A Modified Fuzzy Bag-of-Words classification method for Document Clustering

R.REVATHI., RESEARCH SCHOLAR,DEPARTMENT OF COMPUTER SCIENCE, PADMAVANI ARTS AND SCIENCE COLLEGE FOR WOMEN,SALEM

D.M.CHITRA.,ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE, PADMAVANI ARTS AND SCIENCE COLLEGE FOR WOMEN,SALEM

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. 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 sdocument classification datasets in comparison with document representation learning methods have shown that our methodsFBoW and FBoWC achieve the highest classification accuracies.

Keywords: *Document clustering, Bag-of-Words, Fuzzy Bag-of-WordClusters(FBoWC), Fuzzy Bag-of-Words (FBoW).*

I. INTRODUCTION:

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems [Rij79, Kow97] and as an efficient way of finding the nearest neighbors of a document [BL85]. More recently, clustering has been proposed for use in browsing a collection of documents [CKPT92] or in organizing the results returned by a search engine in response to a user's query [ZEMK97]. Document clustering has also been used to automatically generate hierarchical clusters of documents [KS97].

Object categorization through Bag of Words model is one of the most popular representation methods for object categorization. Bag of Words (BoW) approach has shown acceptable performance because of its fast run time and low storage requirements [4]. The key idea is to quantize each extracted key point into one of visual word, and then represent each image by a histogram of the visual words. For this purpose, a clustering algorithm like K-means is generally used for generating the visual words. Appropriate datasets are required at all stages of object recognition research, including learning visualmodels of object and scene categories, detecting and localizing instances of these models in images, and evaluating the performance of recognition algorithms. Image databases are an essential element of object recognition research[4]. They are required for learning visual object models and for testing the performance of classification, detection, and localization algorithms.A common and effective approach to document display is the bag-of-words (BoW) model. The BoW model assigns a vector to a document as d = (x1; x2; ::; xl), where xi denotes the normalized number of occurrences of the ithbase term and l the size of the collection of bases. It should be noted that the base terms are the high frequency words in a corpus, and the number of base terms or the dimensionality of BoW vectors is less than the size of the vocabulary [7], [8], [9], [10]. BoW is a simple but effective way to map a document into a fixed-length vector. However, the mapping function in the BoW model is hard or binary, i. H. The crisp binary relationship that represents only the presence or absence of a base term in the document. The hard-mapping function has several limitations. First, the learned vector is extremely sparse because a document contains only a very small portion of all base terms. Second, the BoW representations can not effectively capture the semantics of documents because semantically similar documents with different word uses under BoW map to very different vectors.

In this work, we suggest fuzzy BoW models to learn more dense and robust document images that code more Semantics. To overcome the limitations of the original BoW Model as discussed above, we propose to replace the original Hard mapping through a fuzzy mapping, and develop the fuzzy BoW (FBoW) model. Unlike BoW, which works exactly Word matching to basics, FBoW introduces vagueness in the correspondence between words and the basic concepts. Fuzzy Mapping allows a word semantically similar to a basic term be activated in the BoW model. The membership function a basic term in the FBoW model assigns membership Values to words according to their semantic similarity to the Basic runtime. The intuition behind such a membership function lies in the fact that the affiliation values should be proportional the semantic similarity between the word in documents and the basic concepts. In our proposed model, word embeds Technique is introduced to evaluate the semantic similarity.

Trained on a large corpus word embeds code word Meanings in vectors and thus the semantic similarity between Two words can be conveniently evaluated using cosine Similarity between the corresponding word embeds [11].

The cosine similarity measure can be interpreted as the degree a word semantically appropriate to another word. To illustrate the comparative advantages of our proposed blurred BoW= the original BoW, Fuzzy BoW is applied to the same toy Example, as shown in Figure 1 (b). Due to the assumed fuzzy Mapping, bank in set d1 and huskies in set d2= can be assigned to the basic table or the dog, and their values are proportional to the semantic similarity.

