JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 413-427 (2022)
DOI: 10.6688/JISE.202203_38(2).0008

Revisiting Supervised Word Embeddings

DIEU VU[1], KHANG TRUONG, KHANH NGUYEN,
NGO VAN LINH AND KHOAT THAN[+]
School of Information and Communication Technology
Hanoi University of Science and Technology
Hanoi 100000, Vietnam
[+]*E-mail: khoattq@soict.hust.edu.vn*
[1]*Faculty of Electrical and Electronic Engineering*
Phenikaa University
Hanoi 12116, Vietnam

Word embeddings are playing a crucial role in a variety of applications. However, most previous works focus on word embeddings which are either non-discriminative or hardly interpretable. In this work, we investigate a novel approach, referred to as *SWET*, which learns *supervised word embeddings using topic models* from labeled corpora. SWET inherits the interpretability of topic models, the discriminativeness of supervised inference from labels. More importantly, SWET enables us to directly exploit a large class of existing unsupervised and supervised topic models to learn supervised word embeddings. Extensive experiments show that SWET outperforms unsupervised approaches by a large margin, and are highly competitive with supervised baselines.

Keywords: supervised word embeddings, topic models, supervised learning, supervised topic models, word vectors

1. INTRODUCTION

Word embeddings refer to vector representations of words that can capture their meanings. Those vectors can be applied in diverse NLP tasks [1]. Recently, several studies present the effectiveness of word embeddings in a variety of applications [2], such as text classification, language modeling, named entity recognition, parsing, and tagging, *etc*. Word2Vec [3] and Glove [4] are two of the most well-known methods that can learn powerfully word embeddings from large-scale corpora.

While there is an enormous literature on unsupervised learning for word embeddings, there are few approaches on *supervised word embeddings (SWE)* that have capacity for encoding the supervision from labeled data. Unsupervised word embeddings are often non-discriminative and therefore, undesirable for supervised tasks [2, 5, 6]. A large class of embedding methods based on deep neural networks [3, 7] can model the local context of a word well. Remarkably, some recent approaches such as ELMO [8], BERT [9] take advantages of contextual information to learn word embeddings which are extremely rich in semantic knowledge. Nevertheless, those methods require large computational resource. In fact, they are difficult and complex to implement on devices of low capacity

Received April 11, 2020; revised June 20, 2020; accepted August 17, 2020.
Communicated by Meng Chang Chen.
[+]Corresponding author

[10]. Furthermore, those methods often produce continuous vectors that hardly support interpretability [11].

Interpretability of model is crucial in various practical applications [12, 13]. To obtain interpretable word embeddings, some methods use sparsity constraints [14, 15] and rotation techniques [16, 17]. Another work [18] exploits informative priors to create interpretable and domain-informed dimensions for probabilistic word embeddings. Recently, Word2Sense [19] extends topic models to refine the representation of a polysemous word in a short context. Besides, many studies proposed to combine the benefits of topic models [20] and deep neural networks [21–25]. Although those approaches can target interpretability, they ignore discriminativeness.

There are few efforts to develop methods that can learn discriminative embeddings. L-SVD [2] was proposed to encode labels into the co-occurrence matrix of terms and then used SVD or Glove to learn discriminative embeddings. Besides, another proposal [26] considered each word has many embeddings, each of which associates with a class label. Other works [6, 27] tried to fine-tune the universal embeddings for specific tasks and achieved promising results. Recently, LEAM [28] jointly learns label and word embeddings in the same latent space. This framework uses the text-label compatibility to learn an attentive model for text representation. All of those approaches succeed in capturing the supervision from labels, but lack interpretability.

In this work, we are interested in learning word embeddings which are both discriminative and interpretable. Moreover, the embeddings should be easy and light to train and test. Therefore, our contributions are as follows:

- We propose *SWET*, which can learn supervised word embeddings using topic models from labeled corpora. SWET inherits not only the interpretability of topic models but also the discriminativeness of supervised inference from labels. More importantly, SWET enables us to directly exploit a large class of existing unsupervised [29–31], supervised [32–41] topic models to learn supervised word embeddings. This property is really beneficial in practice.
- We provide a theoretical analysis which shows the rationale behind SWET.
- We did an extensive experiments to evaluate SWET and compare with various baselines. We find that SWET outperforms unsupervised approaches by a large margin, and are highly competitive with supervised state-of-the-art baselines.

The remainder of this paper is organized as follows: Section 2 presents some backgrounds. In Section 3, we present SWET, instantiate its application to some classes of topic models, and present the rationale behind SWET. The experiments and evaluation are presented in Section 4. Section 5 discusses future work and concludes the paper.

