Fuzzy Bag-of-Words Model for Document Representation

Rui Zhao and Kezhi Mao

Abstract—**One key issue in text mining and natural language processing (NLP) is how to effectively represent documents using numerical vectors. One classical model is the Bag-of-Words (BoW). In a BoW-based vector representation of a document, each element denotes the normalized number of occurrence of a basis term in the document. To count the number of occurrence of a basis term, BoW conducts exact word matching, which can be regarded as a hard mapping from words to the basis term. BoW representation suffers from its intrinsic extreme sparsity, high dimensionality, and inability to capture high-level semantic meanings behind text data. To address the above issues, we propose a new document representation method named Fuzzy Bag-of-Words (FBoW) in this paper. FBoW adopts a fuzzy mapping based on semantic correlation among words quantified by cosine similarity measures between word embeddings. Since word semantic matching instead of exact word string matching is used, the FBoW could encode more semantics into the numerical representation. In addition, we propose to use word clusters instead of individual words as basis terms and develop Fuzzy Bag-of-WordClusters (FBoWC) models. Three variants under the framework of FBoWC are proposed based on three different similarity measures between word clusters and words, which are named as FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$, respectively. Document representations learned by the proposed FBoW and FBoWC are dense and able to encode high-level semantics. The task of document categorization is used to evaluate the performance of learned representation by the proposed FBoW and FBoWC methods. The results on seven real word document classification datasets in comparison with six document representation learning methods have shown that our methods FBoW and FBoWC achieve the highest classification accuracies.**

Index Terms—**Document Representation, Fuzzy Similarity, Word Embeddings, Document Classification.**

I. Introduction

With exponential growth of the Internet, more than one exabyte (10^{18}) of data are created on the Internet each day. Among various types of data, text or document data accounts for a large portion. Different from structured data such as signals acquired by physical sensors, text data belong to unstructured data type. It is well known that effective machine learning algorithms should be supported by structured numerical representation of data and the good representation determines the upper bound of the performance of machine learning algorithms [1]. To draw insights from unstructured text data which is the goal of text mining and natural language processing, the important and core step is transforming these unstructured text data into numerical vectors. This problem is referred to as text representation learning. Representation learning of text is the keystone for various text mining tasks including text classification, document clustering and so on [2], [3], [4], [5], [6]. In this paper we will focus on representation learning of documents including short texts.

One commonly adopted and effective approach for document representation is the Bag-of-Words (BoW) model. BoW model assigns a vector to a document as $\boldsymbol{d} = (x_1, x_2, \ldots, x_l)$, where x_i denotes the normalized number of occurrence of the i-th basis term and l is the size of the collection of basis terms. It should be noted that the basis terms are the high frequency words in a corpus, and the number of basis terms or the dimensionality of BoW vectors is less than the size of vocabulary [7], [8], [9], [10]. BoW is a simple but effective method to map a document into a fixed-length vector. However, the mapping function in BoW model is hard or binary, i.e., the crisp binary relation, which only represents presence or absence of a basis term in the document. The hard mapping function has several limitations. First, the learned vector is extremely sparse since a document only contains a very small portion of all basis terms. Second, the BoW representations may not effectively capture semantics of documents since semantically similar documents with different word usages are mapped to very different vectors under BoW. To give an intuitive illustration of the above limitations of the BoW method, a toy example is provided in Figure 1. There exists a strong semantic correlation between two sentences d_1 *"a dog is above the bench"* and d_2 *"the huskies is on the table"*. Here, sentences are regarded as short documents. If the BoW employs the following basis terms: $\{$*"above","on","table","dog"*$\}$, the two sentences d_1 and d_2 will be projected to $\boldsymbol{d}_1 = (1, 0, 0, 1)$ and $\boldsymbol{d}_2 = (0, 1, 1, 0)$ respectively, as shown in Figure 1.(a). For sentence d_2 *"the huskies is on the table"*, due to the hard mapping of BoW, the informative word *huskies* is neglected, though it is semantically similar to *dog*. Similarly, the word *bench* in d_1 is not captured in the BoW representations. In addition, *above* and *on* are considered as two different dimensions in the BoW model since BoW adopts exact word matching. Thus, the cosine similarity measure between two learned vectors d_1 and d_2 is 0, which is contrary to the true semantic similarity between the two sentences. In summary, BoW model may not effectively capture semantics of documents, and this may eventually hinder performance in classification or regression problems.

In this paper, we propose fuzzy BoW models to learn more dense and robust document representations encoding more semantics. To overcome the limitations of the original BoW model as discussed above, we propose to replace the original hard mapping by a fuzzy mapping, and develop the Fuzzy BoW (FBoW) model. In contrast to BoW that employs exact

R. Zhao and K. Mao are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.
E-mail: rzhao001,ekzmao@ntu.edu.sg

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2690222, IEEE Transactions on Fuzzy Systems

word matching to basis terms, FBoW introduces vagueness in the matching between words and the basis terms. Fuzzy mapping enables a word semantically similar to a basis term to be activated in the BoW model. The membership function of a basis term in the FBoW model assigns membership values to words according to their semantic similarity to the basis term. The intuition behind such a membership function lies in that the membership values should be proportional to the semantic similarity between the word in documents and the basis terms. In our proposed model, word embeddings technique is introduced to help evaluate semantic similarity. Trained on a large corpus, word embeddings encode word meanings into vectors and thus the semantic similarity between two words can be conveniently evaluated using the cosine similarity between the corresponding word embeddings [11]. The cosine similarity measure can be interpreted as the degree of one word semantically matching another word. To illustrate the comparative advantages of our proposed fuzzy BoW over the original BoW, fuzzy BoW is applied to the same toy example, as shown in Figure 1.(b). Due to the adopted fuzzy mapping, *bench* in sentence d_1 and *huskies* in sentence d_2 can be mapped to the basis terms *table* and *dog* respectively, and their values are proportional to the semantic similarity. The fuzzy BoW produces the following vectors for the two toy sentences: $\boldsymbol{d}_1 = (1, 0.7, 0.8, 1)$ and $\boldsymbol{d}_2 = (0.7, 1, 1, 0.8)$. Obviously, the FBoW model produces two similar vectors for two semantically similar sentences. Based on FBoW, a fuzzy Bag-of-WordClusters (FBoWC) model is proposed. Different from fuzzy BoW (FBoW) model whose basis terms are single words, FBoWC uses clusters of words as the basis terms, where each cluster consists of semantically similar words. The fuzzy membership function is based on the similarity between words and word clusters. Three different similarity measures including mean, maximum and minimum between words and clusters are investigated, and this leads to three variants named FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$, respectively. The main contributions of our work can be summarized as follows:

* Our proposed fuzzy Bag-of-Words (FBoW) model is able to reduce sparsity, improve robustness and encode more semantic information than the original BoW model. Instead of binary mapping in BoW, FBoW adopts fuzzy mapping, in which connections between words in documents and basis terms in BoW space are fuzzy numbers. Fuzzy membership value can be determined according to word semantic similarity, which can be easily calculated based on word embeddings in our model. By introducing fuzzy mapping, our proposed FBoW can be regarded as a more general formulation of BoW.
* The proposed fuzzy Bag-of-WordClusters model (FBoWC) produces representation with lower dimensions than FBoW. In FBoWC, words are clustered based on semantic similarity between words, and the word clusters are used as basis terms. Similar to FBoW model, a fuzzy membership function is constructed based on the similarity measures between words in corpus and clusters.
* Comprehensive experiments on several real-life document categorization datasets have verified the performance of our proposed models. Quantitative analysis is provided to prove that the representation learned by the FBoW and FBoWC models can capture more semantic information.

