Volume 1, Issue 2 (2019)
Article No. 5
PP 1-6

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Survey on Text Prediction Techniques

Deepal S. Thakur[1], Rajiv N. Tarsarya[1], Akshay A. Vaskar[1], Ashwini Save[1]
[1](*Computer Engineering Department, VIVA Institute of Technology, India)*

Abstract: *Writing long sentences is bit boring, but with text prediction in the keyboard technology has made this simple. Learning technology behind the keyboard is developing fast and has become more accurate. Learning technologies such as machine learning, deep learning here play an important role in predicting the text. Current trending techniques in deep learning has opened door for data analysis. Emerging technologies such has Region CNN, Recurrent CNN have been under consideration for the analysis. Many techniques have been used for text sequence prediction such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Recurrent Convolution Neural Networks (RCNN). This paper aims to provide a comparative study of different techniques used for text prediction.*

Keywords *– CNN, Deep Learning, LSTM, Machine Learning, RCNN, RNN*

1. INTRODUCTION

Word prediction tools were developed which can help to communicate and also to help the people with less speed typing. The research on word prediction has been performing well [5]. Word prediction technique does the task of guessing the preceding word that is likely to continue with few initial text fragments. Existing systems work on word prediction model, which suggests the next immediate word based on the current available word [4]. These systems work using machine learning algorithms which has limitation to create accurate sentence structure [6]. Developing technologies has been producing more accurate outcomes than the existing system technologies, models developed using deep learning concepts are capable of handling more data efficiently and predicts better than ML algorithms [2].

2. LITERATURE REVIEW

S. Lai, et. al. [1] have proposed the context-based information classification; RCNN is very useful. The performance is best in several datasets particularly on document level datasets. Depending on the words used in the sentences, weights are assigned to it and are pooled into minimum, average and the max pools. Here, max pooling is applied to extract the keywords from the sentences which are most important. RNN, CNN and RCNN when compared with other traditional methods such as LDA, Tree Kernel and logistic regression generates high accurate results.

A. Hassa, et. al. [9] have proposed RNN for the structure sentence representation. This tree like structure captures the semantic of the sentences. The text is analyzed word by word by using RNN then the semantic of all the previous texts are preserved in a fixed size hidden layer. For the proposed system LSTM plays important role, being a memory storage it holds the characters which helps in predicting the next word.

J. Y. Lee, et. al. [7] have proposed that text classification is an important task in natural language processing. Many approaches have been developed for classification such as SVM (Support Vector Machine), Naïve Bayes and so on. Usually short text appears in sequence (sentences in the document) hence using information from

Volume 1, Issue 2 (2019)
Article No. 5
PP 1-6

preceding text may improve the classification. This paper introduced RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network) based model for text classification.
V. Tran, et. al. [5] have proposed that n-gram is a contiguous sequence of 'n' items from a given sequence of text. If the given sentence is 'S' ,we can construct a list on n-grams from 'S' , by finding pairs of words that occurs next to each other. The model is used to derive probability of sentences using the chain rule of unconditional probability.

Z. Shi, et. al. [4] have defined that recurrent neural network has input, output and hidden layer. The current hidden layer is calculated by current input layer and previous hidden layer. LSTM is a special Recurrent Neural Network. The repeating module of ordinary RNN has a simple structure instead LSTM uses more complex function to replace it for more accurate result. The key element in the LSTM is the cell state which is also called as hidden layer state.
J. Shin, et. al. [10] have defined that understanding the contextual aspects of a sentence is very important while its classification .This paper mainly focuses on it. Various approaches like SVM, T-LSTM, and CNN have been previously used for sentence classification. But, the proposed C-CNN (Contextual-CNN) gives better results i.e. The C-CNN achieves state-of-theart accuracy 52.3% on the fine-grained sentiment prediction task and 95.2% on the TREC question categorization task.

