Text Mining using Deep Learning Article Review

Nehad Mohamed Ibrahim
Lecturer
College of Computer Science & Information Technology
Imam Abdulrahman Bin Faisal University
Email: nmaibrahim@iau.edu.sa

Abstract

Deep Learning has efficient and accurate methods of learning which come back to the research area again after rapidly developments in the hardware, Also the text learning either supervised or unsupervised open area for the research. This paper aims to provide the researcher in (deep learning for text learning supervised or unsupervised) domain by comprehensive knowledge in this domain, it represents an overview of important articles over the last five years and discus methods that used and the conclusion. This article conducted to address relevant researches about the deep learning use in text mining by using the Google Scholar to define the period (issued between 2013 and 2018).

Keywords: Deep Learning, Natural Language Processing, Machine Learning, Neural Network and CNN

I. Introduction

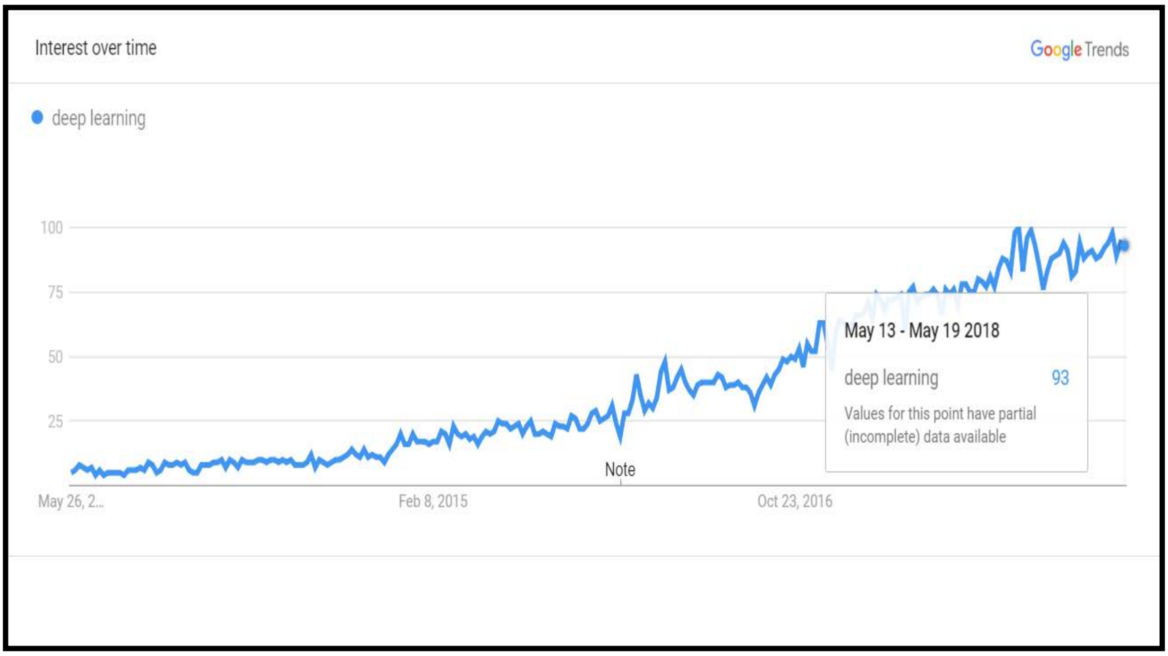
This review article conducted to address relevant researches about the text mining advanced classification techniques, in the last years the deep learning come back to the research area. Also the text mining is the open research area this article to review the last techniques in the deep learning used in text mining. To perform a wide-ranging survey, I was used the Google Scholar to define the period (issued between 2013 and 2018). The search methodology uses the following keywords:

Natural Language Processing AND Deep Learning AND Classification.

1.1. Research Articles Selection

Articles with these characteristics were included (1) deep learning and (2) the classification, clustering or text mining or experimentally-induced conditions. Review articles, news, editorials, letter, or case studies were excluded from the review.

1.2. Data Extraction and Management



This work partitioned into three stages, as the following: first stage; any article not match from the title will separate. The second stage the rest articles were separated and articles that does not match the core of survey. In the third stage the rest articles were read carefully and remove any articles that did not match the core of survey.
The following information were recorded from the survey: Reference, learning type, Methods used in the learning, and (4) Results of the review.
Deep learning methodology is growing in pattern recognition and computer vision. Recent Natural Language Processing research is now increasingly concentrating by using new methodology deep neural learning.

The deep learning is kind of buzz word right now as seen in Fig. 1, which explain the google trend over last five years by using search item deep learning.

We introduce this survey on the natural language processing using deep machine learning, deep learning is the method that use a neural network with multi layers of nodes between input and output. The multi layers of nodes between input and output do processing in a series of stages and feature identification, just as our brains seem to.

The performance of text mining processes need to be increased, to enhance the performance of text mining, it needs to research new technologies and the new text mining methods. Deep learning is a new learning method in the text mining; it can improve the performance of the text mining processes to access the desired text information quickly. The deep learning has an important significance in the text mining.