This paper is organized as follows. In Section II, some related work including document representation learning and fuzzy systems are reviewed. Our proposed fuzzy BoW model for document representation are presented in Section III. In Section IV, experimental results on document classification tasks are illustrated. Finally, concluding remarks are given in Section V.

II. Related Work

Our work aims to incorporate fuzzy theory into the original BoW model to learn dense, robust and effective representations for documents. Therefore, document representation learning and fuzzy system are both related to our work. In the following, the previous work in these two areas are briefly reviewed, respectively

A. Document Representation Learning

As mentioned in the Introduction, document representation is the keystone for various text mining and NLP tasks. The most established Bag-of-words (BoW) model is often criticized for its extreme sparsity, high dimensionality and inability to capture semantics. Some works have been proposed to improve BoW model including latent semantic analysis (LSA) and topic models [12], [13], [14]. These models transform the BoW representation into low-dimension representations to capture the latent semantic structure behind documents. In LSA, singular value decomposition (SVD) is applied to the original BoW representation to obtain a new representation, where each new latent dimension is a linear combination of all original dimensions. In topic models including probabilistic latent semantic analysis [13] and latent dirichlet allocation [14], probability distributions are introduced to describe words and the generation process of each word in a document. The assumption behind topic models is that word choice in a document will be influenced by the topic of the document probabilistically. However, in these models, the derived latent dimension lacks semantic interpretation. For example, LSA regards a latent dimension as a linear combination of all original terms in vocabulary, which is counter intuitive because only a small part of the vocabulary is actually relevant to a certain topic. In addition, these two approaches both utilize the word occurrence of documents to perform dimensionality reduction. However, the occurrence statistics may not be able to capture the true semantic information underlying a document. Different from BoW model and BoW-enhanced models such as LSA and topic models that employ exact word matching and hard mapping, our proposed FBoW and FBoWC models adopt semantic matching and fuzzy mapping to project the words occurred in documents to the basis terms. In our proposed fuzzy BoW models, word embeddings is introduced to help evaluate semantic similarity between words. Since

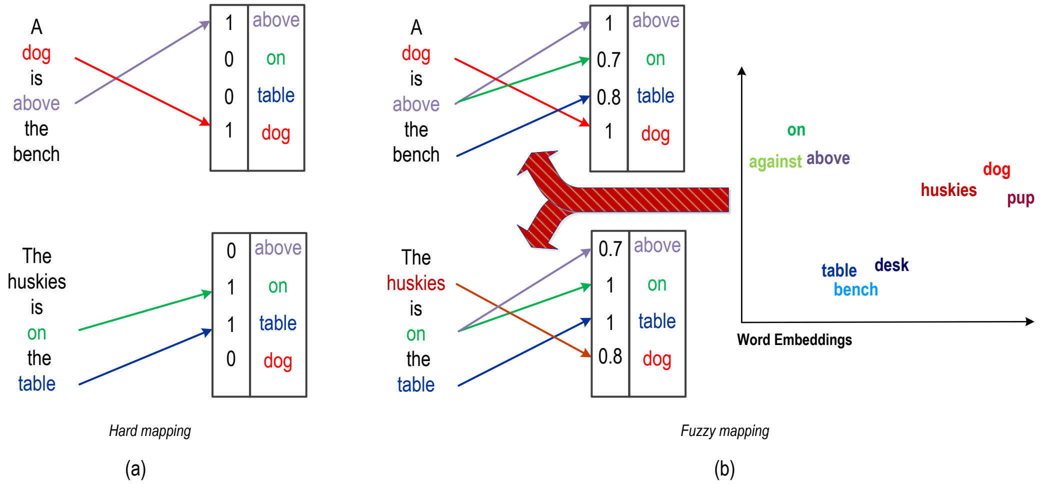

Fig. 1. Illustration and Comparison between Bag-of-Words and our proposed Fuzzy Bag-of-Words models based on a toy example : (a) Bag-of-Words; (b) Fuzzy Bag-of-Words.

word embeddings are trained on very large-scale corpus, it is believed that the captured similarity information is more accurate and general than that extracted from word occurrence statistics underlying a document in previous BoW-based approaches. In addition, our proposed fuzzy BoW models can also be used in conjunction with the LSA method to reduce the dimensionality of the FBoW representation.

In recent years, compositional models based on word embeddings have been proposed to learn text representation. Word embeddings use real-valued, dense and low-dimensional vectors to represent semantics of words [15], [16]. Different from one-hot word representation, a extreme sparse 1-of-V vector (V is the vocabulary size) in BoW model, word embeddings are able to encode semantic and syntactic information shared by words into a low-dimensional and dense space. In addition, word embeddings can be learned through neural language models from a large-scale text corpus without any annotations in a very efficient way [15], [17], [18]. For document representation, compositional models should be applied by treating a document as a sequence of words. The compositional models can be categorized into two types: one is shallow model that performs fixed algebraic operation including average-pooling or max-pooling on documents' word embeddings to derive their representations. The shallow model is based on the assumption that the semantic space is shared by documents and words. This assumption may be impractical since expressive power of documents are more complex and sophisticated than that of words. Another is deep model: multi-layers neural networks with various forms including recursive [19], [20], recurrent [21], [22] and convolution [23], [24], [25] neural networks to learn and perform compositions over word embeddings. Although the deep models consider the order of words in the documents, they have several limitations including heavy computation, demands of sufficient and high-quality annotated training corpus, and tricky hyperparameter settings. In addition, the deep models might get lost in the long-distance dependency existing in documents. Therefore, deep models are extensively adopted in sentence modeling instead of document modeling. Our proposed FBoW models are unsupervised and much more efficient than the deep models for document modeling.

In [26], a novel dissimilarity measure between two documents named word mover's distance (WMD) is proposed. The new measure is defined as the minimum distance that words in one document should travel to the words in another documents, where word distance is quantified based on word embeddings. Our proposed FBoW and FBoWC models share the same view with [26] that BoW representation may not capture the semantic information and word embeddings can be used to address this problem. But WMD attempts to alleviate the issue, while our works aim to solve the problems of BoW from their root cause.