The blurred BoW generates the following vectors for the two Toy sets: d1 = (1; 0: 7; 0: 8; 1) and d2 = (0: 7; 1; 1; 0: 8), the FBoW model generates two similar vectors for two semantically similar sentences. Based on FBoW, a fuzzy the Bag-of-WordClusters model (FBoWC) is proposed. Different from the Fuzzy BoW (FBoW) model, whose basic terms are single words, FBoWC uses clusters of words as basic terms, 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 FBoWCmean, FBoWCmax or FBoWCmin.

II. RELATED WORK

1.Fuzzy based Multiple Dictionary Bag of Words for Image Classification

In this paper Bag of Words model has been implemented for visual categorization of images using Harriscorner detector for extracting features and Scale Invariant Feature descriptor (SIFT) for representing theextracted features. After obtaining local features called descriptors, a codebook is generated to representthem. The codebook is a group of codes usually obtained by clustering over all descriptors. Clustering isthe process of assigning a set of objects into groups so that the objects of similar type will be in onecluster. Clustering can be classified as hard clustering and soft clustering. The performance of BoWdepends on the dictionary generation method, dictionary size, histogram weighting, normalization, anddistance function. In this paper the method of generation of the dictionary of visual words is beingfocused. A novel method, Multiple Dictionaries for BoW (MDBoW) [18] using soft clustering algorithmFuzzy C-means, that uses more visual words is implemented. This method significantly increases theperformance of the algorithm when compared to the baseline method for large scale collection of images.Unlike baseline method, more words are used from different independent dictionaries instead of addingmore words to the same dictionary. The resulting distribution of descriptors is quantified by using vectorquantization against the pre-specified codebook to convert it to a histogram of votes for codebook centers.K nearest neighbor algorithm (KNN) is used to classify images through the resulting global descriptorvector.

2.A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification

We propose a fuzzy similarity-based self-constructing feature clustering algorithm, which is an incremental feature clustering approach to reduce the number of features for the text classification task. The words in the feature vector of a document set are represented as distributions, and processed one after another. Words that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is created for this word. Similarity between a word and a cluster is defined by considering both the mean and the variance of the cluster. When all the words have been fed in, a desired number of clusters are formed automatically. We then have one extracted feature for each cluster. The extracted feature corresponding to a cluster is a weighted combination of the words contained in the cluster. Three ways of weighting, hard, soft, and mixed, are introduced. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial-and-error for determining the appropriate number of extracted features can then be avoided. Experiments on real world data sets show that our method can run faster and obtain better extracted features than other methods.

3. Document Representation Learning

As mentioned in the Introduction, document representationis the keystone for various text mining and NLP tasks. Themost established Bag-of-words (BoW) model is often criticizedfor its extreme sparsity, high dimensionality and inabilityto capture semantics. Some works have been proposed toimprove BoW model including latent semantic analysis (LSA)and topic models [12], [13], [14]. These models transformthe BoW representation into low-dimension representations tocapture the latent semantic structure behind documents. InLSA, singular value decomposition (SVD) is applied to theoriginal BoW representation to obtain a new representation,where each new latent dimension is a linear combination ofall original dimensions. In topic models including probabilisticlatent semantic analysis [13] and latent dirichlet allocation[14], probability distributions are introduced to describe wordsand the generation process of each word in a document. Theassumption behind topic models is that word choice in adocument will be influenced by the topic of the documentprobabilistically. However, in these models, the derived latentdimension lacks semantic interpretation. For example, LSAregards a latent dimension as a linear combination of alloriginal terms in vocabulary, which is counter intuitive becauseonly a small part of the vocabulary is actually relevant to acertain topic. In addition, these two approaches both utilizethe word occurrence of documents to perform dimensionalityreduction. However, the occurrence statistics may not beable to capture the true semantic information underlying adocument. Different from BoW model and BoW-enhancedmodels such as LSA and topic models that employ exact wordmatching and hard mapping, our proposed FBoW and FBoWCmodels adopt semantic matching and fuzzy mapping to projectthe words occurred in documents to the basis terms. In ourproposed fuzzy BoW models, word embeddings is introducedto help evaluate semantic similarity between words. Sinceword embeddings are trained on very large-scale corpus, itis believed that the captured similarity information is moreaccurate and general than that extracted from word occurrencestatistics underlying a document in previous BoW-based approaches.In addition, our proposed fuzzy BoW models canalso be used in conjunction with the LSA method to reducethe dimensionality of the FBoW representation.

