Documents Classification Based On Deep Learning

Aalaa Abdulwahab, Hussein Attya, Yossra Hussain Ali

Abstract : Every day a large number of digital text information is generated, the effectively searching, exploring and managing text data has become a main task. The Text Classification has areas in Sentiment Analysis, Subjectivity/Objectivity Analysis, and Opinion Polarity the Convolution Neural Networks (CNN's) has a good performance and accuracy therefore it gained special attention. Latent Dirichlet Allocation (LDA) is a classic topic model that able to extract latent topic from high dimensions and large-scale multi-class textual data(large data corpus).In this paper, we present a comparison among CNN ,traditional LDA and modified LDA with TF-IDF algorithm to classify a large pool of documents as a data set, it's 20 news group. Experiment results show that the accuracy performance of CNN (94%) is better than the modified LDA approach (74.4%) and traditional LDA (60%).The time to perform dataset classification by Traditional LDA is 4.04m, Modified LDA is 3.02m was less than time of CNN model 11.52m.

Keyword: Topic modeling, LDA, CNN,TF-IDF, Deep learning.

1.INTRODUCTION

Text classification is a supervised learning task that is known as the identification of categories of new documents based on the probability proposed by a specified training corpus of previously labeled documents. while the amount of textual information of new documents online is increases, the management to classify them precisely becomes more difficult, because the ability to effectively recover the correct categories for new documents depends on the amount of labeled documents already exists for reference [1]. Traditional classification uses information retrieval techniques such as continuous bag of-words or TF-IDF to represent the documents. These techniques are used widely in natural language processing (NLP) that assist in provided a simplified representation of documents through different features, the goal of natural language processing (NLP) is to process text by computers in so as to analyze it, to extract information and to represent the same information differently[2]. In the continuous bag-of words ignoring the grammar and word order but multiplicity is kept and TF-IDF is a statistical method with a high accuracy and reflect the importance of a word to a particular document in a collection of documents[1]. The principle of TF-IDF technique if a word appears in the document at a high frequency and is rarely exists in other documents, that is mean the word has a good class distinction and is suitable for the classification. Recently, researchers show a big interest in exploring new text representation models for enhancing the efficiency and accuracy of text processing. The theoretical idea of the topic model is that document is a mixture of different topics, each containing multiple terms of word distribution. Topic model obtains the collection of semantic related topics hidden in the document by common word information in the document. It transforms the document from word space to topic space and expresses the document in a lower dimension space. The topic model originated from latent semantic indexing (LSI), and then the topic model evolved into a variety of forms, particularly latent models based on Dirichlet allocation (LDA)[3]. LDA, an unsupervised generative probabilistic method for modeling a corpus, is the most commonly used topic modeling method[4]. In[2014] LDA algorithm is used for topic text classification by adding topic-category distribution parameter to LDA, which can make the document created from the most relevant category. Gibbs sampling is used to perform approximate inference, and the results of the experiments in two datasets show the efficacy of this method[5]. In[2015] LDA's parameterization of "topics" as categorical distributions over ambiguous word types are replace with multivariate Gaussian distributions on the embedding space, this allow the model to group words that are a priori known to be semantically related into topics[6]. In[2016] A general framework for short text classification is presents by learning vector representations of both words and hidden topics together, LDA is used to build topic model that classify a large-scale external data collection named "corpus" which is topic consistent with short texts to be classified[7]. In [2018] proposed probabilistic generative model based on LDA, it called ED-LDA algorithm model .It is include the environmental data relationship and structure that help in find out the useful information and analysis to mine the relationship between users and their posted environmental data on social network to better understand the meaning of the data for environmental management [8]. In [2019] developed the dynamic latent Dirichlet allocation (D-LDA) and word embedding that are combine in a generative model of document called the D-ETM. In the developed model each word with a categorical distribution, the parameter of which is given by the internal product between the word embedding and the embedding representation of the topic at a given time step [14]. In the last few years, deep learning has led to very good performance on a variety of issues, such as speech recognition, visual recognition and natural language processing. Convolution Neural Networks are among the different types of deep neural networks. Research on convolution neural networks has been developed quickly and achieved state-of - the-art results on various tasks [10]. In [2014] show that a simple CNN with little hyper parameter tuning and static vectors achieves excellent results on multiple benchmarks and Propose a simple modification of the architecture to allow the use of both specific and static vectors[11]. In[2015] CNN studies on text categorization to exploit the 1D structure (i.e. word order) of text data for accurate prediction by directly applying CNN to high-dimensional text data which leads to the direct learning of small text areas for use in classification[9].In [2017] proposed text classification method named document matrix convolution neural networks (DMCNN) this model is based on the n-dimensional word embedding obtained in advance, taking each entry of the word embedding as a pixel, and converting the text into a 2-dimensional document matrix