Various deep learning types: deep neural networks (DNN), convolutional deep neural networks (CDNN) have been applied to many fields; will show in state-of-the-art results. CNNs have been heavily explored in the image recognition and computer vision fields, offering improvements over DNNs on many tasks. [14], [15]

This literature review will discuss set of features that has direct impact in deep machine learning and in text mining.
The following (Table 1) conclude the title of study, type of learning, methods that used in each article and finally the results.

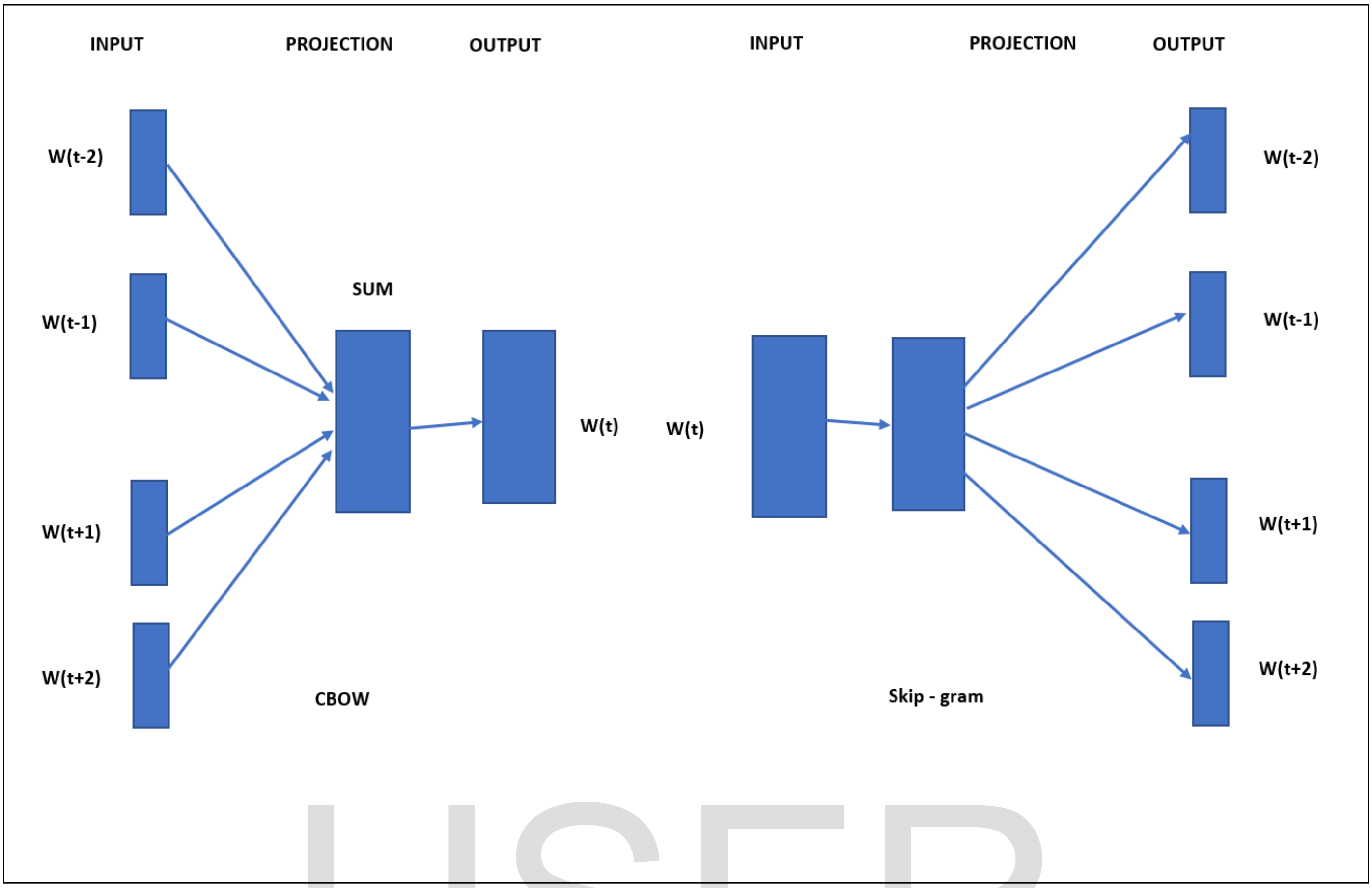


Table 1: List of articles conclusion

Ref.	Title of study	Type of Study	Methods or Models used	Results
[1]	Study of vector of words representations by various models.	Supervise learning	This work introduced the Skip-gram and Continuous Bag-of-Words (CBOW) model, these models an efficient approach to high quality learning by using vector representations of words and apply on the large amounts text data. (See Figure 2) Models: RNNLM, NNLM, CBOW and Skip-gram	The CBOW Faster to train than the skip-gram model But skip-gram performs better than CBOW