III. PROPOSED WORK

Our proposed fuzzy Bag-of-Words models are presented. Since the fuzzy membership function is based on word embedding, we begin with a brief review of the word embeddings.

A. Embed Words

The core idea behind word embedding is the assignment of sucha dense and low-dimensional vector representation for everyonethat semantically similar words are close to each otherin vector space. The merit of the word embedding is that thesemantic similarity between two words can be convenientbased on the cosine similarity measure betweencorresponding vector representations of the two words. In thatpopular word embeds word2vec [15], [11], [18], a two-layered version the language model of the neural network has been developed to learnVector representations for each word. The word2vec framework contains two separate models including continuous Bag byWords (CBoW) and skip-grams with two opposite training goals.CBoW tries to predict a word with the surrounding wordswhile Skip-Gramm tries to predict a window of words given asingle word. Because of its surprisingly efficient architecture andunmonitored training protocol, about which word2vec can be traineda large unannotated body efficiently. word2vec is capableto encode meaningful linguistic relationships between wordsin learned words embedding. Usually the cosine resemblance measure between word embeds is used to measure thatsemantic similarity between two words:

cos (wi; wj) =wi? W Jkwikkwjk(1)where wi and wj denote word embedding of two words wiand wj respectively. The cosine similarity measure is positivewhen the words are close to each other and negative ifThe words have the opposite meaning. The measure is zero under aCouple of two completely random words. To give an illustration,book the top 5 similar words to two sample words andPupils and their cosine similarity values are given in the table. In our proposed FBoW models, cosine similarity measurebased on Wohltps: are used to construct fuzzy membership functions for mappingthe words in documents to basic terms. It should be notedthat our proposed models do not take into account the polysemic problem,since the individual prototype word embeds are used as input.There are some ambiguous words embedded in theLiterature that is disambiguation process of every word sensequite challenging and therefore hinders the application ofMulti-sense word embedding [38], [39], [40]. In addition,Documents usually contain many words, the effects of neglectPolysemy is less important than at the word or sentence level.However, it still makes sense to look into Multi-SenseWords embedded in our proposed FBoW models in the futurejob.

B. Modified Fuzzy Bag-of-Words Model

First, some accepted notations in our proposed methodsare introduced. Let D = fw1; :::; wvg is the vocabularyall words that are present in the body text, and vthe vocabulary size. W 2 Rv? D denotes a well-trained wordEmbedding matrix, where its ith row wi 2 Rd thed-dimensional word embedding for word wi. Every documentIn the text, the corpus is represented by a BoW vector whose elements indicate the number of occurrences of basic terms in the document.In a large corpus only the top 1 high-frequency wordsare usually chosen as basic terms in the BoW model for reductionthe Sparsity and dimensionality in BoW representations, andthe BoW basis terms T = ft1; :::; tlg is therefore a subset ofthe corpus vocabulary.Traditional BoW representations map documentsin vectors by exact match of the words in the documentsto the basic concepts. Exact word match is equivalent to perform a hard or sharp assignment. If a word w matchesa basic term ti, is the output of the sharp mapping function 1,and is zero otherwise.Fuzzy Membership Function: To address the problem causedby exact word matching in BoW, we propose to use semanticmatching, which matches two words based on semantic similarity.