2. BACKGROUND

2.1 Topic Models

Consider a corpus D consisting of M documents and a vocabulary of V terms. A topic model assumes a corpus is composed from K topics β, and each topic $\beta_k = (\beta_{k1}, ..., \beta_{kV})$

Algorithm 1 : Two-phase SDR

Phase 1: Learn an unsupervised model to get K topics as an initialization: $\beta_1, \cdots, \beta_K$.
Phase 2: (finding discriminative space)

1. for each document **d** in class c, select a set N_d of its nearest neighbors in c.
2. infer new representation θ_d^* for each document d in class c by using the Frank-Wolfe algorithm [35] to maximize sum of log likelihood of document **d** and its neighbors in N_d
3. compute new topics as: $\beta_{kj}^* \propto \sum_{d \in D} d_j \theta_{dk}^*$.
   Finally, $\Omega_* = span\{\beta_1^*, ..., \beta_K^*\}$ is the discriminative space.

is a probability distribution on the vocabulary of K terms, meaning that $\sum_j \beta_{kj} = 1$ and $\beta_{kj} \geq 0$ for any k, j. Each document $d = (d_1, ..., d_V)$ (d_j is a count of term w_j in document d) is a mixture of those K topics. Each vector $\theta_d = (\theta_{d1}, ..., \theta_{dK})$ represents the topic proportion in document d, such that $\sum_k \theta_{dk} = 1$ and $\theta_{dk} \geq 0$ for any k. The target of learning a topic model is often to discover the hidden structures $(\beta, \theta_1, ..., \theta_D)$ from the given corpus. While β shows popular topics in the corpus, θ_d tells the importance/proportion of each topic in document d. *Probabilistic latent semantic analysis* [30] and *latent Dirichlet allocation* (LDA) [29] are popular topic models. When K needs to be pre-specified by users, those models are parametric.

2.2 Supervised Topic Models

Supervised topic models [32,34,36] aim to incorporate side information such as class labels into topic models. Label information is injected into the topical space and makes the space more discriminative for each class. Those models are effective for supervised tasks such as document classification or regression.

Supervised LDA (sLDA) [36] assigns each document to a response variable. Generative process for each document d of length N is described as below:

1. Draw topic proportion $\theta \sim \text{Dir}(\alpha)$.
2. For the n-th word in d:
   draw topic assignment $z_n \sim Mult(\theta)$, then draw word $w_n \sim Mult(\beta_{z_n})$
3. Draw class label $y \sim softmax(\frac{1}{N} \sum_{n=1}^{N} z_n, \eta)$

A difference between sLDA and LDA is y, which is an observed variable representing for the label of a document. Inference of sLDA also uses variational methods to approximate posterior distribution given a pair of document and label. Nevertheless, sLDA needs much memory for all parameters and exorbitant computations.

Another effective framework [35] called *supervised dimension reduction* (SDR) succeeds in incorporating labels into an unsupervised topic model to find a low-dimensional representation for documents. The framework is briefly described in Algorithm 1. The SDR framework learns a new space β^* encoding three features: label, document manifold and the semantic topics initialized in Phase 1. The label data and document manifold are utilized in Phase 2 to learn a low-dimensional topical space which is discriminative.

3. SUPERVISED WORD EMBEDDINGS WITH TOPIC MODELS (SWET)

In this section, we present SWET for learning supervised word embeddings, which are interpretable and discriminative. We will also explain why our approach is reasonable and discuss some of its key properties.

3.1 Method

SWET contains two steps:

- **Step 1:** Learn a supervised topic model to obtain topics β^* of size $K \times V$, where K is the number of topics and V is the vocabulary size.
- **Step 2:** Form the embedding of word j by taking the jth column of β^* and then normalizing it by a normalization method, such as $L1$, $L2$, $softmax$.

Next we discuss two different approaches to obtaining topics in Step 1.

3.1.1 Supervised approach

Various supervised topic models [32–41] can be used in Step 1 of SWET. Note that SWET can be applied in a variety of situations because the side information of documents may be categories, tags, ratings, *etc*. Word embeddings, learned in those cases, can be applied to classification problems or recommendation systems.

3.1.2 SDR-based approach

Supervised dimension reduction (SDR) [35] is the simple framework that boosts unsupervised topic models to work well with supervised tasks. It exploits the local structure of each class and the document manifold to learn a discriminative topical space.

