat arbitrary.

Another limitation concerns the use of assigned diagnosis codes as class labels to train and evaluate predictive models, since it is well known that coding errors exist in EHR databases, resulting in noisy data. When diagnosis codes are used to indicate the presence or absence of ADEs, the problem is further exacerbated by the fact that ADEs are heavily underreported in EHRs, which means that it cannot be excluded that some of the negative examples should, in fact, be positive. Given that the task was cast as a binary classification problem, where a specific ADE represents the positive class, the risk of this happening is reasonably low. For access to higher-quality training data and a proper reference standard, the data needs to be verified by a domain expert.

VI. CONCLUSIONS

It was here demonstrated that improved predictive performance can be obtained on the task of detecting ADEs by creating representations of healthcare episodes with ensembles

of semantic spaces, built with different context window sizes. Representing the data in this manner addresses important challenges in using EHR data for predictive modeling, such as high dimensionality and sparsity, and also more holistically captures the deeper semantics of clinical data. Applying machine learning is a promising way of enabling meaningful (secondary) use of EHR data in the endeavor of improving health care, not least by supporting pharmacovigilance.

ACKNOWLEDGMENT

This work was supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection, ref. no. IIS11-0053.

REFERENCES

[1] B. Sibbald, "Rofecoxib (vioxx) voluntarily withdrawn from market," *Canadian Medical Association Journal*, vol. 171, no. 9, pp. 1027–1028, 2004.

[2] C. D. Furberg and B. Pitt, "Withdrawal of cerivastatin from the world market," *Curr Control Trials Cardiovasc Med*, vol. 2, no. 5, pp. 205–207, 2001.

[3] R. Howard, A. Avery, S. Slavenburg, S. Royal, G. Pipe, P. Lucassen, and M. Pirmohamed, "Which drugs cause preventable admissions to hospital? a systematic review," *British Journal of Clinical Pharmacology*, vol. 63, no. 2, pp. 136–147, 2007.

[4] L. Hazell and S. A. Shakir, "Under-reporting of adverse drug reactions," *Drug Safety*, vol. 29, no. 5, pp. 385–396, 2006.

[5] A. J. Forster, A. Jennings, C. Chow, C. Leeder, and C. van Walraven, "A systematic review to evaluate the accuracy of electronic adverse drug event detection," *JAMIA*, vol. 19, no. 1, pp. 31–38, 2012.

[6] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman, "Novel data-mining methodologies for adverse drug event discovery and analysis," *Clinical Pharmacology & Therapeutics*, vol. 91, no. 6, pp. 1010–1021, 2012.

[7] J. Zhao, A. Henriksson, and H. Boström, "Detecting adverse drug events using concept hierarchies of clinical codes," in *Proceedings of IEEE International Conference on Healthcare Informatics*. IEEE, 2014, pp. 285–293.

[8] J. Zhao, A. Henriksson, L. Asker, and H. Boström, "Detecting adverse drug events with multiple representations of clinical measurements," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2014, pp. 536–543.

[9] J. Zhao, A. Henriksson, and H. Boström, "Cascading adverse drug event detection in electronic health records," in *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2015.

[10] A. Henriksson, M. Kvist, H. Dalianis, and M. Duneld, "Identifying adverse drug event information in clinical notes with distributional semantic representations of context," *Journal of Biomedical Informatics*, vol. 57, pp. 333–349.

[11] A. Henriksson, J. Zhao, H. Boström, and H. Dalianis, "Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection," in *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2015.

[12] M. Sahlgren, "The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces," PhD Thesis, Stockholm University, 2006.

[13] G. Lapesa, S. Evert, and S. S. im Walde, "Contrasting syntagmatic and paradigmatic relations: Insights from distributional semantic models," in *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, 2014, pp. 160–170.

[14] G. Lapesa and S. Evert, "A large scale evaluation of distributional semantic models: parameters, interactions and model selection," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 531–545, 2014.

[15] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[16] Z. S. Harris, "Distributional structure." *Word*, 1954.

[17] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[18] P. D. Turney, P. Pantel *et al.*, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[19] T. Cohen and D. Widdows, "Empirical distributional semantics: methods and biomedical applications," *Journal of Biomedical Informatics*, vol. 42, no. 2, pp. 390–405, 2009.

[20] A. Henriksson, "Semantic spaces of clinical text: leveraging distributional semantics for natural language processing of electronic health records," Licentiate Thesis, Stockholm University, 2013.

[21] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravicius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," *Journal of Biomedical Semantics*, vol. 5, no. 6, pp. 1–25, 2014.

[22] A. Henriksson, H. Dalianis, and S. Kowalski, "Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2014, pp. 450–457.

[23] A. Henriksson, "Learning multiple distributed prototypes of semantic categories for named entity recognition," *International Journal of Data Mining and Bioinformatics*, vol. 13, no. 4, pp. 395–411, 2015.

[24] D. Austen-Smith and J. S. Banks, "Information aggregation, rationality, and the condorcet jury theorem," *American Political Science Review*, vol. 90, no. 01, pp. 34–45, 1996.

[25] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[26] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and cctive learning," in *Advances in Neural Information Processing Systems*. MIT Press, 1995, pp. 231–238.

[27] H. Dalianis, M. Hassel, A. Henriksson, and M. Skeppstedt, "Stockholm EPR Corpus: a clinical database used to improve health care," in *Swedish Language Technology Conference*, 2012.

[28] J. Stausberg and J. Hasford, "Drug-related admissions and hospital-acquired adverse drug events in germany: a longitudinal analysis from 2003 to 2007 of icd-10-coded routine data," *BMC Health Services Research*, vol. 11, no. 1, p. 134, 2011.

[29] R. Östling, "Stagger: an open-source part of speech tagger for swedish," *Northern European Journal of Language Technology (NEJLT)*, vol. 3, pp. 1–18, 2013.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Worshop*, 2013.

[31] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Association for Computational Linguistics*, vol. 1, 2014, pp. 238–247.

[32] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[33] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[34] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2010.

[35] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[36] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[37] H. Boström, "Feature vs. classifier fusion for predictive data mining – a case study in pesticide classification," in *International Conference on Information Fusion*, 2007, pp. 121–126.