B. Fuzzy System and Its Applications in Text Mining

The core idea behind the fuzzy system is the employment of membership functions [27]. Fuzzy system is composed of three successive modules including fuzzification, inference, and defuzzification, which can be regarded as knowledge-based nonlinear system. Fuzzy system has been extensively applied in many fields, including automatic control, clustering, image processing, optimization and so on [28], [29], [30], [31], [32]. It should be noted that [29] also proposed a model named Fuzzy BoW model for computer vision tasks, which shares the same name with our work. The FBoW in [29] attempted to address feature vector ambiguity problem for image representation, while our FBoW solves semantic and expression diversity problem of natural language for document representation. In addition, our models are different from theirs in fuzzy membership constructions, motivations and design of *codebook* or basis terms under the context of documents. Since our work focuses on text data, some previous papers incorporating fuzzy logic into text mining problems are reviewed here [33], [34], [35], [36], [37]. Tjhi et.al [33] proposed a dual fuzzy-possibilistic coclustering for document clustering, in which document clusters and fuzzy-possibilistic word memberships functions are generated jointly. Chiang et.al [34] presented a hierarchical web document clustering method named fuzzy latent semantic clustering (FLSC). FLSC introduced co-occurring named entity associations to build a fuzzy linguistic topological space. Lee et.al [36] presented a fuzzy based method named ML-FRC to address multi-class text categorization problems. In ML-FRC, fuzzy transformation is adopted to obtain low-dimensional fuzzy relevance vectors and the fuzzy clusterings are linearly mapped to labels. Based on the linear mapping, a set of thresholds are computed on training corpus and used on testing corpus to classify unseen documents. Martin et.al [37] proposed fuzzy grammar fragments to extract structured grammar components from unstructured text. These above approaches all focused on classification and clustering methods for text data. However, our approaches address the representation learning in text mining. Jiang et.al [35] proposed a fuzzy self-constructing feature clustering algorithm (FFC) to learn text representation methods. The core idea that a cluster of words is regarded as a dimension for document vector behind FFC is partially similar to our proposed FBoWC. However, FFC is a supervised method that requires labels to generate word patterns and its applications are limited to multi-class document categorization. Our proposed FBoW models are able to derive document representation in an unsupervised way. Furthermore, the document vectors learned by our FBoW can be applied to various text mining applications including information retrieval, text categorization and text clustering etc.

III. Fuzzy Bag-of-Words

In this section, our proposed fuzzy Bag-of-Words models are presented. Since fuzzy membership function is constructed based on word embeddings, we begin with a brief review of word embeddings.

A. Word Embeddings

TABLE I
Top5 similar terms to words: *book* and *student*. They are retrieved based on word embeddings. Their corresponding cosine-similarity scores are also shown.

Inquiry	Similar Words	Cosine Similarity Scores
book	*tome*	0.748
	books	0.738
	memoir	0.730
	paperback_edition	0.686
	autobiography	0.674
student	*students*	0.729
	teacher	0.631
	faculty	0.608
	school	0.605
	undergraduate	0.602

The core idea behind word embeddings is to assign such a dense and low-dimensional vector representation to each word that semantically similar words are close to each other in the vector space. The merit of word embeddings is that the semantic similarity between two words can be conveniently evaluated based on the cosine similarity measure between corresponding vector representations of the two words. In the popular word embeddings *word2vec* [15], [11], [18], a two-layer neural network language model is designed to learn vector representations for each word. The *word2vec* framework contains two separate models including Continuous Bag of Words (CBoW) and Skip-gram with two reverse training goals. CBoW tries to predict a word given the surrounding words, while Skip-gram tries to predict a window of words given a single word. Due to its surprisingly efficient architecture and unsupervised training protocol, *word2vec* can be trained over a large-scale unannotated corpus efficiently. *word2vec* is able to encode meaningful linguistic relationships between words into learned word embeddings. Usually, the cosine similarity measure between word embeddings is used to measure the semantic similarity between two words:

$$cos(w_i, w_j) = \frac{\mathbf{w_i} \cdot \mathbf{w_j}}{\|\mathbf{w_i}\| \, \|\mathbf{w_j}\|} \tag{1}$$

where $\mathbf{w_i}$ and $\mathbf{w_j}$ denote word embeddings of two words w_i and w_j, respectively. The cosine similarity measure is positive when the words are close to each other, and is negative when the words have reverse meaning. The measure is zero under a pair of two completely random words. To give a illustration, the top 5 similar words to two example words *book* and *student* and their cosine similarity scores are given in Table I. In our proposed FBoW models, cosine similarity measure based on wohttps://arxiv.org/pdf/1301.3781.pdfrd embeddings are utilized to construct fuzzy membership functions to map the words in documents to basis terms. It should be noted that our proposed models do not consider polysemy issue, since the single prototype word embeddings is used as input. Although there are some multi-sense word embeddings in the literature, the disambiguation process of each word sense is quite challenging, and therefore hinders the application of multi-sense word embeddings [38], [39], [40]. In addition, documents usually contain lots of words, the effects of neglecting polysemy is less vital than in word-level or sentence-level. However, it is still meaningful to look into multi-sense word embeddings in our proposed FBoW models in the future work.

B. Fuzzy Bag-of-Words Model

Firstly, some adopted notations in our proposed methods are introduced. Let $D = \{w_1, \ldots, w_v\}$ be the vocabulary covering all the words existing in the text corpus, and v is the vocabulary size. $\mathbf{W} \in \mathbb{R}^{v \times d}$ denotes a well-trained word embeddings matrix, where its i-th row $\mathbf{w}_i \in \mathbb{R}^d$ represents the d-dimensional word embeddings for word w_i. Each document in text corpus is represented by a BoW vector whose elements denote the number of occurrence of basis terms in the document. In a large corpus, only the top l high-frequency words are usually selected as basis terms in BoW model to reduce the sparsity and dimensionality in BoW representations, and the BoW basis terms $T = \{t_1, \ldots, t_l\}$ is therefore a subset of the corpus vocabulary, i.e., $T \subset D$.

For traditional BoW representations, documents are mapped into vectors by exact matching of the words in the documents to the basis terms. Exact word matching is equivalent to performing a hard or crisp mapping. If a word w matches a basis term t_i, the output of the crisp mapping function is 1, and is zero otherwise. Exact word matching is equivalent to employing the following membership function:

$$A_{t_i}(w) = \begin{cases} 1 & \text{if } w \text{ is } t_i, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

Figure 2. (a) also illustrates this hard mapping. BoW has two immediate impacts on document representations. First, the vector is very sparse. Second, the vector does not encode much semantic information. As illustrated in Section I and Figure 1, two semantic similar sentences with different word usages are mapped to two uncorrelated BoW vectors, which is counter intuitive.

Fuzzy Membership Function: To address the problem caused by exact word matching in BoW, we propose to use semantic matching, which matches two words based on semantic similarity. To implement semantic matching, we employ fuzzy membership function as shown in Figure 2.(b). The corresponding equation is given as follows:

$$A_{t_i}(w) = \begin{cases} cos(\mathbf{W}[t_i], \mathbf{W}[w]) & \text{if } cos(\mathbf{W}[t_i], \mathbf{W}[w]) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{W}[w]$ denotes word embeddings for word w. Here, we adopt *word2vec* as our word embeddings, and the fuzzy membership degree that measures the similarity between attribute (words in documents) and set (basis terms in BoW space) is approximated by their cosine similarity score. The membership function Eq. (3) is intuitive since only words that are semantically similar to basis terms (positive correlated) should be counted. In contrast to hard membership function that only counts words identical to basis terms, fuzzy membership function allows words semantically similar to the basis terms to be counted. As shown above, the fuzzy membership function of one basis term is determined by the cosine similarity between its corresponding word embeddings and other word embeddings. A word which is similar to basis terms should be mapped to a large degree of membership. In the following, we will present how the fuzzy membership function can be incorporated into the BoW model to provide a semantic document representation.