W. Yin, et. al. [11] have defined various classification tasks are important for Natural language processing applications. Nowadays CNN are increasing used as they are able to model long range dependencies in sentence, the systems used are with fixed-sized filters. But, the proposed MVCNN approach breaks this barrier and yields better results when applied to multiple datasets: binary with 89.4%, Sentiment140 with 88.2% and Subjectivity classification dataset (Subj.) with 93.9% accuracy. Multichannel initialization brings two advantages: 1) Frequent words can have c representations in the beginning (instead of only one), which means it has more available information to leverage 2) A rare word missed in some embedding versions can be "made up" by others (we call it "partially known word").
I. Sutskever, et. al. [12] have defined deep learning being the newest technology in the era has advanced in many fields. One of the techniques called as Deep Neural Networks are very powerful machine learning models and have achieved efficient and excellent performance on many problems like speech recognition, visual object detection etc. due to its ability to perform parallel computation for the modest no of steps. Many attempts have been made to address the problems with neural network. The results showed that a large deep LSTM with a limited vocabulary can outperform a standard SMT-based system.

W. yin, et. al. [3] have defined that deep neural networks have revolutionized the field of natural language processing. Convolutional Neural Network and Recurrent Neural Network, the two main types of DNN architectures, are widely explored to handle various NLP tasks. CNN is supposed to be good at extracting position invariant features and RNN at modelling units in sequence. RNNs are well suited to encode order information and long-range context dependency. CNNs are considered good at extracting local and position-invariant features and therefore should perform well on TextC, but in experiments they are outperformed by RNNs.

K. C. Arnold, et. al. [6] have proposed an approach that presents phrase suggestion instead of word predictions. It says phrases were interpreted as suggestions that affect the context of what the user write more than then the conventional single word suggestion. The proposed system uses statistical language modelling capable of accurately predicting phrases and sentences. System used n-gram sequence model and KenLM for language model queries which used Kneser-Ney smoothing. It pruned the n grams that repeated less than two times in the dataset, by marking the start-of-sentence token with some additional flags to indicate the start of the sentence. The work demonstrated the phrase completions were accepted by users and were interpreted as suggestions rather than the predictions.

M. Liang, et. al. [8] have defined that in past years, deep learning techniques has achieved great success in many computer vision tasks. The visual system of the brain shares many properties with CNNs and hence they have inspired neuroscience to a great extent. CNN is typical feed forward architecture while in the visual systems recurrent connections are abundant. So incorporating recurrent connection in each convolutional layer the following system was proposed for Object Recognition. The proposed model tested several benchmark object detection datasets. RCNN achieved better results over CNN.

P. Ongsulee [2] has defined that machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish baseline behavioural profiles for various entities and then used to find meaningful anomalies.

3. ANALYSIS TABLE

Sr. No.	Technical paper name	Technique Used	Dataset Used	Accuracy
1.	Recurrent Convolutional Neural Networks for Text Classification [1]	RNN, CNN, R-CNN.	20Newsgroups, Fudan set, ACL anthology network, Sentiment Standford TreeBank.	20Newsgroups: 96.49% Fudan s 95.20% ACL anthology network: 49.19% Sentiment Standford TreeBank: 47.21%
2.	Deep Learning for Sentence Classification [9]	RNN, LSTM.	Stanford Large Movie Review Dataset IMDB and Stanford Sentiment	--

			Treebank (SSTb) dataset, Word2vec.	
3.	Sequential ShortText Classification with Recurrent and Convolutional Neural Networks [7]	RNN, CNN, LSTM, ANN.	DSTC 4, MRDA, SwDA.	For DSCT 4: LSTM- 66% CNN- 65.5% For MRDA: LSTM- 84.3% CNN-84.6% For SwDA: LSTM- 69.6% CNN-73.1%
4.	A Vietnamese Language Model Based on Recurrent Neural Network [5]	RNN, LSTM.	1500 random movies collected from internet.	83.8%
5.	The prediction of character based on Recurrent Neural network language model [4]	RNN, LSTM.	--	--
6.	Contextual CNN:A Novel Architecture Capturing Unified Meaning for Sentence Classification [10]	Contextual CNN(C-CNN).	Sentiment Standford TreeBank: SST-5, SST-2, TREC.	SST-5:52.3% SST-2: 89.2% TREC: 95.2%
7.	Multichannel Variable-Size Convolution for Sentence Classification [11]	MV-CNN.	Sentiment Standford TreeBank(Binary, Fine-grained), Sentiment140[2], Subj[3].	Binary:89.4% Fine-grained: 49.6% Sentiment140[2]: 88.2% Subj[3]: 93.9%
8.	Sequence to sequence learning using Neural Network [12]	LSTM, RNN, DNN.	WMT'14 English to French.	LSTM: 37%
9.	Comparative study of CNN and RNN for Natural Language Processing [3]	CNN, RNN.	Sentiment Standford TreeBank, SemEval 2010 task 8, SNLI, WikiQA.	RNN (BiLSTM): 94.35% CNN: 94.18
10.	On Suggesting Phrases vs. Predicting Words for Mobile Text Composition [6]	n-gram Model, KenLM.	Yelp academic dataset.	--
11.	Recurrent Convolutional Neural Network for Object Recognition [8]	CNN, RNN, RCNN,	CIFAR-10, CIFAR-100, MNIST,	--

