Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 147 (2019) 131–135

Procedia Computer Science

www.elsevier.com/locate/procedia

2018 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2018

Deep Learning-Based Sentiment Analysis for Roman Urdu Text

Hussain Ghulam[a], Feng Zeng[a,*], Wenjia Li[b], Yutong Xiao[a]

[a]School of Software, Central South University,Changsha 4180083,China
[b]Department of Computer Science, New York Institute of Technology, New York, NY 10023,USA

Abstract

Sentiment Analysis has significant attention due to its versatile approach to analysis user's sentiments on various social networks, forums, e-marketing sites and blogs. Sentiments related data on the web has great importance and impact on customer's, readers and business firms.Reccurent Neural Network has been widely applied to perform Natural Language Processing tasks because it is designed for modeling the sequential data efficiently.
In this paper we used Deep Neural Long-short time memory model (LSTM).It has extraordinary capability to Capture long-range information and solve gradient attenuation problem, as well as represent future contextual information, semantics of word sequence magnificently. This paper is the foundation of adapting Deep learning methods to perform Roman Urdu Sentiment Analysis. Our experimental results shows the significant accuracy of our model and surpassed accuracy of baseline Machine learning methods.

Peer-review under responsibility of the scientific committee of the 2018 International Conference on Identification, Information and Knowledge in the Internet of Things.

Keywords: Recurrent Neural Nework(RNN) ; Long Short-term Memory(LSTM); Roman Urdu Sentiment Analysis ; Word embedding.

1. Introduction

Sentiment Analysis play vital role in Natural Language Processing (NLP) to acquire and process the meaningful iformation from user's reviews, which they expresses through online Communities and collaborative media. The opinionated data is becoming increasingly very important from applicative point of view, exclusively in Scientific and business intelligent systems to surmount and improve their services and products quality.[1] [2]
We have performed our model on Roman Urdu Corps. Roman Urdu [3] is popular language with Roman script which is easy to write and read by non-native Urdu speakers around the globe. There were some studies found to perform both Urdu and Roman Urdu Sentiment Analysis. Most of work has been done by implementing Lexicon based and Machine Learning baseline approaches.The limitation of this all approaches are the low applicability to new data and

* Corresponding author. Tel.: +86- 13397645839 ; fax: 0731-82656877
E-mail address: fengzeng@csu.edu.cn

Peer-review under responsibility of the scientific committee of the 2018 International Conference on Identification, Information and Knowledge in the Internet of Things.
10.1016/j.procs.2019.01.202

finite number or words in the lexicons and fixed no of assignation incapable to perform sentiment analysis on huge dataset.

In recent decade Deep learning has great achievement in various fields like Image recognition, self-driving cars and other intelligent system. Recurrent Neural network (RNN) [4], due to its capability of having memory to capture long term dependencies in sequential data has outperformed in Natural language processing Tasks (NLP) .The performance of Recurrent Neural network (RNN) is significantly better then classical approaches of Lexicon, Machine learning and statistical algorithms like Hidden Markov Model (HMM) [5]

RNN have LSTM network [6] which is consider as great effort to modeling sequential data like text and speech.[7].It has capability to capture global sequence features. Previous literature review shows evidential success of Long Short-Term Memory in sequential models. Wang et al 2015, used Long-Short-Term LSTM into POS tagging, chunking and NER tasks and internal representations are learnt from unlabeled text for all tasks [20]. Sundermeyer et al. analyzed LSTM neural network by modeling English and French [7]. Quan-Hoang Vo ; Huy-Tien Nguyen,(2017) Deep Neural Networks are employed to generate information channels for Vietnamese sentiment analysis.[8]

In Roman Urdu there are not such studies which used Deep neural Networks Models for Roman Urdu Sentiment Analysis due to lack of resources. Our Model is the first novel approach to apply Deep neural Network LSTM to perform Sentiment Analysis on Roman Urdu dataset. In other domains of NLP like machine translation, OCR Recognition, They have used deep neural network. Mehreen Alam et al.2017 has used LSTM sequence to sequence Network to translate Roman Urdu to Urdu Nastaliq [9] Adnan Ul-Hasan et al 2013 used Long Short Term Memory (LSTM) architecture with Connectionist Temporal Classification (CTC) output layer was employed to recognize printed Urdu text.[10]. The existing work covers Sentiment Analysis by using classical approaches and its sub topics like polarity Analysis [11], [12], [13], Lexicon based Sentiment analysis for Urdu Sentiment Sen-ti units.[14] , Roman Urdu opinion mining system (RUOMIS) [15], Urdu Sentiment Analysis by using Naïve Bayesian and decision tree [16],performing natural language processing (NLP) on Roman Urdu data set [17], Processing Informal, Romanized Pakistani Text Messages [18]

In our propose approach, in order to ameliorate the state of art performance of the Roman Urdu Sentiment Analysis, we deployed long short-term memory LSTM networks [6] to word segmentation task, the Contribution of approach can be follow:

1) Our work is the first to used Deep Neural Network long short-term memory (LSTM) to capture long-term sentence dependencies for Roman Urdu Sequential Modeling task.