Representation Learning: Here, the fuzzy membership function is adopted to count the number of occurrence of basis terms in a document. For a document, the FBoW model representation is denoted by $\mathbf{z} = [z_1, z_2, \ldots, z_l]$, where the i-th element z_i is the sum of the membership degrees that all words semantically match the i-th basis term, i.e.:

$$z_i = c_i \sum_{w_j \in \mathbf{w}} A_{t_i}(w_j) x_j \tag{4}$$

where $\mathbf{w}$ denotes a set of all words in the document, t_i is the i-th basis term and x_j denotes the number of occurrence of w_j. It should be noted that c_i is a controlling parameter defined by different weighting schemes in BoW model. For example, $c_i = 1$ if the counting scheme is adopted, while c_i is the inverse-document frequency if the TF-IDF is adopted. For simplicity, we adopt the counting scheme as our weighting scheme and c_i is fixed to 1. As shown in Eqs. (2) and (4), BoW model can be regarded as a special case of our proposed fuzzy model. In BoW, x_i is determined by the term frequency only, which is equivalent to the utilization of hard membership function.

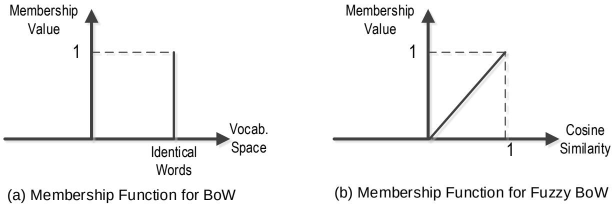

Fig. 2. Illustrations of membership functions behind Bag-of-Words and our proposed Fuzzy Bag-of-Words models : (a) Bag-of-Words; (b) Fuzzy Bag-of-Words.

In the following, a matrix formulation of the above fuzzy BoW model is presented. The numerical vector representation of a document under fuzzy BoW model is given by:

$$\mathbf{z} = \mathbf{x}\mathbf{H} \tag{5}$$

where

$$\mathbf{x} = [x_1, x_2, \ldots, x_v] \tag{6}$$

x_i denotes the number of occurrence of word w_i in the document;

$$\mathbf{H} = \begin{bmatrix} A_{t_1}(w_1) & A_{t_2}(w_1) & A_{t_3}(w_1) & \ldots & A_{t_l}(w_1) \\ A_{t_1}(w_2) & A_{t_2}(w_2) & A_{t_3}(w_2) & \ldots & A_{t_l}(w_2) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ A_{t_1}(w_v) & A_{t_2}(w_v) & A_{t_3}(w_v) & \ldots & A_{t_l}(w_v) \end{bmatrix} \tag{7}$$

where $A_{t_i}(w_j)$ is given by Eq. (3). Via Eq. (5), the final representation for document can be obtained. Each element of $\mathbf{x}$ is the weighted sum of word occurrence in the documents, where the weights are the membership values of the words in documents belonging to the corresponding basis terms in BoW model.

C. Fuzzy Bag-of-WordClusters Model

It is well acknowledged that BoW model has three limitations, including sparsity, high dimensionality, and lack of capability to encode high-level semantics. The fuzzy BoW model developed in Section III-B addressed the issues of sparsity and semantics, but the high dimensionality problem remains. Actually, the high dimensionality also means redundancy. This is the reason why BoW is often combined with LSA to reduce dimensionality. Certainly, FBoW can also be combined with LSA to reduce the dimensionality and redundancy of FBoW representation. In this study, we propose a plausible method to solve the high dimensionality and redundancy problem of FBoW model. The basic idea of the new method is to use word clusters instead of individual words as basis terms. A word cluster is a group of semantic similar words. Since basis terms are clustered, the dimension and redundancy can be significantly reduced. In FBoWC, even the entire vocabulary can be used as basis terms without increasing dimensionality of learned representations. We proposed three measures to evaluate the similarity between words and clusters and developed three variants of FBoWC models named as FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$ by utilization of mean, max and min operation to obtain similarity between words and clusters, respectively. The three different similarity measures are illustrated in Figure 3, where max, min and mean scores are used to evaluate the similarity between a certain word and word cluster 1, 2 and 3, respectively. The details of the three models are presented next.

Word Cluster Construction: Since word embeddings can capture the semantics of words, K-means clustering algorithm is applied to group these embeddings, i.e., discover word clusters. The number of clusters, which can be set by users, is equal to the dimensionality of the learned document representation.

Fuzzy Membership Function for FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$: To derive fuzzy membership functions as given in Eq. (3), similarity measures between clusters and words should be defined. Here, three measures have been proposed, which lead to three models including FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$. It is assumed the i-th word cluster $\mathbf{t}_i$ contains a collection of word $\mathbf{t}_i = \{w_{i1}, \ldots, w_{ik_i}\}$, where k_i denotes the number of words in the i-th word cluster. To evaluate the similarity between cluster $\mathbf{t}_i$ and word w, we firstly compute the cosine-similarity between embeddings of words in cluster $\mathbf{t}_i$ and word w. Then, FBoWC$_{\text{mean}}$, FBoWC$_{\text{max}}$ and FBoWC$_{\text{min}}$ take mean, maximum and minimum value, respectively:

$$A_{\mathbf{t}_i}(w) = \begin{cases} \text{mean}(\mathbf{q}_{\mathbf{t}_i}) & \text{for FBoWC}_{\text{mean}}, \\ \max(\mathbf{q}_{\mathbf{t}_i}) & \text{for FBoWC}_{\text{max}}, \\ \min(\mathbf{q}_{\mathbf{t}_i}) & \text{for FBoWC}_{\text{min}}, \end{cases} \tag{8}$$

where

$$\begin{aligned} \mathbf{q}_{\mathbf{t}_i} =& [\max(cos(\mathbf{W}[w_{i1}], \mathbf{W}[w]), 0), \ldots \\ & \max(cos(\mathbf{W}[w_{ik_i}], \mathbf{W}[w]), 0)] \end{aligned} \tag{9}$$

$\mathbf{q}_{\mathbf{t}_i}$ is a vector whose elements are the similarity measures between words in cluster $\mathbf{t}_i$ and word w. In the max operation in FBoWC$_{\text{max}}$, the most similar word in word cluster $\mathbf{t}_i$ to word w is identified, and the similarity measure with the most similar word is taken. This is to make sure that the similarity measure takes the value of 1 when the basis word is mapped to itself. Max operation has the effect of reducing intra-class spared in the dimension where the words in a word cluster are very relevant to the category of the document. Contrary to max operation, the min operation identifies the least similar word in a word cluster and takes the similarity measure with the least similar word in the word cluster as the output. Min operation also has the effect of reducing intra-class spread in the dimension where the words in the word cluster are NOT relevant to the category of the document. Both max and min operations have the effect of reducing intra-class spread, which is beneficial to pattern classification. The mean operation is equivalent to taking the similarity value with the centroid of a cluster, i.e. the word of a codebook in BoW for image representation [29].