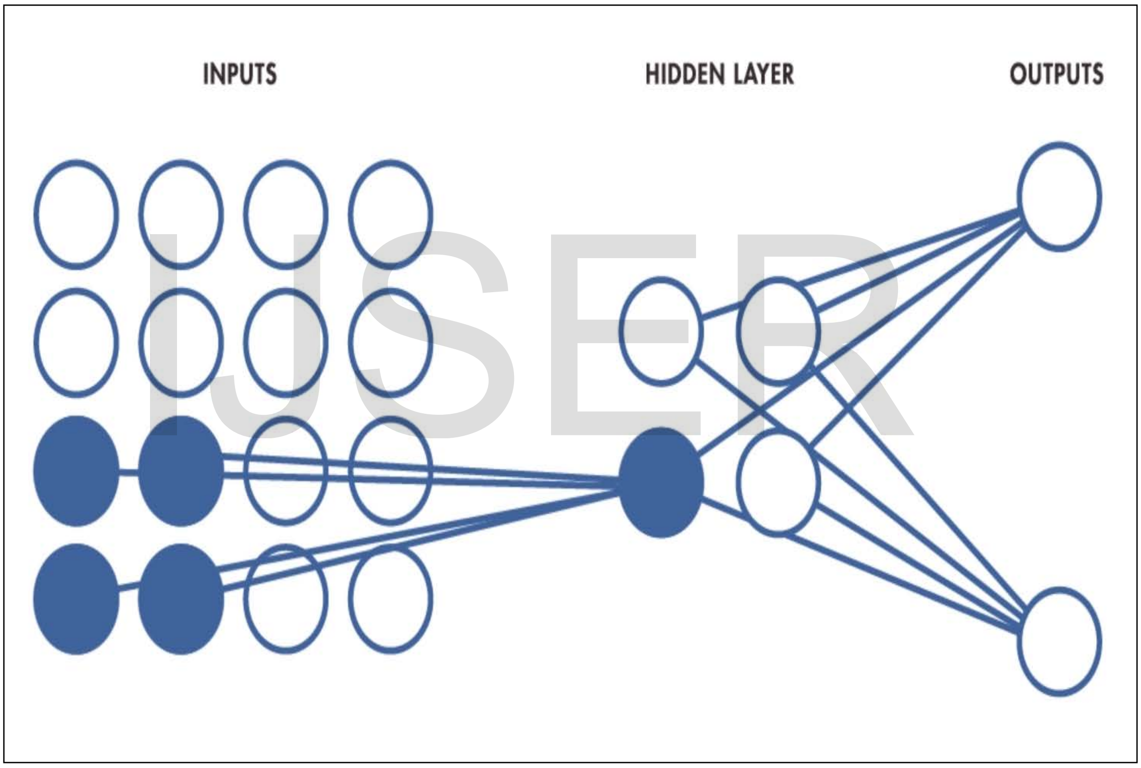
[2]	Deep learning Analysis in text mining.	Un-supervise Learning	Neural network model of deep learning.	Applications of deep learning in text mining increase the speed, quality and accuracy of the text mining.
[3]	Deep Learning methods for Subject Text Classification of Articles	Supervise Learning	This work presents a method of classification of text documents using deep neural network by two approaches: the first LSTM (long short-term memory) units. The second CNN based on more sophisticated word2vec method. Models: CNN and LSTM	The most promising results of classification were obtained with word2vec vector space.
[4]	Sequence Learning with Neural Networks	Supervise Learning	Use the Recurrent Neural Network (RNN) in machine translation by set input sequence by English and the output sequence is French. Models: RNN	By using a multilayered Long Short-Term Memory (LSTM) to map the input order to a vector of a fixed dimensionality, and then another deep LSTM to interpret the target order from the vector.
[5]	Distributed Representations of Phrases and Words and their Compositionality	Supervise learning	Using expression vectors instead of the word vectors. To train distributed representations of words and expressions with the Skip-gram model and demonstrate that these representations linear structure that makes detailed analogical reasoning possible. Models: Skip-gram	When trained without subsampling the Hierarchical Soft-Max will achieve the lower performance, it became the best performing method when down sampled the many words.

[6]	Designing a better data representation for deep neural networks and text classification	Supervise Learning	This work presents a new method to improve the training convolutional neural networks CNN from text by using character encoding. By using tweet sentiment data, the networks trained by character encoding, and measured the accuracy and training time. Models: CNN	By using the new approach of character encoding, implied as log (m), this approach allow training the neural network faster by 4.85 times. Also, this method achieved Meaningfully improve the network design performance by using log (m) encoding compared to 1-of-m encoding.
[7]	Distributed Representations of Sentences and Documents	Un-Supervise learning	This work describes Vector of paragraph, Algorithm that use vector representations in learning the sentences and documents. to predict the nearby words in contexts the paragraph. Models: Paragraph Vector	The advantages of Paragraph Vector in get the semantics of paragraphs is high performance. Also, this method overcome many weaknesses of BOW models.

[8]	Short Text Clustering via Convolutional Neural Networks	Un- Supervise learning	This study presents the combining Convolutional Neural Networks and semantic constraint, by using word embedding in unsupervised learning task on the short text. Models: CNN	When use CNN with word embedding showing this collection gives us the better performance than some other existing approaches, such as Laplacian eigenvectors , average embedding , and term frequency-inverse document frequency for clustering.