2) It achieves highest accuracy on Roman Urdu Sentiment binary classification as compare to Baseline Machine Learning and Lexicon Based Approach's.

The paper is formed as follows. Section 2 elaborates architecture Our Proposed network. Next, in Section 3 Shows our experiments on Roman Urdu Dataset and summarizes our experimental results. Finally, we summarize key conclusions in Section 4.

2. Our Model

2.1. Recurrent Neural Network

Recurrent Neural Networks (RNNs) [4] have shown Promising performance in machine translation tasks. Typical RNN where X^t is an input,h^t is an output and A is the neural network which acquired information from the previous step in a loop. The output of one unit refers to the next one and the information send forward. RNN is not capable to capture long-term dependencies because of large updates to neural network model weights it cause error gradient accumulation and weights overflow if value larger 1 or vanishing happens when values are less than 1.

$$o^t = f(h^t; \sigma) \tag{1}$$

$$h^t = f(h^{t-1}, x^t; \sigma) \tag{2}$$

Where o^t, is the output of RNN at time t,x^t is the input of RNN at time t, and h^t is the state of the hidden layer(s)

2.2. Long Short-Term Memory Network (LSTM)

To surmount the overflow or vanishing error gradient and capturing long term dependencies , Long Short Term Memory (LSTM)[6] As an emerging variant of the RNN model widely used . LSTM can overcome this problem by using its gates to manage the error gradient. The Mathematical representation of LSTM can be showed as:

$$h_t = f(W_{xt} + Uh_{t-1} + b) \tag{3}$$

$$i_t = \sigma(W^i x_t + U^i h_{t-1} + b^i) \tag{4}$$

$$f_t = \sigma(W^f x_t + U^f h_{t-1} + b^f) \tag{5}$$

$$o_t = \sigma(W^o x_t + U^o h_{t-1} + b^o) \tag{6}$$

$$g_t = \sigma(W^g x_t + U^g h_{t-1} + b^g) \tag{7}$$

$$c_t = f_t \Theta c_{t-1} + i_t \Theta g_t \tag{8}$$

$$h_t = o_t \Theta tanh(c_t) \tag{9}$$

where σ Θ ,

respectively denote a logistic sigmoid function and element-wise multiplication; W_i,Ui, bi are respectively two weights matrices and a bias vector for input gate i. The denotation is similar to forget gate f, output gate o, tanh layer u, memory cell c and hidden state h. spontaneously, the forget gate decides which previous information should be forgotten, while the input gate controls what new information should be stored in the memory cell. Finally, the output gate decides the amount of information from the internal memory cell should be exposed. This gate units help a LSTM model remember significant information over multiple time steps.

2.3. Training Model

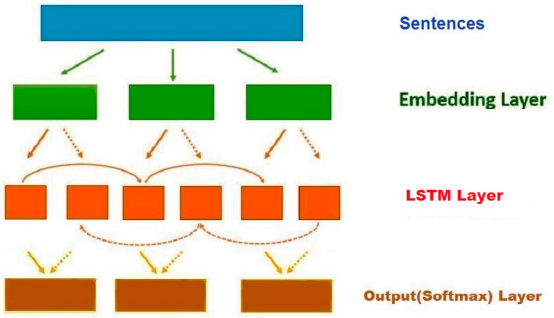

Fig. 1. Our proposed Deep Neural Model

In our evaluation of the proposed scheme, each classifier is implemented as a deep learning model having four layers, as illustrated in Figure 1, and is described as follows.

- The Input (Embedding) Layer [19] the input layer's size is defined by the number of inputs for that classfier.This number equals the size to the word vector plus the number of additional features. The word vector dimension was set to 300 so that to be able to encode every word in the vocabulary used.