The procedure of our proposed FBoW and FBoWC for document representation learning is given in Algorithm 1.

Algorithm 1 Fuzzy Bag-of-Words Frameworks

```
Input: a text corpus with $n$ documents; the vocabulary $\mathbf{D}$ and
    its corresponding word embeddings matrix $\mathbf{W} \in \mathbb{R}^{v \times d}$,
    where $v$ is the vocabulary size and $d$ is the dimensionality of word
    embeddings. Required dimensionality for document vectors: $l$
Output: learned document vectors for the corpus: $\mathbf{Z} \in \mathbb{R}^{n \times l}$

Based on the corpus vocabulary $\mathbf{D}$, obtain data matrix $\mathbf{X} \in \mathbb{R}^{n \times v}$ that each row $\mathbf{x} \in \mathbb{R}^{v}$ is the $i$-th document vector whose $j$-th element is the number of occurrence of word $w_j$ in the corresponding document, as shown in Eq. (6);
if FBoW is performed then
    Based on term frequencies over the corpus, select the top-$l$ words with highest frequency as our models' BoW space $\mathbf{T}$ and the corresponding word embeddings are obtained as $\mathbf{W}_T \in \mathbb{R}^{L \times d}$;
    Construct transformation matrix $\mathbf{H}$ based on $\mathbf{W}$ and $\mathbf{W}_T$ using Eqs. (3) and (7);
else if FBoWC is performed then
    Apply K-means algorithm to cluster words based on word embeddings matrix $\mathbf{W}$ by setting the number of clusters to $l$. Then, the embeddings of words in each clusters are obtained and the cosine similarity between these clusters' words and word in documents are computed as shown in Eq. (9);
    Construct transformation matrix $\mathbf{H}$ based on $\mathbf{W}$ and $\mathbf{q}_{\mathbf{t}_i}$ using Eqs. (8) and (7);
end if
Calculate learned data matrix $\mathbf{Z}$ according to Eq. (5), which can be used to represent the corpus.
return Z
```

D. Relationships with Previous Text Representation Methods

Word embeddings are introduced to capture the semantic relationships among words, and the derived semantic similarity and fuzzy mapping are then incorporated into the original BoW model. As a result, the learned document representations are more dense and able to capture more semantic information. In this subsection, we analyze the connections between our proposed FBoW frameworks including FBoW and FBoWC with two typical text representation learning models including dimensionality reduction methods and a deep composition model: convolutional neural network (CNN).

Relationships with Dimensionality Reduction: Dimensionality reduction techniques seek to reduce the rank of vectors. Through dimensional reduction, sparse and high-dimension

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2690222, IEEE Transactions on Fuzzy Systems

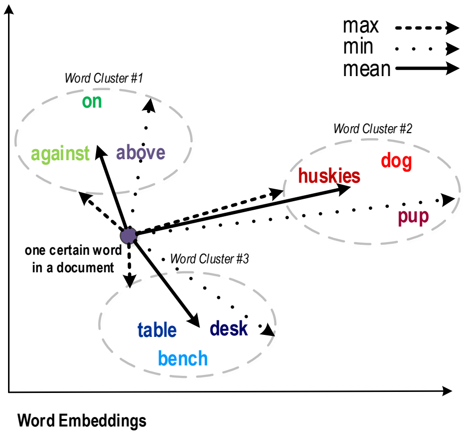

Fig. 3. Illustrations of mean, maximum and minimum similarity measures between clusters and words. It is noted that a high similarity measure denotes a small distance shown in the Figure.

BoW vectors can be transformed into dense and low-dimensional ones, which in turn boosts the performance of subsequent tasks such as classification, information retrieval, etc. Some models including latent semantic analysis (LSA) and random projection (RP) are applied extensively in many text mining applications [12], [41]. LSA and RP are linear dimensionality reduction methods, and the key issue is to find the mapping matrix. For LSA, the mapping matrix is learned via maximizing the preservation of variance of the original feature space. Since the input information for LSA can be regarded as occurrence statistics between documents and words, LSA may fail to model the true semantic information and the resulting dimensions may not have interpretable meaning in natural language [42]. For RP, the mapping matrix is generated randomly. Some experimental results have shown that RP can achieve a significant speedup in computation time will little distortion of pairwise information of data. However, without data-based parameter tuning, RP may not capture the semantic information underlying the natural language.

In FBoWC representations, each dimension corresponds to word clusters which are subsets of the entire vocabulary. By contrast, each dimension in LSA and RP is a linear combination of all words in the vocabulary. As the mapping matrix of FBoWC in Eq. (7) directly measures the semantic similarity between words and basis terms based on word embeddings, it can capture high quality semantic information. In addition, word embeddings are pre-trained and publicly available, the computational cost is not a potential problem for FBoWC.

Relationships with Convolutional Neural Network: With the development of deep learning, recent works on text representation learning have been focused on deep compositional models [43], [20], [23]. The essence of deep compositional models is the application of multi-layers neural network of various forms including recursive [19], recurrent [21] and convolution [23] neural networks to learn and perform compositions over word embeddings. Among these models, convolutional neural network-based compositional models have been proven to be very efficient for sentence or short text modeling. Similar to CNN, the projection process adopted in our FBoW models can be regarded as a convolutional operation whose filter vector is the word embeddings of basis terms and window size is fixed to one gram. After sliding the filter through the whole sentence, the summation pooling method is adopted to get the activation of the filter. The output of each filter corresponding to each basis term are concatenated together to form the final document representation.

CNNs are suitable for sentence or short text modeling, but not suitable for document modeling due to the limits of window size. Our FBoW and FBoWC are effective for document or short text modeling. In addition, our approach belongs to unsupervised methods with limited computational cost, while CNN requires a large high-quality labeled training corpus. Furthermore, the filter parameters in CNN are tuned via back-propagation and the yielded representation lacks the semantic interpretability, but in our proposed FBoW and FBoWC, each dimension corresponds to a word or a cluster of words and is therefore interpretable.

IV. Experiments

In this section, we use document categorization tasks to evaluate the performance of our proposed Fuzzy Bag-of-words models.

A. Descriptions of Datasets

The task of document categorization is to assign a class label or category to a document. Seven real-life datasets are used in the experiments.

20Newsgroups is a collection of nearly 20,000 newsgroup documents, which is organized into 20 different classes. Here, we adopted the version of 20 Newsgroups (20NG) sorted by the removal of duplicates and some headers[1]. The whole corpus has 18846 documents, and the vocabulary size is 32716, excluding the removed words whose document frequencies are less than five. Actually, the removal of low frequency words were performed for all the seven datasets used in the experiments. We followed predefined training and testing splitting. The statistics of 20NG are given in Table II.