Representation Learning:

Here, the fuzzy membership function is used to count the number of occurrences of bases in a document. For a document, the FBoW model representation with z = [z1; z2; :::; zl], where thei-tes element zi is the sum of the degrees of membership where all words semantically agree with the ith base term, i.

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

W denotes a set of all words in the document, ti is the i-th base term, and xj denotes the number of occurrences of wj. It should be noted that ci is a control parameter defined by different weighting schemes in the BoW model. For example, ci = 1 if the count scheme is assumed while ci is the reverse document frequency when the TF-IDF is accepted. For the sake of simplicity, we take the counting scheme as our weighting scheme and ci is set to 1. As in Eq. (2) and (4) the BoW model can be considered a special case of our proposed fuzzy model. In BoW, xi is only determined by the term frequency, which corresponds to the use of the hard-membership function. In the following, a matrix formulation of the above fuzzy BoW model is presented.

C. Fuzzy Bag-of-WordClusters Model

It is well acknowledged that BoW model has three limitations,including sparsity, high dimensionality, and lack ofcapability to encode high-level semantics. The fuzzy BoWmodel developed in Section III-B addressed the issues ofsparsity and semantics, but the high dimensionality problemremains. Actually, the high dimensionality also means redundancy.This is the reason why BoW is often combinedwith LSA to reduce dimensionality. Certainly, FBoW canalso be combined with LSA to reduce the dimensionality andredundancy of FBoW representation. In this study, we proposea plausible method to solve the high dimensionality and redundancy problem of FBoW model.

Algorithm 1 Fuzzy Bag-of-Words Frameworks

Input: a text corpus with n documents; the vocabulary D andits corresponding word embeddings matrix W 2 Rvd,where v is the vocabulary size and d is the dimensionalityof word embeddings. Required dimensionality fordocument vectors: l

Output: learned document vectors for the corpus: Z 2 Rnl

1: Based on the corpus vocabulary D, obtain data matrixX 2 Rnv that each row x 2 Rv is the i-th documentvector whose j-th element is the number of occurrence ofword wj in the corresponding document, as shown in Eq.(6);

2: if FBoW is performed then

3: Based on term frequencies over the corpus, select thetop-l words with highest frequency as our models' BoWspace T and the corresponding word embeddings areobtained as WT 2 RLd;

4: Construct transformation matrix H based on W andWT using Eqs. (3) and (7);

5: else if FBoWC is performed then

6: Apply K-means algorithm to cluster words based onword embeddings matrix W by setting the numberof clusters to l. Then, the embeddings of words ineach clusters are obtained and the cosine similaritybetween these clusters' words and word in documentsare computed as shown in Eq. (9);

7: Construct transformation matrix H based on W andqti using Eqs. (8) and (7);

8: end if

9: Calculate learned data matrix Z according to Eq. (5), which can be used to represent the corpus.

10: return Z

D. Relationships with Previous Text Representation Methods

Word embeddings are introduced to capture the semanticrelationships among words, and the derived semantic similarityand fuzzy mapping are then incorporated into the originalBoW model. As a result, the learned document representationsare more dense and able to capture more semantic information. In this subsection, we analyze the connections between ourproposed FBoW frameworks including FBoW and FBoWCwith two typical text representation learning models includingdimensionality reduction methods and a deep compositionmodel: convolutional neural network (CNN).Relationships with Dimensionality Reduction: Dimensionalityreduction techniques seek to reduce the rank of vectors.Through dimensional reduction, sparse and high-dimensionbetween clusters and words. It is noted that a high similarity measure denotesa small distance shown in the Figure.BoW vectors can be transformed into dense and lowdimensionalones, which in turn boosts the performance ofsubsequent tasks such as classification, information retrieval,etc. Some models including latent semantic analysis (LSA)and random projection (RP) are applied extensively in manytext mining applications [12], [41]. LSA and RP are lineardimensionality reduction methods, and the key issue is to findthe mapping matrix. For LSA, the mapping matrix is learnedvia maximizing the preservation of variance of the originalfeature space. Since the input information for LSA can be regardedas occurrence statistics between documents and words,LSA may fail to model the true semantic information and theresulting dimensions may not have interpretable meaning innatural language [42]. For RP, the mapping matrix is generatedrandomly. Some experimental results have shown that RP canachieve a significant speedup in computation time will littledistortion of pairwise information of data. However, withoutdata-based parameter tuning, RP may not capture the semanticinformation underlying the natural language. In FBoWC representations, each dimension corresponds toword clusters which are subsets of the entire vocabulary.By contrast, each dimension in LSA and RP is a linearcombination of all words in the vocabulary. As the mappingmatrix of FBoWC in Eq. (7) directly measures the semanticsimilarity between words and basis terms based on wordembeddings, it can capture high quality semantic information.In addition, word embeddings are pre-trained and publiclyavailable, the computational cost is not a potential problemfor FBoWC.