		RMLP.	SVHN.	
12.	Artificial Intelligence, Machine Learning and Deep Learning [2]	Supervised Learning, Unsupervised Learning, Semisupervised Learning, Reinforcement, ANN, CNN	--	--

Paper titled, "Recurrent Convolutional Neural Networks for Text Classification" used Recurrent Convolutional Neural Networks to classify the text as per context i.e. right context or the left context. It gave accuracy of 96.46% for a given dataset, other techniques gave accuracy less than the RCNN in terms of context. Another paper, "Deep Learning for Sentence Classification" compared the two techniques LSTM and RNN, both techniques standout with different result on same dataset, results show that a simple LSTM with one single layer perform well with unsupervised word vectors, which become an important feature in natural language processing and deep learning. Also paper "Sequential ShortText Classification with Recurrent and Convolutional Neural Networks" gave 84.3% and 84.6% accuracy on a certain data using LSTM and CNN techniques; this gives a slight upper edge to CNN which is a deep learning technique.

4. CONCLUSION

The above paper mainly discussed deep learning techniques like RCNN, RNN, and CNN which gives different results on different datasets giving varied accuracy. Large number of datasets were used for proper prediction. The research shows text sequence prediction can be implemented through deep learning techniques which can change the scenario of typing whole sentences.

The study of various papers gives a clear edge stating deep learning might provide better results when compared with other techniques. Previously, machine learning and natural language processing were used in prediction but deep learning models produced better accuracy.

Volume 1, Issue 2 (2019)
Article No. 5
PP 1-6

REFERENCES

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

[2] P. Ongsulee, "Artificial Intelligence, Machine Learning and Deep Learning", 15th International Conference on ICT and Knowledge Engineering (ICT&KE), 2017

[3] W. Yin, K. Kann, Mo Yu and H. Schütze, "Comparative study of CNN and RNN for Natural Language Processing", Feb-17.

[4] Z.Shi, M. Shi and C. Li, "The prediction of character based on Recurrent Neural network language model", IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017

[5] V. Tran, K. Nguyen and D. Bui, "A Vietnamese Language Model Based on Recurrent Neural Network", 2016 Eighth International Conference on Knowledge and Systems Engineering.

[6] K. C. Arnold, K.Z. Gajos and A. T. Kalai, "On Suggesting Phrases vs. Predicting Words for Mobile Text Composition";https://www.microsoft.com/enus/research/wpcontent/uploads/2016/12/arnold16suggesting.pdf, 2016

[7] J. Lee and F. Dernoncourt, "Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks", Conference paper at NAACL 2016.

[8] M. Liang and X. Hu, "Recurrent Convolutional Neural Network for Object Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015

[9] A. Hassan and A.Mahmood, "Deep Learning for Sentence Classification", IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2017

[10] J. Shin, Y. Kim and S. Yoon, "Contextual CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification",IEEE International Conference on Big Data and Smart Computing (BigComp), 2018

[11] W. Yin and H. Schutze, "Multichannel Variable-Size Convolution for Sentence Classification", 19th Conference on Computational Language Learning,Association for Computational Linguistics, 2015

[12] I.Sutskever, O.Vinyals and Q. V. Le, "Sequence to Sequence Learning with Neural Networks", Dec-14.

[13] Y. Zhang, B. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification", arXiv: 1510.03820v4 [cs.CL], 2016.

[14] A. Salem, A. Almarimi, G Andrejková, "Text Dissimilarities Predictions Using Convolutional Neural Networks and Clustering" World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018

[15] Y. Lin, J. Wang, "Research on text classification based on SVM-KNN" IEEE 5th International Conference on Software Engineering and Service Science, 2014

[16] A. Hassan, A. Mahmood, "Convolutional Recurrent Deep Learning Model for Sentence Classification", IEEE Access, 2018