(DM) according to the proposed method [12]. In [2018] proposed a convolution neural network model with part-of-speech tagging and word double embedding to deal with text multi-classification problem. The chunk max pooling is added to the sampling layer for down sampling to enhance the ability of the feature extraction. And in pre-processing text, the knowledge base of the word segmentation is expanded to the content of the data set to improve the accuracy of the pre-processing text model[13]. In [2019] Propose a method of features extraction to better investigate the text space. Specifically, create a model composed of an attention mechanism a convolution neural network and a recurrent neural network to extract multi-view characteristics[20]. In this paper, we present how to implement multi-text classification (sentence classification) on large number of documents with Convolutional neural network ,traditional LDA model and Modified LDA model .We use TF-IDF algorithm to modify the traditional LDA model, then comparing the result of three models .Experiment results prove the effectiveness of CNN model than other two models. The outline of this work is shown as follows. In Sect. 2, a brief introduction of LDA model and TF-IDF algorithm is explained. A specific description of the proposed approach is followed in next section. In Sect. 4, a brief introduction of CNN model. The experiment results are shown in Sect. 5, and conclusions are made in the final section.

2- LATENT DIRICHLET ALLOCATION

The LDA is a generative model, suggested by Blei in 2003, which has a powerful theoretical structure and experiments to be used efficiently in many text classification applications. This model was used to discover a hidden thematic structure in large collections of documents[5]. According to this model assumes the document is composed by a set of topics, which words are grouped into “topics” using vector dimension reduction. This also helps put words and documents map into a lower dimension space, so as to make a good performance in latent topic extract process[15]. A graphical model for LDA is shown in Figure1:

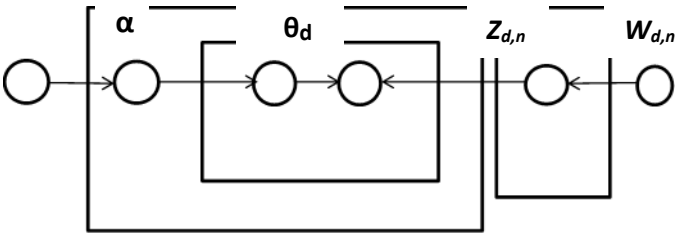


Figure1:*Graphical model of LDA*

α , η : are proportion parameter and topic parameter, respectively. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary. The topic proportion for the d th document are θ d , where $\theta_{d,k}$ is the topic proportion for topic k in document d . The topic assignments for the d th document are Z_d , where $Z_{d,n}$ is the topic assignment for the nth word in document d . Finally, the observed words for document d are w_d , where $w_{d,n}$ is the nth word in document d , which is an element from the fixed vocabulary. With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables:

$P((\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_i^k p(\beta_i) \prod_{d=1}^D p(\theta_d) (\prod_{n=1}^N p(Z_{d,n}|\theta d) p(W_{d,n}|\beta_{1:K}, Z_{d,n}))$

Notice this distribution specifies a number of dependencies. The topic assignment $Z_{d,n}$ depends on the per-document topic distribution θ_d ; and the word $W_{d,n}$ depends on all of the topics $\beta_{1:K}$ and the topic assignment $Z_{d,n}$. Basic steps of the traditional LDA model are: we read the corpus and preprocess the set of documents in the corpus at first, then we use LDA model to classify text of training set and evaluate the model with test set.

Data preprocessing:

The first step of preprocessing is text-cleaning. The goal of text cleaning is to simplify the text data like by splitting it into words and handling punctuation. Because we cannot use raw text straight to fitting a machine learning or deep learning model. In this experiment, there are several steps in text cleaning:

1. Tokenization: Split the text into sentences and the sentences into words. Remove punctuation and lowercase the words.
2. Words that have fewer than 3 characters are removed.
3. All stop words are removed.
4. Lemmatizing : words in third person are changed to first person and verbs in past and future tenses are changed into present
5. Stemmed : words are reduced to their root form.

The next step , convert text into dictionary that each document represent by vector ,each vector represent word and its Id. Bag of word method is used to calculate the co-occurrence of each word in the dictionary and represent them in vectors (Id of word, no. of occurrence) then classify them by LDA model. The algorithm(1) of traditional LDA model is shown below:

Algorithm (1) traditional LDA model:

Input: Labeled Corpus(collection of documents) .

Output: List of topics.