[9]	CNN with word embedding clustering to improving short text classification	Un-supervising learning	This study provides the words embedding trained used to initialize the table, that enable to measure words affinity and introduce extra knowledge and calculate the Euclidean Distance between two vectors and clustering using CNN. Models: CNN	This method developed to compute multiple units of scale applied on short texts. By using the embedding text which will collect similar words together this method enhances the performance in learning algorithms.
[10]	Recurrent Convolutional Neural Networks for Text Classification	Supervise learning	Use RCNN and Compare with the widely used text classification methods: Bag of Words/Bigrams + LR/SVM Average Embedding + LR LDA Tree Kernels Recursive NN CNN Models: RCNN	Neural networks can capture more contextual information of features compared with traditional methods based on BoW model. the convolution-based approaches achieve better results than Recursive NNs.
[11]	Use learning vector to Improve short text classification	Supervising Learning	This study present classification based on words vectors and topics vectors in the short text. also, evaluation the topic model with LDA on the quantity and improve texts with the word topics. On the enhanced corpus, viewing topics as new words, the learning of both words vectors and topics vectors are performed together. Models: TWE (Topical Word embeddings)	By using the learning words vectors and topics vectors enhance and improve text classification. Also, BOW has very small intelligence in semantics of words.
[12]	Dependency-Based Word Embeddings	Un-supervise Learning	In this work the author made generalization of the skip-gram model with negative sampling by replacing the bag-of-words contexts with arbitrary contexts. By perform experiments with dependency-based (syntactic contexts). Models: Word2Vec	This method produces markedly different kinds of similarities.

[13]	CNN for Sentiment Analysis	Unsupervised Learning of Word-Level Embedding	Propose a new deep convolutional neural network in two different domains based on character embedding to perform sentiment analysis of short texts. Models: Char SCNN, SCNN, RNTN, MV-RNN, RNN, NB and SVM	Character-level information has a greater impact for Twitter data. Using unsupervised pre-training, Character to Sentence Convolutional Neural Network provides an absolute accuracy improvement of 1.2 over SCNN.



II. Results & Discussion

1.3. Search Results

This article review identified 30 relevant articles. The articles address the natural language processing using deep machine learning. A total of 11 researches retrieved for further assessment. Nineteen of these articles were omitted because they did not concentrate on text mining and deep learning. The remaining 11 articles fulfill all inclusion and exclusion criteria and were

included for this review.

The main points in this research where divide into two parts:

1.4. Data representation

In this point I will conclude the main methods to data representations, that used as inputs to the classification or clustering process.

A. Word Embedding

Distributed vectors or word embedding which hold the attributes of the neighbor words. The measuring of distance between vectors is available by using similarity measurements as cosine method, also it's used in the data preparation phase and used in the unlabeled big data.

B. Word2Vector

Word2vec is a method of representation of words in multidimensional vector space, used in [3].

III. SUPERVISE LEARNING MODELS

1. CONVOLUTIONAL NEURAL NETWORK

CNN as in Figure 3, CNN is a specific type of AI neural network has an input layer, an output layer and hidden layers. CNNs has two deferent models, sentence and window models. CNNs can apply to image processing, natural language processing and other kinds of detection tasks.

1.1. CNN Approaches:

This CNN model has two types from input perspective:

1.1.1. Sentence Approach:

In general CNN architecture consists of set of filters that called kernels which usually multiple different levels over the embedding word matrix. Each layer is often followed by a max-pooling process, which is a sample based that subsamples. Fig. 4 depicts such a sentence as an input to the CNN framework. Max pooling strategy has two primary reasons:

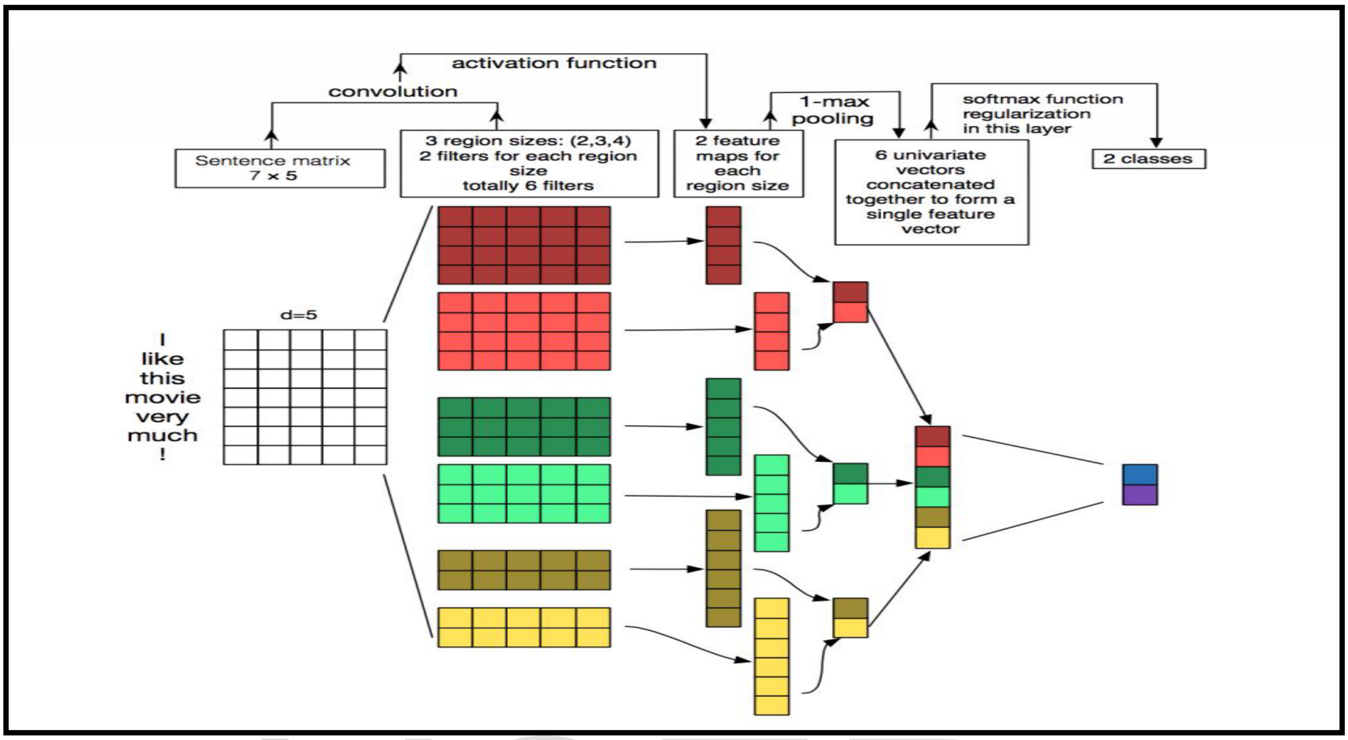
1.1.2. Max pooling process supply a fixed-width result that is required for supervising learning.