IV. EXPERIMENTS

In this section, we use document categorization tasks toevaluate the performance of our proposed Fuzzy Bag-of-wordsmodels.

A. Descriptions of Datasets

The task of document categorization is to assign a class label or category to a document. Seven real-life datasets areused in the experiments.20Newsgroups is a collection of nearly 20,000 newsgroupdocuments, which is organized into 20 different classes. Here,we adopted the version of 20 Newsgroups (20NG) sortedby the removal of duplicates and some headers1. The wholecorpus has 18846 documents, and the vocabulary size is 32716,excluding the removed words whose document frequenciesare less than five. Actually, the removal of low frequencywords were performed for all the seven datasets used inthe experiments. We followed predefined training and testingsplitting. The statistics of 20NG are given in Table II.Reuters 14 and Reuters 8 were both generated from aclassical corpus Reuters-21578 containing newswire articlesand Reuters annotations2. The whole collection has 21,578documents, which are categorized into 90 classes. Since somecategories have only a few documents, we created two datasetscontaining 14 and 8 most frequent classes, respectively. Thepredefined training and testing splitting was adopted. Thestatistics of these two datasets Reuters 14 and Reuters 8 canbe found in Table II.

Amazon 6 is a collection of Amazon reviews for productsof six categories. This 1 dataset was originally published for sentimentanalysis [44], but we used it for categorization. The dataset has been kindly provided at http://qwone.com/jason/20Newsgroups/2The dataset has been kindly provided at http://csmining.org/index.php/six categories are cameras, laptops, mobile phone, tablets, TVsand video surveillance, in which the largest sample numberis 6736 under cameras and the smallest sample number is881 under tablets. To make the dataset more balanced, werandomly selected 1500 samples from categories with morethan 1500 reviews. The corpus used in our experiments has8083 reviews with a vocab size of 10790. The details areshown in Table II.For AE, WMD, FBoW and FBoWC models, the same wordembeddings were used. We utilized the pre-trained word2vecvectors published by Google6. These word embeddings weretrained on a Google News corpus (over 100 billion words)and have a dimensionality of 300. For all the seven documentcategorization tasks, we further fine-tuned the pretrainedword embeddings over the specific dataset. Since AEaverages embeddings of all words, the dimension of documentvector learned by AE is the same as the dimension of wordembeddings, Which is 300. The other settings of WMD methodwere the same as that reported in its original paper [26].

Linear SVM [48] was applied to the document representationslearned by the above mentioned approaches. In linearSVM, we searched the best regularization parameter C fromf0:001; 0:01; 0:1; 1;10; 100g. Since WMD can only derive document distanceinstead of document representations, document classificationbased on WMD used the kNN decision rule [49]. The searchingrange of the neighborhood size k is f1; 3; : : : ; 19g.B.