Begin:

Step1: preprocessing text : For all documents (D) in corpus (C) do

1.1 Apply Tokenization to the text.
1.2 if words < 3 characters removed.
End if
1.3 Remove stop word .
1.4 Apply Lemmatization and Stemming to the text.
Next

Step2: For all the preprocessed documents:

2.1 Build a Dictionary to save word (key) and index (value: words occurrence)
2.2 For sentence in corpus
2.3 If token not in wordfreq. Key
2.4 Wordfreq[token]=1

Else
2.5 Wordfreq[token]=+1
Next
Next

Step 3: For all documents in corpus
Convert documents and words as vectors of features (Bag of Word)

Step 4: Build LDA model(bow-corpus, dictionary, topics no.)
Step 5: Return List of topics for each document, and for each topic get some of the most relevant words with the topic.
End.

3-TF-IDF ALGORITHM

TF-IDF is a well algorithm used in information retrieval field. In recent years, researchers have used TF-IDF algorithm to calculate the weight of features in documents and achieved good results[16] . $W = TF \times IDF = TF \times 1/DF$ (1)

TF is the frequency of word T in document D which is used to calculate the capability of the word to represent the document. The inverse of the frequency of the document D containing the word T in the corpus is IDF, that is used to calculate the capability of the word to recognize the document.

$$IDF = \log N/Ni \quad (2)$$

Where: N refer to the all number of documents in the dataset or class . Ni refer to the number of documents where word i appears in the collection of documents. If the frequency of a word is high in its own document but low in other documents, this word is assigned to a high weight and has a strong ability to distinguish it from other documents and[17].

4- MODIFIED LDA

To improve the performance of traditional LDA we use the TF-IDF algorithm ,the system called Modified LDA. The main idea of this method is the words in the traditional LDA model does not have weights and by TF-IDF algorithm the words have a weight in the training model and has the ability to express the importance of the word to the current document and to distinguish it from other documents. Calculate TF-IDF for the step (4) in the previous LDA algorithm as shown in the algorithm(2):

Algorithm (2) Modified LDA:
Input: Bag of Word (corpus vectors)
Output: TF-IDF Bag of Words model
Begin:
Step1: For all documents in Bag of Word do
1.2 For all words in Bag of Word do

1.3 Compute term frequency (TF).

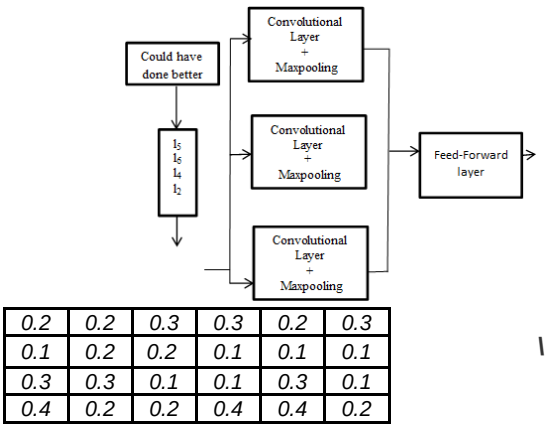
1.4 Compute inverse document frequency (IDF) by using eq(2).
1.5 Compute W=TF*IDF // compute weights of words in BOW by using eq(1).
Next
Next
Step2: Each feature is associated with its weight (TF-IDF) determines the weight and importance of the feature in the document
End

5-CONVOLUTIONAL NEURAL NETWORK:

CNN is multistage trainable Neural Networks architectures sophisticated for classification tasks and it is choose for classification tasks like sentence classification and sentiment classification since sentiment is usually specified by some key phrases. It is a feed-forward neural network with convolution layers interleaved with pooling layers, where the top layer performs classification using the features generated by the layers below in hierarchical architectures[18]. Convolution Neural Network includes a sequence of filters of varying dimensions and shapes that convolve (roll over) the original sentence matrix to decrease it to further low-dimensional matrices. The down sampling method used in the convolution neural network is max pooling. CNN uses an activation function to help it work in the kernel (i.e. high dimensional neural processing space) [9][11].

6- CNN FOR TEXT CLASSIFICATION

The components of a sentence (the words) must to be encoded before input to the CNN therefore a vocabulary may be use. The variability of the documents length depends on the amount of words in the document that need to be addressed , as CNNs need constant input dimensionality. For this purpose, the padding technique is used by filling the document matrix with zeros in order to achieve the maximum length of all documents in terms of dimensionality[18]. Next step, the encoded documents are converted into matrixes for which each row corresponds to a single word. The produced matrices pass through the embedding layer where each word (row) is converted into a low-dimensional representation by a dense vector. The procedure then continues following the standard CNN methodology[19]. The process of CNN text classification is illustrate in figure 2.