It caused the output's dimensions reduction, however keeping the most relevant sentence attributes. Sometimes this doing in a translation fixed method wherever each filter can extract a particular attribute from the sentence and adds it to the last sentence representation. This mixture of convolution layer followed by max-pooling operation is usually called CNN (Convolution Neural Networks).

1.1.3. Window Approach:

ISSN 2229-5518

A window method is use the label of a words depends on its neighboring words. A CNN is applied to this sub-sentence as expounded earlier and predictions are featured to the word in the center of the window.



1.2. CNN Applications

We present some of applications that employed CNNs on NLP tasks. Kim [20] explored a variety of sentence classification tasks, inclusive sentiment, subjectivity and showing competitive results. The researchers were adapted to this work quickly it's simple but its effective network. After train the network for a certain task, the convolutional filters became oriented to attribute detectors that were important for that target task (Fig. 5).

The main defect in this network is the disability to put model depend on the long distances.

Overall, the convolution neural networks are effective in the semantic by using the contextual windows. They include trained large number of parameters that require heavy data trained. Another problem within convolution neural networks is their inability to preserving sequential order in

			POSITIVE			
lovely	comedic	moments	and	several	fine	performances
good	script	,	good	dialogue	,	funny
sustains	throughout	is	daring	,	inventive	and
well	written	,	nicely	acted	and	beautifully
remarkably	solid	and	subtly	satirical	tour	de
			NEGATIVE			
,	nonexistent	plot	and	pretentious	visual	style
it	fails	the	most	basic	test	as
so	stupid	,	so	ill	conceived	,
,	too	dull	and	pretentious	to	be
hood	rats	butt	their	ugly	heads	in

Fig. 5: Top 7-grams by four learned 7-gram filters [19]

their representations [19, 22]. These problems solved in other recursive models.

2. RECURRENT NEURAL NETWORKS

RNNs Recurrent Neural Networks are popular models that used in many NLP operations; it uses the consecutive information method. It's called recurrent in order to it apply the same task in each iteration; the output is dependent on the previous stage. RNNs [23] use the idea of processing consecutive information.

Generally, a fixed-length vector is produced to sequence representation by supply words one by one to a recurrent unit. Overall, RNNs have "storage" over previous computations and use this information in current operation. The templates such as language modeling [2, 24, 25], machine translation [26, 27, 28], speech recognition [29, 30, 31, 32], image captioning [33] are suited for many NLP tasks. All these templates listed RNNs as important for NLP applications in the latest years.

Given RNN that performs processing by demonstrating units in consecutive, it has the capability to save the inherent consecutive presented nature in language; however, units are words, sentences or even characters. Words in a language present their semantic meaning depend on the previous words in the sentence. RNNs are customized for modeling such sequence dependencies in language.

Another factor support RNN's appropriate for series modeling tasks lies in its ability to model variable length of text, including very long sentences, paragraphs and even documents [34]. RNNs network have flexible computational steps over CNNs network that provide better modeling ability and create the prospect to capture unlimited context. This capability to manipulate input of random size became one of the marketing of major works using RNNs [35].

2.1. RNN MODELS

2.1.1. Simple RNN

In the stat of NLP, RNNs are primarily based on Elman network [23] and they are three layer networks. Fig. 6 explain RNN network which is expose across time to adjust a whole sequence. In the figure 6, Xt is taken as the input to the network at time step t and St Represent the hidden state at the same time step.