This paragraph explains how SWET can encode the local structure of each class into word embeddings. SDR (Algorithm 1) attempts to learn a new representation θ^*_d for each document, which remains the structure of each class. Some dimensions of θ^*_d are promoted because θ^*_d captures the topic proportions of nearest neighbors. These promoted dimensions refer to the distinctive topics acknowledged as a local structure of class. Moreover, the representation of word w_j is calculated by normalizing the column vector $\beta^*_{(.)j}$, where $\beta^*_{(.)j} \propto \sum_{d \in D} d_j \theta^*_d$. If a large number of documents in D containing w_j belong to same class, several dimensions of $\beta^*_{(.)j}$ will be promoted. Hence, the representation $\beta^*_{(.)j}$ can capture local structure of a class. We recognize this structure as a global context which helps to regularize word embeddings. Intuitively, global context provides two significant advantages: the meaning of a word is understood more clearly, the representation is potentially discriminative. On the other hand, SDR is a flexible framework. One can use any unsupervised topic models in Phase 1 of SDR (Algorithm 1) and make supervised word embeddings.

3.2 Rationale of the Supervised Embeddings

In this section, we explain why we can obtain a word representation by taking columns of matrix β. We mention the relationship between word representation and topic

modeling. To see this aspect clearly, we need other views about learning topic models. The prevailing approaches use an approximation inference to find a maximum likelihood solution for given corpus. However, learning topic models can be viewed as non-negative matrix factorization [42]. The corpus can be represented by a $M \times V$ matrix document-by-term D (M documents, V terms and each element d_{ij} of matrix D is a count of terms w_j in document i). Learning a topic model is to find a topic matrix β with non-negative entries and a stochastically generated matrix θ such that $D_{[M,V]} \approx \theta_{[M,K]}\beta_{[K,V]}$. With the constraints on θ and β: $\begin{cases} \sum_{k=1}^{K} \theta_{dk} = 1 & \text{for each document } d. \\ \sum_{j=1}^{V} \beta_{kj} = 1 & \text{for each topic } k. \end{cases}$

Before going into the detail of word representation, we consider a solution to find two matrices θ and β. They are learned by minimizing a cost function that quantify the quality of the approximation between D and $\theta\beta$. Instead of using $L1$ or $L2$ distance to build the cost function, KL-divergence [43] can be applied. The KL-divergence between two matrices is calculated as below:

$$KL(D||\theta\beta) = \sum_{d \in M} \sum_{j=1}^{V} (d_j \log \frac{d_j}{\sum_{k=1}^{K} \theta_{dk}\beta_{kj}} - d_j + \sum_{k=1}^{K} \theta_{dk}\beta_{kj})$$

Note that $\sum_{j=1}^{V} \sum_{k=1}^{K} \theta_{dk}\beta_{kj} = 1$ due to (3.2). Minimizing KL-divergence is equivalent to:

$$\arg\max_{\theta,\beta} \sum_{d \in M} \sum_{j=1}^{V} (d_j \log \sum_{k=1}^{K} \theta_{dk}\beta_{kj})$$

Regarding to word embedding, we can consider topic modeling as a representation learning by re-writing the objective function as follows:

$$\min_{\theta,\beta} KL(D||\theta\beta) = \min_{\theta,\beta} \sum_{j} KL(D_{(.)j}||\theta\beta_{.j}) \tag{1}$$

where $D_{(.)j}$ is the jth column of matrix D and $\beta_{(.)j}$ is the j-th column of matrix β.

There are some reasons why our word embeddings are suitable. Firstly, two words that are highly co-occurrent have a higher similarity in our method. $D_{(.)j}$ is the statistic of term w_j in the whole corpus and is characteristic for w_j. While $\beta_{(.)j}$ is considered as the code or hidden representation of $D_{(.)j}$ as well as term w_j, θ can be regarded as a transformation matrix which maps the feature vector $D_{(.)j}$ to the code $\beta_{(.)j}$. If term w_i and term w_j have similar statistics *i.e.* $D_{(.)j}$ is close to $D_{(.)i}$ or (w_i, w_j) co-occurs in many contexts, their representations should be close to each other. It is clear that the vectors $\beta_{(.)j}$ and $\beta_{(.)i}$ will be close to each other when optimizing Eq. (1) because of the similarity between $D_{(.)i}$ and $D_{(.)j}$.

The second reason is the benefit of taking a column of β to obtain word embeddings. Each row of the matrix β is a topic. Hence, if we look into each column of β, we see how the meaning of the word is related to the topics. In other words, each column of β indicates the semantic information of a word. Moreover, the column vector $\beta_{(.)j}$ captures both local context (word co-occurrence patterns implicitly encoded by topic models) and global context (as mentioned in Section 3.1.2), therefore, this representation is reasonable.

3.3 The Properties of SWET

We next analyze some key properties of SWET.