Reuters_14 and *Reuters_8* were both generated from a classical corpus Reuters-21578 containing newswire articles and Reuters annotations[2]. The whole collection has 21,578 documents, which are categorized into 90 classes. Since some categories have only a few documents, we created two datasets containing 14 and 8 most frequent classes, respectively. The predefined training and testing splitting was adopted. The statistics of these two datasets *Reuters_14* and *Reuters_8* can be found in Table II.

Amazon_6 is a collection of Amazon reviews for products of six categories. This dataset was originally published for sentiment analysis [44], but we used it for categorization. These

[1] The dataset has been kindly provided at http://qwone.com/~jason/20Newsgroups/

[2] The dataset has been kindly provided at http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html

six categories are cameras, laptops, mobile phone, tablets, TVs and video surveillance, in which the largest sample number is 6736 under cameras and the smallest sample number is 881 under tablets. To make the dataset more balanced, we randomly selected 1500 samples from categories with more than 1500 reviews. The corpus used in our experiments has 8083 reviews with a vocab size of 10790. The details are shown in Table II.

Amazon_4 is a multi-domain sentiment dataset, which contains product reviews crawled from Amazon for 4 product types: Kitchen, Books, DVDs and Electronics [45]. Here, we used it for classification of product types instead of sentiment analysis. The statistics of this dataset are given in Table II.

BBC is a collection of 2225 documents from the BBC news website [46]. The documents are crawled from 2004-2005 in five topical areas including business, entertainment, politics, sport and tech. The vocab size is 8865.

BBCSport is a collection of 737 documents crawled from the BBC Sport website in five topical areas including athletics, cricket, football, rugby, tennis [46]. The vocab size is 3669.

Since training and testing subsets have not been given explicitly in their original sources in *Amazon_6*, *Amazon_4*, *BBC* and *BBCSport*, we split the corpus into training and testing domains equally. The train/test splitting was repeated 15 times. Therefore, for each corpus, 15 training/testing sub-datasets were created. The mean accuracy and standard deviation of each compared approach on 15 sub-datasets of each corpus were reported in the paper.

To preprocess the above corpus, a English stop-word list built in the library *sklearn*[3] was used for stop-word removal, and all characters were converted to lowercase because stop-word and case sensitivity may not be informative to document categorization tasks. In addition, stemming and lemmatization were not conducted, since words embeddings assign vectors to all forms of words instead of only their base forms.

B. Experimental Setup

Our proposed FBoW and FBoWC models are compared with the following document representation methods:

* BoW: the raw BoW representations are directly fed into the classifier. Here, the top high-frequency words, which is a subset of the vocabulary, are used as basis terms.
* BoW_{full}: Different from BoW model, BoW_{full} use the entire vocabulary as basis terms.
* LSA: Latent Semantic Analysis [47].
* LDA: Latent Dirichlet Allocation [14]. Our implementation of LDA is based on *Gensim*[4].
* AE: Average Embeddings take the average of word embeddings of all words contained in the document to represent the document.
* WMD: Word Mover's Distance [26]. The code is kindly provided[5].
* FBoW and FBoWC ($\text{FBoWC}_{\text{mean}}$, $\text{FBoWC}_{\text{max}}$ and $\text{FBoWC}_{\text{min}}$): our proposed fuzzy bag-of-words models that utilize top high-frequency words and words clusters of the entire vocabulary as basis terms, respectively.

The input matrix for dimensionality reduction methods including LSA and LDA were both performed on BoW_{full} representation, which uses the entire vocabulary as the basis terms. For a fair comparison, FBoWC also used the clusters of entire vocabulary as basis terms. In BoW model, the top 2000 high-frequency words were used as basis terms and the resulting document representations have a dimensionality of 2000. For FBoW, the same top 2000 high-frequency words were used as basis terms. For LSA and LDA, the number of latent topics are both set to 2000. In LDA, we set hyperparameter α for document topic multinomial and hyperparameter η for word topic multinomial to 1 and 0.01, respectively. For FBoWC, the number of clusters was set to 2000. Thus, the derived representations of LSA, LDA, FBoW and FBoWC have the same dimensionality of 2000.

For AE, WMD, FBoW and FBoWC models, the same word embeddings were used. We utilized the pre-trained *word2vec* vectors published by Google[6]. These word embeddings were trained on a Google News corpus (over 100 billion words) and have a dimensionality of 300. For all the seven document categorization tasks, we further fine-tuned the pre-trained word embeddings over the specific dataset. Since AE averages embeddings of all words, the dimension of document vector learned by AE is the same as the dimension of word embeddings, which is 300. The other settings of WMD method were the same as that reported in its original paper [26].

Linear SVM [48] was applied to the document representations learned by the above mentioned approaches. In linear SVM, we searched the best regularization parameter C from $\{0.001, 0.01, 0.1, 1,$
$10, 100\}$. Since WMD can only derive document distance instead of document representations, document classification based on WMD used the *k*NN decision rule [49]. The searching range of the neighborhood size *k* is $\{1, 3, \dots, 19\}$.

C. Experimental Results on Document Categorization Tasks

In this section, we compare the experiment results of our proposed FBoW and FBoWC models and several other benchmark document representation learning approaches on seven document classification datasets. The classification accuracies of each method on the seven datasets are shown in Table III. For datasets *Amazon_6*, *Amazon_4*, *BBC* and *BBCSport*, the reported mean accuracies and standard deviations of all compared methods are over 15 random training/testing splittings.

First, it is observed that BoW_{full} model performs slightly worse than BoW model in five out of seven datasets. The difference between these two models lies in that BoW_{full} use all words in the vocabulary, while BoW only use top 2000 high-frequency words as basis terms. This means that the high dimensionality and the sparsity of BoW_{full} hinder its performance. Our approaches including FBoW and FBoWC all gain a significant performance improvement compared to original BoW models. This is because of the introduction

[3] http://scikit-learn.org/stable/

[4] https://radimrehurek.com/gensim/index.html

[5] https://github.com/mkusner/wmd

[6] https://code.google.com/archive/p/word2vec/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2017.2690222, IEEE Transactions on Fuzzy Systems

TABLE II
STATISTICAL PROPERTIES OF SEVEN DOCUMENT CATEGORIZATION DATASETS.

Statistics	Document Categorization Tasks						
	20NG	*Reuters_14*	*Reuters_8*	*Amazon_6*	*Amazon_4*	*BBC*	*BBCSport*
Vocab Size	32716	9413	8812	10790	9925	12473	5345
Sample No.	18846	8286	7674	8083	8000	2225	737
Class No.	20	14	8	6	4	5	5
Average Length (Word Counts)	349	105	102	395	146	412	370
Training/Testing Splits	11314/7532	5967/2319	5485/2189	4041/4042	4000/4000	1112/1113	368/369

of fuzzy mapping. Conventional BoW models assume all words are independent and adopt hard mapping, and this makes the learned document representations sparse and lack of high-level semantic information. By contrast, our proposed approaches adopt a fuzzy membership function and word embeddings to account for semantic correlation between words in a document and basis terms to learn dense and semantic document representations.