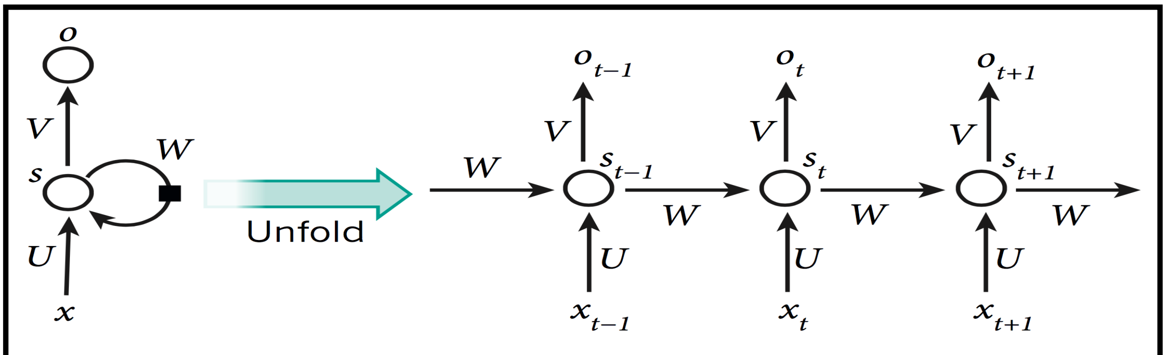
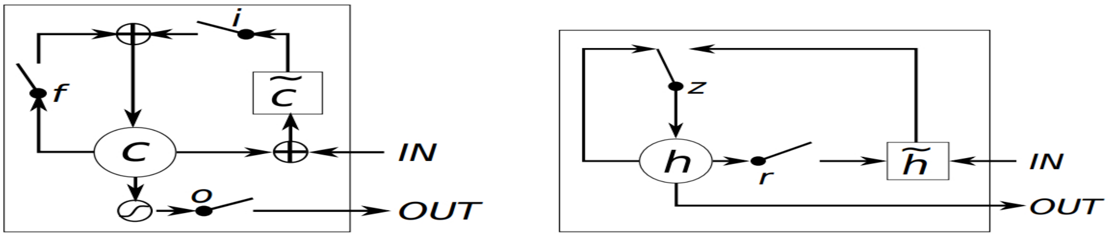


Fig. 6: RNN network [37].

Calculation of St is based as per the equation: St = f(U Xt +WSt-1)



(a) Long Short-Term Memory (b) Gated Recurrent Unit

Fig. 7: LSTM and GRU gate [35].

So, St is computed depended on the current entry and the previous step's state. Also; the function f is considered as a non-linear transformation such as tanh; ReLU and U; V; W account for weights that are shared across time. In the context of NLP, Xt typically involve of one-hot encodings. At times, they can also be abstract representations of textual content. Ot illustrates the output of the network which is also often subjected to non-linear.

The hidden context of the RNN is typically taken to be its most eventual element. As stated before, it can be considered as the network's memory element that aggregate information from other steps. Practically, the simple RNN networks suffer from simplicity learn and tune the parameters of the previously layers in the network.

2.1.2. Long Short-Term Memory:

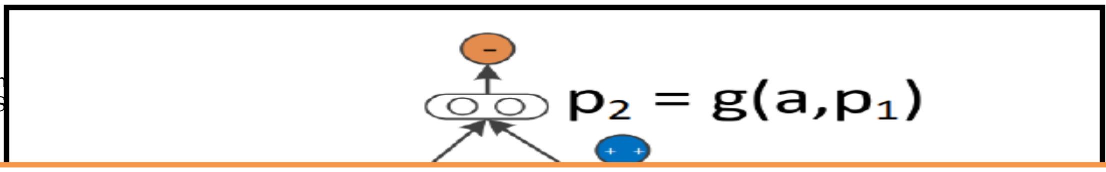
LSTM [38, 39] (Fig. 7) has new "forget" gates to over RNN network. This method allows it to control both the vanishing and exploding gradient problem.

2.1.3. Gated Recurrent Units:

Another approach gated from RNN network variant called GRU [36] (Fig. 7) that enhance LSTM complexity.

GRU cover the two gates the first gate to reset and second gate to update, and holders the information flow as LSTM sans a memory unit. Also, it discoveries the hidden content without any control. GRU can be a more effective RNN than LSTM.

2.2. RNN Applications



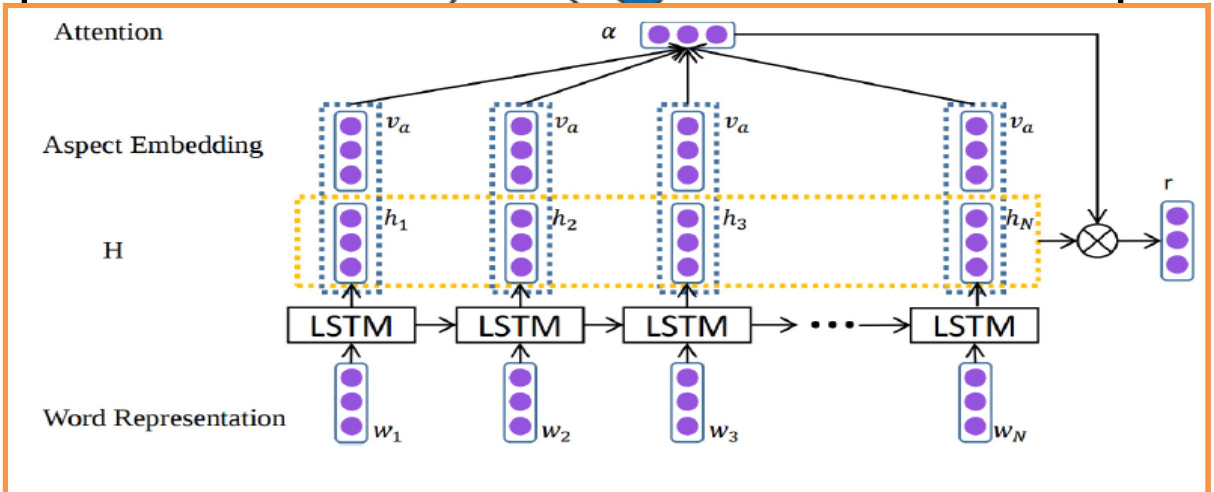


Fig. 8: Classification using attention [44].