3.3.1 Interpretability

Due to inheriting the advantages of topic models, SWET achieves an interpretable ability. It is easy to see that each dimension of the embedding space corresponds to a topic and hence is interpretable. When using L1 normalization, the embedding vector $\beta'_{(.)j} \propto \beta^*_{(.)j}$ of word j can be considered as a probability distribution over topics, and each element β'_{ij} in the embedding vector explains how word j is related to topic i. If β'_{ij} is high, the meaning of word j has a strong connection to topic i. Most existing methods for word embeddings do not have this property.

3.3.2 Discriminativeness

Discriminativeness could be understood as class-attention word embeddings. This property is especially significant for classification tasks. We obverse that if the meaning of word j is strongly connected to a class c (or j has a high contribution to class c) and is different to the remaining classes, its representation should be discriminative for that class. Each class c has its own local structure which is different from those of other classes, and this discriminative information is often encoded in θ^*_d when doing inference. Therefore, the discriminativeness in θ^*_d is inherently translated into the topics after learning, *e.g.*,

in SDR [35]: $\beta^*_{kj} \propto \sum_{d \in D} d_j \theta^*_{dk}$,

in FSLDA [33]: $\beta^*_{kj} \propto \sum_{d \in D} \sum_{n=1}^{N} 1(w_{d,n} = j) \phi^k_{d,n} \approx \sum_{d \in D} d_j \theta^*_{dk}$,

where $\phi_{d,n}$ is the variational multinomial parameter for the topic assignment $z_{d,n}$. As a consequence, the embedding vector $\beta'_{(.)j}$ for each word j is discriminative.

SWET enables us to estimate the contribution of each word to each class. Note that the probability of word j appearing in document d is: $p(w = j|d) = \sum_{k=1}^{K} \theta_{dk} \beta_{kj}$ Hence, the contribution of word j to class c can be approximated by:

$$p(w = j|c) \approx \frac{\sum_{d \in D_c} p(w=j|d)}{\sum_{r=1}^{V} \sum_{d \in D_c} p(w=r|d)} = \frac{\sum_{d \in D_c} \sum_{k=1}^{K} \theta_{dk} \beta_{kj}}{\sum_{r=1}^{V} \sum_{d \in D_c} \sum_{k=1}^{K} \theta_{dk} \beta_{kr}} \quad (2)$$

With each class c, one can select top words which contribute the most to the class. This is extremely significant to understand the classes and provides an excellent interpretation for a class.

4. EVALUATION

In this section, we investigate the main properties of SWET and compare with state-of-the-art baselines. We use seven benchmark datasets including 20NG, R52, R8, OH, and MR as in [44]; AGNews and DBpedia as in [28]. Some statistics of the datasets are described in Table 1.

Three different versions of SWET are used in our evaluation: *SWET-SDR* which uses SDR [35] to learn discriminative topics via dimension reduction; *SWET-FSLDA* which uses FSLDA [33] to learn a supervised topic model.

4.1 Analysis about Interpretability and Discriminativeness

We first want to verify the interpretability and discriminativeness. DBpedia is used in this evaluation.

Table 1. Some statistics of the datasets.

Dataset	#Training	#Test	#Classes
20NG	11314	7532	20
R8	5485	2189	8
R52	6532	2568	52
OH	3357	4043	23
MR	7108	3554	2
AGNews	120000	7600	4
DBPedia	560000	70000	14

Table 2. NPMI of word embeddings methods. Higher is better.

Model	R8	R52	OH	MR
word2vec	-0.82	-0.84	-0.88	-0.94
LEAM	-0.87	-0.88	-0.89	-0.93
SWET-SDR	-0.80	-0.82	-0.68	-0.69
SWET-FSLDA	-0.62	-0.61	-0.71	0.32

Baselines: Word2Vec [3] and LEAM[i] [28] are used in this evaluation. While Word2Vec is a representative for the unsupervised approach, LEAM is the state-of-the-art supervised method for word embeddings.

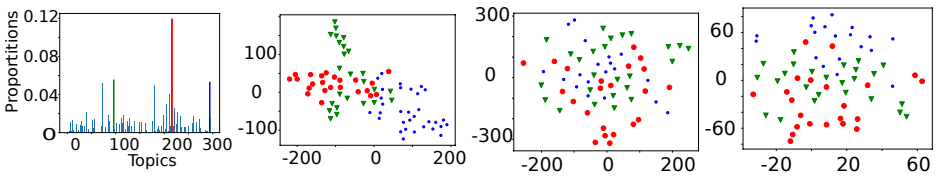
Settings: For SWET-SDR, we set $K = 300$ topics, other parameters are chosen as in [35]. For Word2Vec, we use Google's pre-trained model[ii] including word vectors for a vocabulary of 3 million words. The dimensionality of the embedding vectors is 300.