LSA, LDA and our proposed FBoWC models all are dimensionality reduction methods. For a fair comparison, the dimension of latent space in these methods is fixed to 2000. It can be seen that our proposed FBoWC models outperform LSA and LDA on all seven datasets. This is because our proposed FBoWC models utilize semantic matching and fuzzy membership function.

The other two methods AE and WMD both employ word embeddings as our methods do. The results have shown AE and WMD are unable to achieve comparative performance. For AE, document representations are obtained via averaging the embeddings of all words occurred in the document. However, the averaging operation is too simple to model the complex semantic composition process from words to document. For WMD, the distance between words based on word embeddings is utilized to learn distance between documents instead of document representation, which limits its application. In document categorization applications, only distance-based classifiers such as *k*NN model can be used, while our proposed FBoW and FBoWC produce document representations that can be combined with any classifier such as SVM, which usually outperforms *k*NN.

The performance of FBoW and FBoWC is also compared. FBoW adopts top 2000 high-frequency words as basis terms, while FBoWC use 2000 word clusters of the entire vocabulary as basis terms. As shown in Table III, FBoWC models usually outperform FBoW. The better performance of FBoWC may be owing to two reasons. First, FBoWC representation is more informative than FBoW representation, though they have the same dimensionality, i.e., 2000 in the experiment. In FBoW, the 2000 dimensions correspond to the top 2000 high-frequency words, while the 2000 dimensions in FBoWC correspond to the entire vocabulary grouped into 2000 clusters. Second, FBoWC representations contain less redundancy. In FBoW, even two basis terms are semantically similar, they will create their own dimensions in the vector representation, while in FBoWC, semantically similar words are grouped to one cluster and derive only one dimension in the vector representation. Thus, FBoWC contains less redundancy than FBoW representation, which in turn produces better performance than FBoW. Among the three variants of FBoWC models, the performance of FBoWC$_{\text{mean}}$ and FBoWC$_{\text{max}}$ over different datasets is more stable and robust than FBoWC$_{\text{min}}$.

To further verify the significance of the better performance achieved by our methods, paired *t*-test was conducted to compare population means of two approaches. Under these four adopted corpus including *Amazon_4*, *Amazon_6*, *BBC* and *BBCSport*, the best performing method proposed by us and the most competitive one from other 6 methods were selected for paired *t*-test. For each dataset, 15 random training/testing splittings were performed and the 15 training/testing splits were kept identical for all compared methods. The difference of mean accuracy and standard deviation of 15 results was computed as $\bar{d}$ and σ_d, respectively. Then, the *t*-statistic was calculated as:

$$t = \frac{\bar{d}}{\sigma_d / \sqrt{n}} \qquad (10)$$

where $n = 15$. Under the null hypothesis that the performance difference of the methods pair is zero, the *t*-statistic follows a *t*-distribution with $n - 1 = 14$ degrees of freedom. From the table of the *t*-distribution, p-values of the four selected method pairs are all less than 0.001, which means the best performing model in our proposed FBoW frameworks outperform the other comparative models at nearly 100% confidence level.

D. Experimental Analysis of Mapping Bounds

The membership function Eq. (3) is a special case of the following more general membership function with $\lambda = 0$:

$$A_{t_i}(w) = \begin{cases} cos(\mathbf{W}[t_i], \mathbf{W}[w]) & \text{if } cos(\mathbf{W}[t_i], \mathbf{W}[w]) > \lambda, \\ 0 & \text{otherwise,} \end{cases} \qquad (11)$$

where λ is the mapping bound or threshold. In the experiment, the influence of the parameter λ on the performance of FBoW and FBoWC was investigated. Parameter λ was set from 0 to 1 with a step size of 0.1. FBoWC$_{\text{min}}$ is not considered here since the applied threshold λ could easily set its fuzzy membership values to zero. Since clustering conducted in FBoWC$_{\text{mean}}$ and FBoWC$_{\text{max}}$ requires random initialization, we performed the above two models five times and recorded their mean accuracies and standard deviations. Since the trends over all datasets are almost the same, we only show the results of *Amazon_4* in Figure 4. It is observed when parameter λ increases from 0 to 1, the performance of our proposed approaches general degrades. FBoW, FBoWC$_{\text{mean}}$ and FBoWC$_{\text{max}}$ all perform best when $\lambda = 0$. The increased λ increases the sparsity of the vector representation of a document. This means that the membership function Eq. (3) is effective. Another merit of

TABLE III
CLASSIFICATION ACCURACIES (%) FOR COMPARED METHODS ON SEVEN DOCUMENT CATEGORIZATION DATASETS. BOLD FACE INDICATES BEST PERFORMANCE.

Datasets	BoW	BoW_{full}	LDA	LSA	AE	WMD	FBoW	$\text{FBoWC}_{\text{mean}}$	$\text{FBoWC}_{\text{max}}$	$\text{FBoWC}_{\text{min}}$
20NG	68.0	70.1	73.2	74.2	72.5	74.1	77.7	**78.9**	78.5	75.5
Reuters_8	96.6	95.3	96.2	96.5	96.5	94.5	97.1	97.2	97.3	**97.5**
Reuters_14	96.0	94.2	92.8	96.1	96.3	96.1	96.7	97.1	97.0	**97.2**
Amazon_4	91.3 ± 0.3	90.1 ± 0.2	91.1 ± 0.3	92.0 ± 0.2	92.2 ± 0.3	92.1 ± 0.3	93.5 ± 0.2	93.5 ± 0.3	$\mathbf{93.9 \pm 0.3}$	93.5 ± 0.3
Amazon_6	90.8 ± 0.2	90.9 ± 0.2	90.6 ± 0.3	91.0 ± 0.3	91.1 ± 0.3	91.6 ± 0.2	92.1 ± 0.2	$\mathbf{92.9 \pm 0.1}$	92.0 ± 0.2	92.0 ± 0.2
BBCSport	98.1 ± 0.7	97.5 ± 0.5	97.1 ± 0.6	98.3 ± 0.8	98.2 ± 0.7	97.6 ± 0.5	98.5 ± 0.8	98.5 ± 0.7	$\mathbf{98.9 \pm 0.8}$	98.3 ± 0.8
BBC	97.1 ± 0.4	96.5 ± 0.4	97.3 ± 0.3	97.3 ± 0.4	97.2 ± 0.3	96.1 ± 0.3	97.8 ± 0.4	97.9 ± 0.3	$\mathbf{98.2 \pm 0.4}$	97.5 ± 0.3

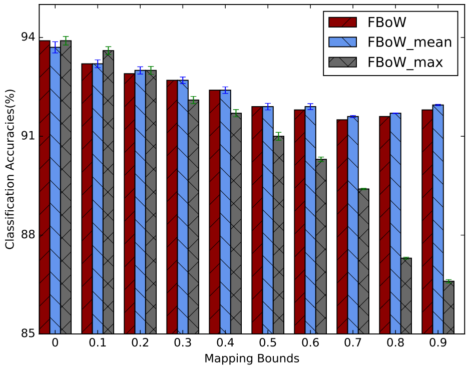

Fig. 4. Performance of FBoW, $\text{FBoWC}_{\text{mean}}$ and $\text{FBoWC}_{\text{max}}$ for different mapping bounds: λ.