RNN network for word-level classification: mostly RNNs research using in word-level classification. [40] Proposed to use bidirectional LSTM. The network do save the long term information around the target word resulting in two fixed-size vector, on top of which another fully-connected layer was built. They used a CRF layer at last for the final entity tagging.

[41] Proposed deep RNNs network where multiple layers of hidden states. This work proposed the usage of RNNs on tasks related to the context of NLP. [42] compared the result gained by RNN. Another important issue is statistical machine translation [43].

RNN for sentence-level classification: Wang et al. [18] proposed training entire tweets with LSTM, whose hidden state is used for predicting sentiment polarity. Another strategy is more complex DCNN structure by [19] designed to provide CNN models by saving long-term dependencies.

[44] Proposed sentiment analysis solution that used embedding to add support during classification (Fig. 8).

3. RECURSIVE NEURAL NETWORKS

RNN represent a method that model orders. also, language display a natural recursive structure, where sub-phrases and words combine into phrases in a hierarchical manner. Such structure can be displayed as a constituency parsing tree.

3.1. Recursive Neural Network Models

RNN structure in Fig. 9, g represent a synthetic function in the RNN network that represent a words or phrases (b; c or a; p1) to compute the higher level phrase (p1 or p2). Same form is used to represent all nodes. g is defined as:

$$p1 = \tanh\left(W \begin{bmatrix} b \\ c \end{bmatrix}\right), p2 = \tanh\left(W \begin{bmatrix} a \\ p2 \end{bmatrix}\right)$$

Another enhancement is the MV-RNN [45]. He was proposed representation every phrase and word as both a vector and a matrix.

Another variation is the recursive neural tensor network (RNTN) that introduce relations between the input vectors without parameters very large like MVRNN. [45] Propose classified semantic relationships between nominals in a sentence. [46] Proposed to classify the logical relationship between sentences with recursive neural networks. [47] Proposed LSTM units to avoiding the gradient vanishing problem.

The authors were used LSTM models to improve the sentence representation and improve sentiment analysis.

IV. UNSUPERVISE LEARNING APPROACHES

1. Reinforcement Learning for Sequence Generation

The reinforcement learning is a training method of an agent to execute separate actions before obtaining a reward. Usually in NLP, language generation tasks doing as reinforcement learning problems.

Given the current hidden state-run and the previous tokens, RNN language generators are always trained by maximizing the likelihood of each token in the ground-truth sequence. [48, 49] Present this discrepancy between training and inference, termed "exposure bias".

Another new approach for supervision learning by sequence-level is to use the adversarial training technique [50], where the training objective for the language generator is to another discriminator trained to separate generated sequences from real sequences.

The discriminator D and the generator G are trained together in a min-max game which ideally leads to G, generating sequences indistinguishable from real ones. This approach can be seen as a variation of generative adversarial networks in [50], where G and D are conditioned on certain stimuli (for example, the source image in the task of image captioning).

In practice, the above scheme can be realized under the reinforcement learning paradigm with policy gradient. For dialogue systems, the discriminator is analogous to a human Turing tester, who discriminates between human and machine produced dialogues [51].

2. Unsupervised Sentence Representation Learning

Sentence distributed representation can be learned in an unsupervised learning. Auxiliary activity has to be defined for the learning process [16], Also, [52] propose learning sentence representation by using skip-thought model , that predict two adjacent sentences based on the given sentence, similar to the skip-gram model [8] for learning word embedding.

3. Deep generative models

We review recent research, first research to discover structure in natural language with variation auto encoders (VAEs) [53], second research to generative adversarial networks (GANs) [50].

GAN is another type of generative model collected of two networks. First network called generative neural network which decodes latent representation to a data instance, second network called discriminative network to discriminate between instances from the data distribution and synthesized instances produced by the generator.

V. CONCLUSION

Deep learning offers a method to conduct complex computation and data [37]. The most popular practices in deep learning research for NLP based on the supervised learning. However, any unlabeled data which need unsupervised or semi-supervised learning. [54] Expect more research in the real world language.

Finally, Depending on Deep learning used to enhance a decision based on past experience.

VI. REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space," in Proceedings of Workshop at ICLR,

Sep 2013.

[2] H. Wang, M. Jiang, J. Qi, X. Zhang, Q. Wang, Y. Zhou, M. Bai, L. Liu, Z. Pei. "Application of Deep Learning in Text Mining,". International Conference on Mechatronics, Control and Electronic Engineering. 2014.

[3] P. Semberecki, H. Maciejewski. "Deep Learning methods for Subject Text Classification of Articles," Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 357–360. 2017.