Interpretability: We want to quantitatively assess the interpretability of each dimension. We select top words of each dimension by getting a set of words having the highest value on the dimension. Then we compute NPMI [45] for selected top words. NPMI essentially measures the coherence of the words in a given set, and is often used for evaluation about interpretability. The results are shown in Table 2. It is obvious that SWET-based approaches have higher NPMI than others. It suggests that each dimension of the word embeddings learned by SWET is more interpretable than that by the other methods. This is not difficult to explain because the value of each dimension of SWET is meaningful and strongly relates to a topic.

We further perform an qualitative assessment by selecting a particular word and its learnt embedding, then we get top 3 dimensions/topics with the highest value of the embedding vector. We extract 30 words, which have the highest probability in each of the three topics, and visualize their embeddings by t-SNE [46]. Fig. 1 shows an example of the interpretability for representation of the word "advertisement". Fig. 1a shows the topic proportions of word "advertisement", in which the topics 216 (red), 96 (green), 294 (blue) are the highest ones. Table 3 presents some top words of the topics illustrating in Fig. 1. It is obvious that top characteristic words of a topic should be represented close to each other because all of them are related to the same topic. Word representation learned by SWET-SDR can capture label and manifold information, therefore, the separation among three topics is noticeable. However, Word2Vec or LEAM cannot separate the topics, it means that they are unable to interpret the word "advertisement" based on each dimension of the embedding.

Discriminativeness: With SWET, we can extract top characteristic words for each class by using Eq. (2). In Word2vec, we compute and choose words having the high frequency

[i] https://github.com/guoyinwang/LEAM
[ii] https://code.google.com/archive/p/word2vec/



(a) Topic proportions of word "advertisement"

(b) Word embedding learned by SWET-SDR

(c) Word embedding learned by Word2Vec

(d) Word embedding learned by LEAM

Fig. 1. Illustration about the interpretability for the representation of the word "advertisement"; (a) shows the contribution of each topic to word "advertisement", learned by SWET-SDR; (b) shows the embeddings of some words representing topics 216 (red), 96 (green), 294 (blue), which have highest contributions to "advertisement"; (c) and (d) visualize the embeddings of those words which are learned by Word2Vec and LEAM respectively.

Table 3. Illustration for the top characteristic words of some topics, learned by SWET-SDR from DBpeadia.

Topic	Top characteristic words
216	center district street states city united county places national building.
96	church built hospital river museum mall school places building house.
294	science academy university education private research students public international college.

in each class. For LEAM, we compute the similarity between word and class embeddings by the cosine measure. Then we choose the words having the highest similarity. We evaluate on DBPedia dataset. Table 4 presents top 10 words and Fig. 2 visualizes the embeddings in 2-dimensional space, using t-SNE [46]. We observe that top words by SWET-SDR robustly relate to the corresponding classes, while there are many "noisy" words extracted by Word2vec and LEAM. In SWET-SDR, the embeddings of the characteristic words belonging to the same class are close to each other. Meanwhile, observing the results of Word2Vec and LEAM, the words characterizing a class distribute chaotically. Moreover, it is difficult to see separation among classes from the embeddings by Word2Vec or LEAM.

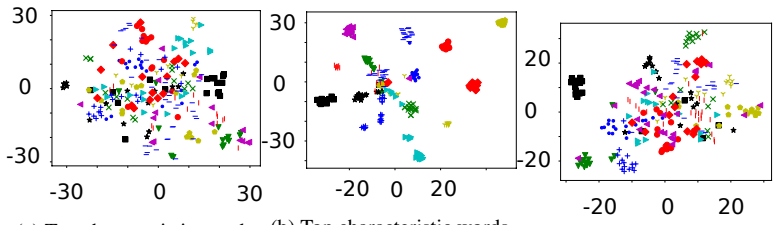
4.2 Document Classification Task

In this section, we evaluate SWET via classification task.

Baselines: Word2Vec [47] is a popular word embedding method. We use the Google's pre-trained Word2Vec[ii]. **fastText** [iii] [48] is the fast and simple method for text representation. Document embeddings were built by averaging word/n-grams embeddings. **SWEM** [iv] (Simple word embedding models) [49] uses pooling method operated over word embedding. **LEAM** (label embedding attentive models) [28] learns jointly word and label embeddings with a compatibility metric between words and labels. The classifier used is MLP layer with a sigmoid or softmax function. **LSVD** [2] is the method that can

[iii] https://github.com/facebookresearch/fastText/

[iv] https://github.com/dinghanshen/SWEM



(a) Top characteristic words of classes learned by Word2vec.

(b) Top characteristic words of classes learned by SWET-SDR.

(c) Top word embedding of classes learned by LEAM.

Fig. 2. Comparison of SWET-SDR, Word2Vec, and LEAM in term of the discriminativeness. The points having the same color and shape are the characteristic words in the same class.