Eq. (3) is that it is parameter-free and easy to use without requiring parameter tuning.

V. CONCLUSION

In this work, we have proposed Fuzzy Bag-of-Words models including FBoW and FBoWC to address issues of sparsity and lack of high-level semantics of BoW representation. Word embeddings are utilized to measure semantic similarity among words and construct fuzzy membership functions of basis terms in BoW space over words in the task-specific corpus. Since *word2vec* embeddings can be trained over billions of words, word embeddings adopted in our methods are able to capture high-quality and meaningful semantic information that are not contained by the task-specific corpus alone. To determine basis terms in BoW space, FBoWC utilizes word clusters, while FBoW directly regards high term-frequencies words as original BoW does. The adoption of word clusters in FBoWC can reduce feature redundancy and improve feature discrimination. Three different measures have been designed to evaluate similarity between clusters and words, and three corresponding variants of FBoWC models as $\text{FBoWC}_{\text{mean}}$, $\text{FBoWC}_{\text{max}}$ and $\text{FBoWC}_{\text{min}}$ have been developed. The performance of our approaches has been experimentally verified through seven multi-class document categorization tasks. As a next step work, document structure or word order information will be considered in document representation learning. In addition, the effects of multi-sense word embeddings and different term weighting schemes will be explored in future.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[3] R. Zhao and K. Mao, "Supervised adaptive-transfer plsa for cross-domain text classification," in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 2014, pp. 259–266.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[5] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[6] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[7] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[8] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, April 2009.

[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[10] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 13, 2008.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR 2013*. ICLR, 2013. [Online]. Available: https://arxiv.org/pdf/1301.3781.pdf

[12] S. Dumais, G. Furnas, T. Landauer, S. Deerwester, S. Deerwester *et al.*, "Latent semantic indexing," in *Proceedings of the Text Retrieval Conference*, 1995.

[13] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1-2, pp. 177–196, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[15] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *The Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[17] A. Mnih and G. E. Hinton, "A scalable hierarchical distributed language model," in *Advances in neural information processing systems*, 2009, pp. 1081–1088.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[19] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 151–161.

[20] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the conference on empirical*

methods in natural language processing (EMNLP), vol. 1631. Citeseer, 2013, p. 1642.

[21] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models." in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 2013, pp. 1700–1709.

[22] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1700–1709.

[23] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1746–1751.

[24] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 655–665.

[25] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in Neural Information Processing Systems*, 2014, pp. 2042–2050.

[26] M. Kusner, Y. Sun, N. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 957–966.

[27] D. Dubois and H. Prade, "An introduction to fuzzy systems," *Clinica Chimica Acta*, vol. 270, no. 1, pp. 3–29, 1998.

[28] Y. Zhao, J. Lam, and H. Gao, "Fault detection for fuzzy systems with intermittent measurements," *Fuzzy Systems, IEEE Transactions on*, vol. 17, no. 2, pp. 398–410, 2009.

[29] Y. Li, W. Liu, Q. Huang, and X. Li, "Fuzzy bag of words for social image description," *Multimedia Tools and Applications*, pp. 1–20, 2014.

[30] J.-P. Mei and L. Chen, "A fuzzy approach for multitype relational data clustering," *Fuzzy Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 358–371, 2012.

[31] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, Aug 2005.

[32] Z. Deng, F.-L. Chung, and S. Wang, "Robust relief-feature weighting, margin maximization, and fuzzy optimization," *Fuzzy Systems, IEEE Transactions on*, vol. 18, no. 4, pp. 726–744, 2010.

[33] W.-C. Tjhi and L. Chen, "Dual fuzzy-possibilistic coclustering for categorization of documents," *Fuzzy Systems, IEEE Transactions on*, vol. 17, no. 3, pp. 532–543, 2009.

[34] I.-J. Chiang, C. C.-H. Liu, Y.-H. Tsai, and A. Kumar, "Discovering latent semantics in web documents using fuzzy clustering," *Fuzzy Systems, IEEE Transactions on*, vol. 23, no. 6, pp. 2122–2134, 2015.

[35] J.-Y. Jiang, R.-J. Liou, and S.-J. Lee, "A fuzzy self-constructing feature clustering algorithm for text classification," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 3, pp. 335–349, 2011.

[36] S.-J. Lee and J.-Y. Jiang, "Multilabel text categorization based on fuzzy relevance clustering," *Fuzzy Systems, IEEE Transactions on*, vol. 22, no. 6, pp. 1457–1471, 2014.

[37] T. Martin, Y. Shen, and B. Azvine, "Incremental evolution of fuzzy grammar fragments to enhance instance matching and text mining," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 6, pp. 1425–1438, Dec 2008.

[38] Z. Wu and C. L. Giles, "Sense-aaware semantic analysis: A multi-prototype word representation model using wikipedia," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[39] J. Li and D. Jurafsky, "Do multi-sense embeddings improve natural language understanding?" in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1722–1732.

[40] R. Zhao and K. Mao, "Topic-aware deep compositional models for sentence classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 2, pp. 248–260, 2017.

[41] ——, "Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space," *IEEE Computational Intelligence Magazine*, vol. 10, no. 3, pp. 30–41, Aug 2015.

[42] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, *Handbook of latent semantic analysis*. Psychology Press, 2013.

[43] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1201–1211.

[44] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 618–626.

[45] J. Blitzer, M. Dredze, F. Pereira *et al.*, "Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification," in *ACL*, vol. 7, 2007, pp. 440–447.

[46] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, 2006, pp. 377–384.

[47] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[48] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[49] E. A. Patrick and F. Fischer, "A generalized k-nearest neighbor rule," *Information and control*, vol. 16, no. 2, pp. 128–152, 1970.

Rui Zhao received the BEng in Measurement and Control from Southeast University, Nanjing, China, in 2012. He is currently pursuing the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

His current research interests include machine learning and its applications on text mining and machine health monitoring..

Kezhi Mao received the BEng degree from the Jinan University, Jinan, China in 1989, the MEng degree from Northeastern University, Shenyang, China in 1992, and the PhD degree from the University of Sheffield, Sheffield, U.K. in 1998. He was a lecturer at Northeastern University from March 1992 to May 1995, a research associate at the University of Sheffield from April 1998 to September 1998, a research fellow at the Nanyang Technological University, Singapore, from September 1998 to May 2001, an assistant professor at the School of Electrical and Electronic Engineering, Nanyang Technological University from June 2001 to Sept 2005. He has been an associate professor since October 2005. His areas of interests include computational intelligence, pattern recognition, text mining, and knowledge extraction, cognitive science, and big data and text analytics.