[4] I. Sutskever, O. Vinyals, & Le, Q. V. Le. "Sequence to sequence learning with neural networks,". In Advances in neural

information processing systems. pp. 3104-3112. 2014.

[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in Proceedings of Workshop at ICLR, Oct 2013.

[6] J. D. Prusa and T. M. Khoshgoftaar, " Designing a better data representation for deep neural networks and text classification, " in IEEE 17th International Conference on Information Reuse and Integration, 2016

[7] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014.

[8] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao,F. Wang, and H. Hao, “Short Text Clustering via Convolutional Neural Networks,” Proceedings of NAACL-HLT 2015, pages 62–69

[9] P. Wang, B. Xu, J. Xu, G. Tian, C. Liu, and H. Hao, " Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," 0925-2312 2015 Elsevier B.V.

[10] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," 2015, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[11] H. Zhang and G. Zhong, "Improving short text classification by learning vector representations of both words and hidden topics," Elsevier B.V. Knowledge-Based Systems 102 (2016) 76–86, March 2016

[12] O. Levy, and Y. Goldberg, “Dependency-Based Word Embeddings,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 302–308, Baltimore, Maryland, USA, June 23-25 2014. c 2014 Association for Computational Linguistics

[13] C. Nogueira, and M. Gatti, “Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts,” Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 69–78, Dublin, Ireland, August 23-29 2014.

[14] Y. LeCun, F. Huang, and L. Bottou, “Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting,” in Proc. CVPR, 2004.

[15] S. Lawrence, “Face Recognition: A Convolutional Neural-Network Approach,” IEEE Transactions on Neural Networks, vol. 8, no. 1, pp. 98–113, 1997.

[16] T Young, Devamanyu Hazarika , S Poria, E Cambria. Recent trends in deep learning based natural language processing. 2017

[17] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts et al., “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, 2013, p. 1642.

[18] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory." in ACL (1), 2015, pp. 1343–1353.

[19] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, June 2014.

[Online]. Available: http://goo.gl/EsQCuC

[20] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.

[21] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," arXiv preprint arXiv:1510.03820, 2015.

[22] Z. Tu, B. Hu, Z. Lu, and H. Li, "Context-dependent translation selection using convolutional neural network," arXiv preprint arXiv:1503.02357, 2015.

[23] J. L. Elman, "Finding structure in time," Cognitive science, vol. 14, no. 2, pp. 179–211, 1990.

[24] T. Mikolov, S. Kombrink, L. Burget, J. Cˇ ernocky`, and S. Khudanpur, "Extensions of recurrent neural network language model," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011, pp. 5528–5531.

[25] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 1017–1024.

[26] S. Liu, N. Yang, M. Li, and M. Zhou, "A recursive recurrent neural network for statistical machine translation," Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1491–1500, 2014.

[27] M. Auli, M. Galley, C. Quirk, and G. Zweig, "Joint language and translation modeling with recurrent neural networks." in EMNLP, 2013, pp. 1044–1054.

[28] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems, 2014, pp. 3104–3112.

[29] T. Robinson, M. Hochberg, and S. Renals, "The use of recurrent neural networks in continuous speech recognition," in Automatic speech and speaker recognition. Springer, 1996, pp. 233–258.

[30] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013, pp. 6645–6649.

[31] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 1764–1772.

[32] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," arXiv preprint arXiv:1402.1128, 2014.

[33] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[34] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification." in EMNLP, 2015, pp. 1422–1432.

[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[36] K. Cho, B. Van Merri¨enboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.

[37] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 9th International Conference on Artificial Neural Networks, pp. 850–855, 1999.

[40] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[41] A. Graves, "Generating sequences with recurrent neural

networks," arXiv preprint arXiv:1308.0850, 2013.

[42] M. Sundermeyer, H. Ney, and R. Schl¨uter, "From feedforward to recurrent lstm neural networks for language modeling," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 23, no. 3, pp. 517–529, 2015.

[43] M. Sundermeyer, T. Alkhouli, J. Wuebker, and H. Ney, "Translation modeling with bidirectional recurrent neural networks." in EMNLP, 2014, pp. 14–25.

[44] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attentionbased lstm for aspect-level sentiment classification." in EMNLP, 2016, pp. 606–615.

[45] 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.

[46] S. R. Bowman, C. Potts, and C. D. Manning, “Recursive neural networks can learn logical semantics,” arXiv preprint arXiv:1406.1827, 2014.

[47] K. S. Tai, R. Socher, and C. D. Manning, “Improved semantic representations from tree-structured long short-term memory networks,” arXiv preprint arXiv:1503.00075, 2015.

[48] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in Advances in Neural Information Processing Systems, 2015, pp. 1171–1179.

[49] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, “Sequence level training with recurrent neural networks,” arXiv preprint arXiv:1511.06732, 2015.

[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.

[51] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky, “Adversarial learning for neural dialogue generation,” arXiv preprint arXiv:1701.06547, 2017.

[52] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, “Skip-thought vectors,” in Advances in neural information processing systems, 2015, pp. 3294–3302.

[53] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2013.

[54] T. Baltruˇsaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” arXiv preprint arXiv:1705.09406, 2017.