- The hidden layer: The sigmoid activation was selected for the hidden LSTM layer. Based on preliminary experiments the dimensionality of the output space for this layer was set to 180. This layer is fully connected to both the Input and the subsequent layer.
- The LSTM layer:In this layer the embedding representation is implement to LSTM layer.At each time step t, a recurrent layer takes the input vector x and hidden state h by applying the recursive operation The output of the LSTM was run through an additional layer to improve the learning and obtain more stable output.it overcomes the dependencies of an RNN by augmenting the RNN with a memory cell see figure(1).
- Output layer. output layer with soft-max activation.

3. Experimetal Setup and Results

In our experimental work we performed a 10-fold cross-validation using the balanced binary dataset 8K. In each fold 90% of the dataset are used to build our train set and remaining 10% of the dataset are used to form our each Validation and test set. for our task the word embedding are initialized with 300-d (300 dimensional) Training of all the model parameters is performed with the mini-batch random gradient descent algorithm, which provides an adaptive learning rate. For regularization of the neural networks and to avoid over fitting problem,we apply Dropout, with a dropout rate of 0.1.The NN activation for cross-entropy loss as the loss function. The model was trained by using training set and used Test set to measure the performance of our model. To avoid over-fitting, the model training was allowed to run for a maximum number of 100 epochs.
In our results section we have compared Machine learning Baseline methods and Our Proposed Deep Neural Network Mode. In other part we performed Sentiment Analysis by using Deep learning techniques and used publicly available FastText [19] Urdu word embedding. We used standard metrics for classification accuracy,and to study and understand the problem we evaluated Precision and Recall. We used F1, the F-score is the harmonic mean of precision and recall Additionally we calculate ROC figure2 for each baseline and Deep Learning Method.In figure 2shows Deep learning Model Validation Accuracy. it Shows 0.95 Validation Accuracy and 0.0 Validation lost. Overall Our Deep learning model has revealedsignificant results see table 1. 0.92 Random forest 0.88 and Naive Bayes has lowest 0.77. As we mentioned before we used same dataset for both Machine learning and Deep Learning methods. According to Table1our Model performed best and surpasses the F1 and accuracy of all the Machine Learning baseline methods and achieved 0.95 Accuarcy and 0.94 F1 score.

Table 1. CLASSIFICATION RESULTS.

Classifiers	Precision	Recall	F1 score	Accuracy
NB	0.79	0.77	0.77	0.77240
RF	0.88	0.88	0.88	0.88691
SVM	0.93	0.92	0.92	0.92472
Our DL Model	0.97	0.9287	0.94	0.95180

4. Conclusion

In this paper, we used first time Long Short Term Memory LSTM neural network to train the model for Roman Urdu Sentiment Analysis, LSTM network is very efficient for sequential Data Models. Our Experimental results shows that deep neural networks is best Model to perform sequential data models, as it does not need any prior knowledge,designing and feature engineering. Our Model result has supasses the accuracy of Machine Learning Baseline and Lexicon Based Approaches. We suggest that LSTM networks with word embedding is great approach to perform Sentiment Anlysis.Our Model will help to ensure further exploration.