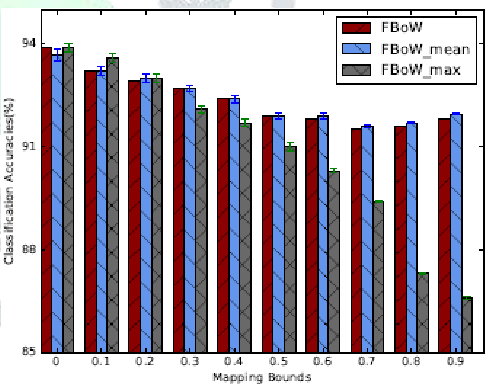

Fig. 4. Performance of FBoW, FBoWCmean and FBoWCmax for differentmapping bounds:ƛ.

V. CONCLUSION

In this work, we have proposed Fuzzy Bag-of-Words modelsincluding FBoW and FBoWC to address issues of sparsityand lack of high-level semantics of BoW representation.

Wordembeddings are utilized to measure semantic similarity amongwords and construct fuzzy membership functions of basisterms in BoW space over words in the task-specific corpus.Since word2vec embeddings can be trained over billions ofwords, word embeddings adopted in our methods are ableto capture high-quality and meaningful semantic informationthat are not contained by the task-specific corpus alone. Todetermine basis terms in BoW space, FBoWC utilizes wordclusters, while FBoW directly regards high term-frequencieswords as original BoW does. The adoption of word clusters inFBoWC can reduce feature redundancy and improve featurediscrimination. Three different measures have been designedto evaluate similarity between clusters and words, and threecorresponding variants of FBoWC models as FBoWCmean,FBoWCmax and FBoWCmin have been developed. The performanceof our approaches has been experimentally verifiedthrough seven multi-class document categorization tasks. As anext step work, document structure or word order informationwill be considered in document representation learning. Inaddition, the effects of multi-sense word embeddings anddifferent term weighting schemes will be explored in future.

REFERENCES:

[1] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A reviewand 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 classificationfrom labeled and unlabeled documents using em,” Machinelearning, vol. 39, no. 2-3, pp. 103–134, 2000.

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

[4] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundationsand 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 documentclustering 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 semanticenhancedmarginalized denoising auto-encoder,” IEEE Transactions onAffective Computing, vol. PP, no. 99, pp. 1–1, 2016.

[7] K. Sparck Jones, “A statistical interpretation of term specificity and itsapplication 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 termweighting methods for automatic text categorization,” IEEE Transactionson Pattern Analysis and Machine

Intelligence, vol. 31, no. 4, pp. 721–735, April 2009.

[9] G. Salton and C. Buckley, “Term-weighting approaches in automatictext 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, “Interpretingtf-idf term weights as making relevance decisions,” ACM Transactionson Information Systems (TOIS), vol. 26, no. 3, p. 13, 2008.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimationof 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. Deerwesteret al., “Latent semantic indexing,” in Proceedings of the Text RetrievalConference, 1995.

[13] T. Hofmann, “Unsupervised learning by probabilistic latent semanticanalysis,” 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 probabilisticlanguage model,” The Journal of Machine Learning Research,vol. 3, pp. 1137–1155, 2003.

[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, andP. Kuksa, “Natural language processing (almost) from scratch,” TheJournal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.

[17] A. Mnih and G. E. Hinton, “A scalable hierarchical distributed languagemodel,” 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 Methodsin Natural Language Processing. Association for ComputationalLinguistics, 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 overa 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 innatural language processing (EMNLP), 2013, pp. 1700–1709.

[22] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, andY. Bengio, “Learning phrase representations using rnn encoder-decoderfor statistical machine translation,” in Proceedings of the conference onempirical 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 NaturalLanguage Processing. Association for Computational Linguistics, 2014,pp. 1746–1751.

[24] K.S.Sujathaa P. Keerthana b S. SugaPriya b E.Kaavya b B.Vinod c “Fuzzy based Multiple Dictionary Bag of Words for Image Classification” in Elsevier 2011.

[25] Michael Steinbach George KarypisVipinKumar “A Comparison of Document