0.2	0.2	0.3	0.3	0.2	0.3
0.1	0.2	0.2	0.1	0.1	0.1
0.3	0.3	0.1	0.1	0.3	0.1
0.4	0.2	0.2	0.4	0.4	0.2

Figure2: *Process of CNN for text classification*

The implemented Convolution Neural Network model has three layers with a single channel model for output. The first layer is Embedding layer which convert the words to its respective embedding vectors by using word embedding learned Glove method. Second layer is the convolution layer(2D) in which main processing of the model occur, the

predefined filters (3,4,5) roll over the sentence matrix and reduce it into low dimensional matrix. The output feature of the convolution layer is down sampling by Maxpooling layer in order to get more features reduction and use the Relu as activation function. Softmax layer is third layer (output layer) which is calculating loss function.

7.EXPERIMENT RESULTS:

We evaluate the performance of CNN that word embedding is used against the Bow approach which is implement text classification in Latent Dirichlet Allocation and modified LDA model by experiments in this section. We used 20newsgroups as dataset to train and test the three models .It has 20 different classes each one has approximately 1000 documents, so there is 20,000 documents.

1.CNN Model Results

We experimented CNN model to check the flexibility of it, List of parameters is as follow: Batch size=30 and number of filters=512 are determined according to the dataset. Embedding dimensions=100 is defined according to the maximum sentence length. Dropout probability=0.5 is used to down sample the output.

Table1: *The classification performance of accuracy and time using CNN*

Epoch no.	Accuracy	Time
1	94.7	11.52
2	94.8	22.5
3	94.86	32.46
4	94.88	43.85

.

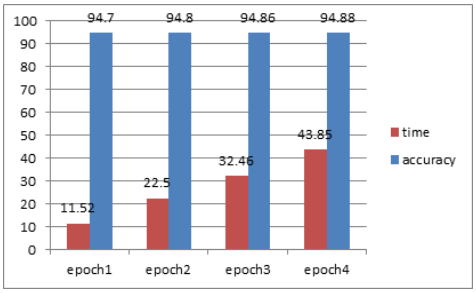


Figure2: *The classification performance of Accuracy and Time*
using CNN .

The above chart explain the relationship between the epoch(mean the entire data set is train on the model at once) , the accuracy and time to implement the classification . At the training and testing of CNN model the increasing in the number of epoch lead to slightly increasing in the accuracy from 94.7% to 94.88% this mean the epoch number has simple effect on the accuracy of CNN model ,it is still higher but its effect on the time of train and test to perform the classification which increased from 11.52 to 43.85 minute .

2.LDA model Results

The training LDA parameters are: passes=10 and number of topics=10,20,30 and 40 and the Dirichlet hyper-parameter β =0.01,α=0.1 and test document number(1000) are used to evaluate the two modules of LDA. The experiment compares the influence of different topic numbers on traditional LDA and modified LDA and the result of time and topics accuracy is shown in Fig. 3 ,4 . The results show that the modified LDA has a better performance than traditional LDA model. When topic number is 10, we get the best performance with topics accuracy74.4% and time 3.02m as shown in table 2:

Topics no.	LDA		Modified LDA	
	Time	Topics accuracy	Time	Topics accuracy
10	4.04	60.8	3.02	74.4
20	4.45	23.01	2.94	39.3
30	4.5	30.15	2.9	68.4
40	5.3	50.28	3.13	53.8

Table2: *The classification performance of Time and Topics accuracy*
using LDA and modified LDA .

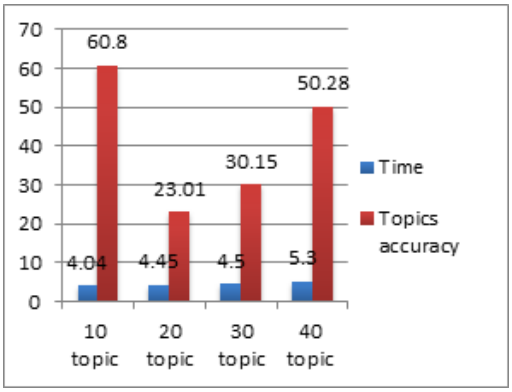


Figure 3: *The classification performance of Time and*

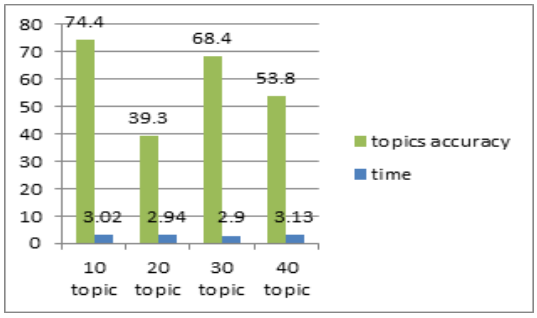


Figure 4: *The classification performance of Time and Topics accuracy using LDA*
Topics accuracy using modified LDA