Table 4. Illustration for the top characteristic words of each class, learned by three methods. The words in Italic seems not characterize the corresponding class. DBPedia is used in this evaluation.

Class name	Word2vec	SWET-SDR	LEAM
Artist	writer best music singer american *known born also end unk.*	writer songwriter artist musician singer author band producer composer english.	sculptor artist painter *nidaros* sculpture printmaker artistic artworks illustrator watercolor.
Building	register built building located house church historic *national end unk.*	*listed register* places museum street hospital mall hotel center style.	renovation plaza building buildings constructed *grosset sitting 91 richardsonian erected.*
Album	rock music records studio band released album *first end unk.*	album released studio music recorded live debut songs compilation release.	album albums soundtrack unreleased *montenegrin richardsonian mixtape ep tour disturbed.*

simultaneously use both local context of words and labels to learn word embeddings. **TextGCN**[v] [44] is the method which learn the representation of words and documents jointly on a graph. **BERT** [9] learns word embeddings by contextual information. With **BERT-last**, we use a pretrained model[vi] , then learn a MLP layer for classification and freeze all other layers. We also include the evaluation about the low-dimensional representation of documents which is learned by **SDR**[vii] [35].

For embedding methods (SWET, Word2Vec and LSVD), we represent a document by concatenating all of word embedding vectors of that document. Such a concatenation preserves the information of the embeddings. Then, LibLinear [50] is used to train SVM classifier, with regularization parameter $C = 0.1$.

Settings: We perform the classification task on 7 benchmark datasets: 5 medium-size datasets and 2 big datasets. The description of these datasets is reported in Table 1. To implement SWET, the parameters are chosen the same as in the original work of SDR and FSLDA. We only tune the parameter K (number of topics) in the training process. When changing K, it means that we change the size of the word embedding vector. For the baselines, we use the default settings as suggested in their original papers.

[v] https://github.com/yao8839836/text_gcn
[vi] https://github.com/google-research/bert
[vii] http://www.jaist.ac.jp/~s1060203/codes/sdr/

Table 5. Test accuracy on document classification task (%).

Model	20NG	R8	R52	Ohsumed	MR	AGNews	DBPedia
word2vec	79.9	96.89	91.12	60.72	75.75	90.61	97.97
fastText	79.38	96.13	92.81	57.7	75.14	92.50	98.60
SWEM	85.16	95.32	92.94	63.12	76.65	92.24	98.42
TextGCN	86.34	97.07	**93.56**	**68.36**	76.74	-	-
LSVD	81.25	96.21	87.97	51.42	71.53	88.93	96.58
LEAM	81.91	93.31	91.84	58.58	76.95	92.45	**99.02**
SDR	80.16	93.92	88.28	56.44	64.24	86.65	94.03
BERT-last	67.90	96.02	89.66	51.17	**79.24**	-	-
SWET-SDR	**86.34**	**97.12**	92.75	66.93	76.45	**92.64**	98.11
SWET-FSLDA	84.88	97.08	93.00	67.30	75.30	91.52	-

Table 6. Accuracy on classification task (%) when using different methods for normalization in SWET. *c* (*a*) **indicates that the methods use concatenation (averaging) of word vectors to represent a document.**

Method	$No-norm^c$	$L2^c$	Max^c	$Softmax^c$	$L1^c$	$L1^a$
MR	57.71	75.66	76.45	66.51	76.45	72.36
R8	85.56	96.16	96.35	95.93	97.12	84.46
R52	73.97	90.19	91.55	86.29	92.75	73.71

Results and Analysis: Table 5 shows the test accuracy of each model on all datasets. SWET-based methods surpass the unsupervised approaches (Word2Vec, SWEM, BERT-last) in most cases. Moreover, SWET-SDR outperforms all of the baselines on 20NG, R8 and AGNews. This can be explained by that embeddings of SWET-SDR are strongly related to topics and discriminative by classes while the classes of documents in 20NG, R8, and AgNews datasets are document topics. Therefore, classification on the three datasets with SWET gains a good result.

TextGCN achieves the highest accuracy on R52 and Ohsumed, and seems to perform best amongst the baselines. It is worth noting that the accuracy gap between SWET-SDR and TextGCN is extremely tight. On the other hand, BERT-last is a representation of contextual approaches. Therefore, it is easy to understand that BERT-last is the best result on a sentiment dataset - MR.

In conclusion, it is undeniable that the deep learning methods have better performance than the others on large datasets while TextGCN and SWET perform better on small datasets. Overall, SWET-based methods are usually one of the top methods that achieve the best results on all datasets.