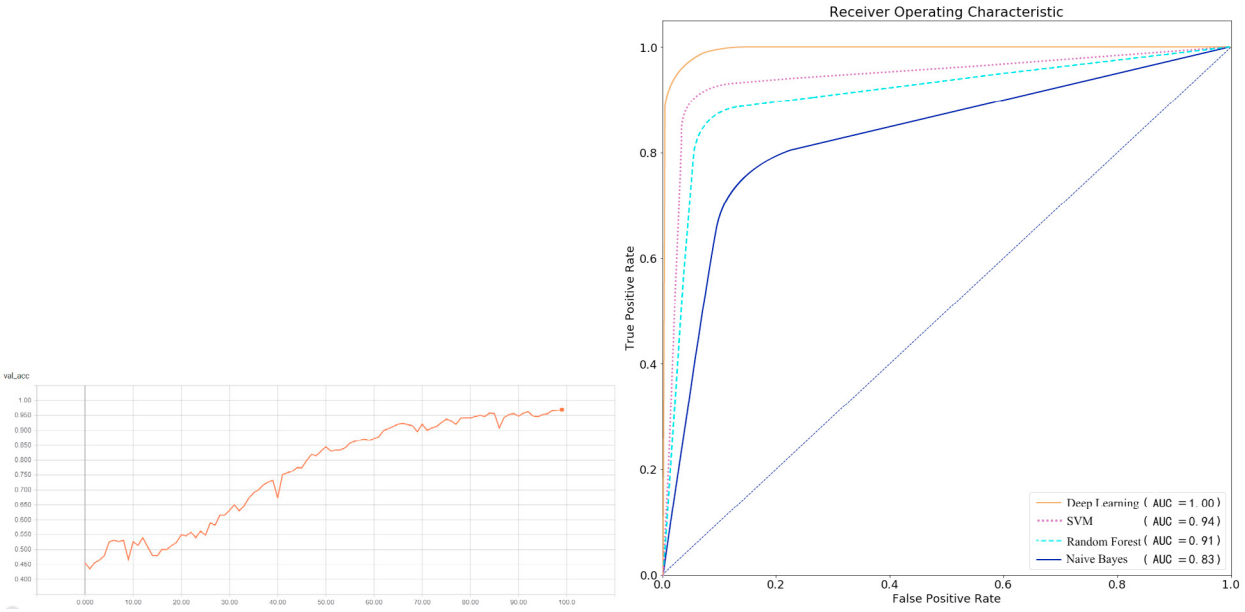

Fig. 2. Validation and Roc Score of Our Model

References

[1] F. Xing , E. Cambria , R. Welsch ,2018,Natural language based financial forecasting: a survey, Artif. Intell. Rev. doi:10.1007/s10462-017-9588-9.

[2] Laércio Dias,2018.” Using text analysis to quantify the similarity and evolution of scientific disciplines”, Royal Society.

[3] https://en.wikipedia.org/wiki/RomanUrdu.

[4] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig,2013. ”Joint language and translation modelling with recurrent neural networks. In EMNLP, volume 3,page 0.

[5] Zhang, 2003,“Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain”.

[6] Sepp Hochreiter and Jurgen Schmidhuber 1997,”Long short-term memory. Neural computation”, 9(8):1735–1780.

[7] Martin Sundermeyer, Hermann Ney, and Ralf Schluter,2015, ”From feedforward to recurrent lstm neural networks for language modeling. Audio, Speech, and Language Processing”, IEEE/ACM Transactions on, 23(3):517–529.

[8] Quan-Hoang Vo ; Huy-Tien Nguyen ; Bac Le ; Minh-Le Nguyen “Multi-channel LSTM-CNN model for Vietnamese sentiment analysis”, (KSE) IEEE,2017.

[9] Mehreen Alam,2017,“Sequence to Sequence Networks for Roman-Urdu to Urdu Transliteration” ,INMIC.

[10] Adnan Ul-Hasan ,2013,”Offline Printed Urdu Nastaleeq Script Recognitionwith Bidirectional LSTM Networks” International Conference on Document Analysis and Recognition.

[11] Mukund S, Srihari R Peterson E, 2010,“An Information-Extraction System for Urdu A Resource Poor Language.ACM Transactions on Asian Language Information Processing “(TALIP),9(4) 15.

[12] Syed AZ, Aslam Martinez Enriquez AM,2010,”Lexicon based sentiment analysis of Urdu text using SentiUnits”, In Advances in Artificial Intelligence Springer Berlin Heidelberg,32-43.

[13] Gule ZulfeNargis and Noreen Jamil,2016,“Generating an Emotion Ontology for Roman Urdu Text,”dline.info.

[14] Afraz Zahra Syed and Aslam Huhammad , 2010, “Lexicon Based Sentiment Analysis of Urdu Text Using SentiUnits,”Mexican International Conference on Artificial Intelligence Advances in Artificial Intelligence pp 32-43.

[15] M. Daud, R. Khan, Mohibullah, and A. Daud, 2014, “Roman urdu opinion mining system,”arXiv preprint arXiv:1501.01386, 2015 - arxiv.org

[16] M. Bilal, H. Israr, M. Shahid, and A. Khan, 2015, “Sentiment classification of roman-urdu opinions using naive bayesian, decision tree and knn classification techniques,”Journal of King Saud University - Computer and Information Sciences Volume 28, Issue 3, July 2016, Pages 330-344

[17] Zareen Sharf, Saifur Rahman, 2018,“ Performing Natural Language Processing on Roman Urdu Datasets,” IJCSNS International Journal of Computer Science and Network Security, VOL.18 No.1.

[18] Ann Irvine, JonathanWeese, Chris Callison-Burch ,2012,“ Processing Informal, Romanized Pakistani Text Messages, Center for Language and Speech Processing Johns Hopkins University” LSM ’12 Proceedings of the Second Workshop on Language in Social Media Pages 75-78.

[19] https://github.com/facebookresearch/fastText

[20] Wang Ling, 2015, ”Finding function in form: Compositional character models for open vocabulary word representation”.arXives.com.