The figures show the effect of topics number generated during document classification by the modified LDA model that achieve good accuracy 74.4% when the model generate 10 topics with short time 3.02 minute while traditional LDA accuracy is not well and take time more than modified LDA. The quality of topics generation in Modified LDA model is better than traditional LDA model.

8 CONCLUSION:-

The paper provides a comparison between deep learning (CNN model) and topics model (LDA and Modified LDA) for text classification using large data set (20 newsgroup) . The experimental results confirm the effectiveness and superiority of the CNN model than other two models .when we modifying LDA with TF-IDF method it achieve better accuracy than traditional LDA that improved from 60% to 74.4% and decreased the time from 4.04 to 3.02. The results prove the two models of LDA implemented with short time better than CNN but the CNN module achieve higher accuracy 94.7%. According to the results, CNN are very effective and useful for text classifications tasks.

9 REFERENCE:

[1] Lilleberg, Joseph, Yun Zhu, and Yanqing Zhang. "Support vector machines and word2vec for text classification with semantic features." 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). IEEE, 2015.

[2] 2-Conneau A., Schwenk H. , Le Cun Y. and Barrault L. ," Very Deep Convolutional Networks for Text Classification", Association for Computational Linguistics,2017, Volume 1, pages 1107–1116, Valencia, Spain.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research3.Jan (2003): 993-1022.

[4] Jelodar, Hamed, et al. "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey." Multimedia Tools and Applications 78.11 (2019): 15169-15211.

[5] Zhao, Dexin, Jinqun He, and Jin Liu. "An improved LDA algorithm for text classification." 2014 International Conference on Information Science, Electronics and Electrical Engineering. Vol. 1. IEEE, 2014.

[6] Das, Rajarshi, Manzil Zaheer, and Chris Dyer. "Gaussian lda for topic models with word embeddings." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015.
7- Zhang H.and Zhong G.," Improving short text classification by learning vector representations of both word and hidden topics". Volume 102, 15 June 2016, Pages 76-86.

[7] Feng, Lei, et al. "Topic Modeling of Environmental Data on Social Networks Based on ED-LDA." International Journal of Environmental Monitoring and Analysis 6.3 (2018): 77.

[8] Johnson, Rie, and Tong Zhang. "Effective use of word order for text categorization with convolutional neural networks."arXiv preprint arXiv:1412.1058 (2014).

[9] Gu, Jiuxiang, et al. "Recent advances in convolutional neural networks." Pattern Recognition 77 (2018): 354-377.

[10] Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

[11] Zhang, Xuemiao, et al. "Text Classification Model Based on Document Matrix Convolutional Neural Networks." 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017). Atlantis Press, 2017.

[12] Tian, Juan, Dingju Zhu, and Hui Long. "Chinese Short Text Multi-Classification Based on Word and Part-of-Speech Tagging Embedding." Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. ACM, 2018.

[13] Dieng, Adji B., Francisco JR Ruiz, and David M. Blei. "The Dynamic Embedded Topic Model." arXiv preprint arXiv:1907.05545 (2019).

[14] Boyd-Graber, Jordan L., and David M. Blei. "Syntactic topic models." Advances in neural information processing systems. 2009.

[15] Salton, Gerard, and Clement T. Yu. "On the construction of effective vocabularies for information retrieval." ACM SIGIR Forum. Vol. 9. No. 3. ACM, 1973.

[16] Ye, Jingyi, Xiaojun Jing, and Jia Li. "Sentiment analysis using modified LDA." International conference on signal and information processing, networking and computers. Springer, Singapore, 2017.

[17] Georgakopoulos, Spiros V., et al. "Convolutional neural networks for toxic comment classification." Proceedings of the 10th Hellenic Conference on Artificial Intelligence. ACM, 2018.

[18] Gal, Yarin, and Zoubin Ghahramani. "A theoretically grounded application of dropout in recurrent neural networks." Advances in neural information processing systems. 2016.

[19] Chen, Si, et al. "Deep Learning Method with Attention for Extreme Multi-label Text Classification." Pacific Rim International Conference on Artificial Intelligence. Springer, Cham, 2019.