Sensitivity analysis: We next assess the influence of normalization and document representation approaches as well as the sensitivity of the number of topics in SWET. We find that Step 2 of SWET plays a key role in document classification as evidenced in Table 6. L1-normalization is suitable for SWET because it shows the best accuracy while no normalization may result in a bad accuracy. We also find that concatenation seems to be better than averaging when representing documents from word embeddings.

The number of topics is the dimensionality of word embeddings and thus affects the performance of SWET on classification task. In Fig. 3, we show the test accuracy

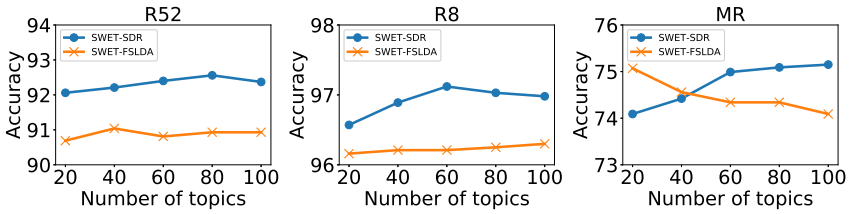


Fig. 3. Accuracy of SWET-SDR and SWET-FSLDA as the number of topics increases.

of SWET-SDR and SWET-FSLDA on three datasets R52, R8 and MR. We observe that SWET-SDR has a higher accuracy than SWET-FSLDA as the number of topic increases. This seems to be due to the fact that SWET-SDR captures manifold information between classes from SDR that makes more discriminativeness for the document representation.

5. CONCLUSION

We investigated supervised word embedding approaches learned by supervised topic models (SWET). SWET can capture the discriminativeness and interpretability. The extensive experiments show that SWET is highly competitive with existing approaches. Nevertheless, SWET still has some drawbacks, for example, the weakness of preserving local context or word order in a text.

ACKNOWLEDGMENT

This research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA18, and by the Office of Naval Research Global (ONRG) under Award Number N62909-18-1-2072, and Air Force Office of Scientific Research (AFOSR), Asian Office of Aerospace Research & Development (AOARD) under Award Number 17IOA031.

REFERENCES

1. J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of ACL*, 2010, pp. 384-394.
2. L. Yang, X. Chen, Z. Liu, and M. Sun, "Improving word representations with document labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, 2017, pp. 863-870.
3. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, 2013, pp. 3111-3119.
4. J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532-1543.

5. D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment embeddings with applications to sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, 2015, pp. 496-509.
6. X. Yang, K. Mao, X. Yang, and K. Mao, "Task independent fine tuning for word embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, 2017, pp, 885-894.
7. Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1137-1155.
8. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of Conference of NAACL: Human Language Technologies*, 2018, No. arXiv:1802.05365.
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of Conference of NAACL: Human Language Technologies*, 2019, No. arXiv:1810.04805.
10. Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a compact task-agnostic BERT for resource-limited devices," in *Proceedings of the 58th Annual Meeting of ACL*, 2020, pp. 2158-2170.
11. L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur, "Semantic structure and interpretability of word embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, pp. 1769-1779.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' explaining the predictions of any classifier," in *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
13. A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD Explorations*, Vol. 15, 2014, pp. 1-10.
14. B. Murphy, P. Talukdar, and T. Mitchell, "Learning effective and interpretable semantic models using non-negative sparse embedding," in *Proceedings of the 24th International Conference on Computational Linguistics*, 2012, pp. 1933-1950.
15. F. Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Sparse word embeddings using l1 regularized online learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016, pp. 2915-2921.
16. S. Park, J. Bak, and A. Oh, "Rotated word vector representations and their interpretability," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 401-411.
17. S. Rothe and H. Schütze, "Word embedding calculus in meaningful ultradense subspaces," in *Proceedings of the 54th Annual Meeting of the ACL*, 2016, pp. 512-517.
18. M. H. Bodell, M. Arvidsson, and M. Magnusson, "Interpretable word embeddings via informative priors," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 6323-6329.
19. A. Panigrahi, H. V. Simhadri, and C. Bhattacharyya, "Word2sense: sparse interpretable word embeddings," in *Proceedings of the 57th Annual Meeting of ACL*, 2019, pp. 5692-5705.
20. D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, Vol. 55, 2012, pp. 77-84.
21. Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical word embeddings," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015, pp. 2418-2424.

22. L.-P. Liu and D. M. Blei, "Zero-inflated exponential family embeddings," in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, 2017, pp. 2140-2148.
23. M. Rudolph and D. Blei, "Dynamic embeddings for language evolution," in *Proceedings of International World Wide Web Conference*, 2018, pp. 1003-1011.
24. R. Das, M. Zaheer, and C. Dyer, "Gaussian lda for topic models with word embeddings," in *Proceedings of the 53rd Annual Meeting of ACL and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 795-804.
25. K. Batmanghelich, A. Saeedi, K. Narasimhan, and S. Gershman, "Nonparametric spherical topic modeling with word embeddings," in *Proceedings of the 54th Annual Meeting of Association for Computational Linguistics*, 2016, pp. 537-542.
26. S. Kuang and B. D. Davison, "Class-specific word embedding through linear compositionality," in *Proceedings of IEEE International Conference on Big Data and Smart Computing*, 2018, pp. 390-397.
27. J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, 2018, pp. 328-339.
28. G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," in *Proceedings of the 56th Annual Meeting of the ACL*, 2018, No. arXiv:1805.04174.
29. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
30. T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, Vol. 42, 2001, pp. 177-196.
31. K. Than and T. B. Ho, "Fully sparse topic models," in *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, pp. 490-505.
32. J. Zhu, A. Ahmed, and E. P. Xing, "Medlda: maximum margin supervised topic models," *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 2237-2278.
33. A. Katharopoulos, D. Paschalidou, C. Diou, and A. Delopoulos, "Fast supervised lda for discovering micro-events in large-scale video datasets," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 332-336.
34. S. Lacoste-Julien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, 2009, pp. 897-904.
35. K. Than, T. B. Ho, and D. K. Nguyen, "An effective framework for supervised dimension reduction," *Neurocomputing*, Vol. 139, 2014, pp. 397-407.
36. J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Proceedings of the 20th International Conference on Neural Information Processing SystemsDecember*, 2008, pp. 121-128.
37. N. Van Linh, N. K. Anh, K. Than, and C. N. Dang, "An effective and interpretable method for document classification," *Knowledge and Information Systems*, Vol. 50, 2017, pp. 763-793.
38. A. M. Dai and A. J. Storkey, "The supervised hierarchical dirichlet process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, 2014, pp. 243-255.

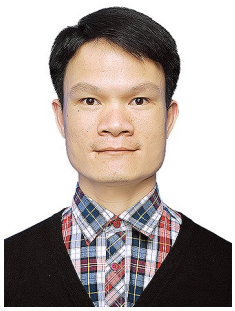
39. D. Kim, S. Kim, and A. Oh, "Dirichlet process with mixed random measures: a nonparametric topic model for labeled data," in *Proceedings of the 29 th International Conference on Machine Learning*, 2012, pp. 675-682.
40. J. Xuan, J. Lu, G. Zhang, R. Y. D. Xu, and X. Luo, "A bayesian nonparametric model for multi-label learning," *Machine Learning*, Vol. 106, 2017, pp. 254-269.
41. J. Xuan, J. Lu, and G. Zhang, "Cooperative hierarchical dirichlet processes: Superposition vs. maximization," *Artificial Intelligence*, Vol. 271, 2019, pp. 43-73.
42. S. Arora, R. Ge, and A. Moitra, "Learning topic models-going beyond svd," in *Proceedings of IEEE 53rd Annual Symposium on Foundations of Computer Science*, 2012, pp. 1-10.
43. D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556-562.
44. L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 7370-7377.
45. G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proceedings of Conference of German Society for Computational Linguistics and Language Technology*, 2009, pp. 31-40.
46. L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, Vol. 9, 2008, pp. 2579-2605.
47. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013, pp. 3111-3119.
48. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of European Chapter of the Association for Computational Linguistics*, Vol. 2, 2017, pp. 427-431.
49. D. Shen, G. Wang, W. Wang, M. Renqiang Min, Q. Su, Y. Zhang, C. Li, R. Henao, and L. Carin, "Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms," in *Proceedings of the 56th Annual Meeting of the ACL*, 2018, No. arXiv:1805.09843.
50. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, Vol. 9, 2008, pp. 1871-1874.

Dieu Vu is currently a graduate student at Hanoi University of Science and Technology (HUST). He received the excellent BS degree from HUST in 2018. His interest is machine learning.

Khang Truong is currently an MS student at Korea Advanced Institute of Science and Technology (KAIST). He received B.S. degree in 2018 from Hanoi University of Science and Technology. His interest includes topic models.

Khanh Nguyen was a student at Hanoi University of Science and Technology (HUST), Vietnam. He also received BS (2015) from HUST. His research interests include topic model and big data.

Ngo Van Linh is a Ph.D. student at Hanoi University of Science and Technology (HUST), Vietnam. He also received BS (2011) and MS (2014) degrees from HUST. His research interest includes topic model.



Khoat Than is currently an Associate Professor at Hanoi University of Science and Technology. He received Ph.D. degree from Japan Advanced Institute of Science and Technology in 2013. His recent research interests include representation learning, topic modeling.