International Journal of Engineering & Technology, 7 (2.27) (2018) 88-92

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Sentimental analysis using recurrent neural network

Merin Thomas [1] *, Latha C.A [2]

[1] Research Scholar, Regional Research Center, Visvesvaraya Technological University.
[2] Head of the Department (CSE), AMCEngineering College, Affiliated to Visvesvaraya Technological University, Bengaluru, India.
**Corresponding author E-mail:merin.jisso@gmail.com*

Abstract

Sentiment analysis has been an important topic of discussion from two decades since Lee published his first paper on the sentimental analysis in 2002. Apart from the sentimental analysis in English, it has spread its wing to other natural languages whose significance is very important in a multi linguistic country like India. The traditional approaches in machine learning have paved better accuracy for the Analysis. Deep Learning approaches have gained its momentum in recent years in sentimental analysis. Deep learning mimics the human learning so expectations are to meet higher levels of accuracy. In this paper we have implemented sentimental analysis of tweets in South Indian language Malayalam. The model used is Recurrent Neural Networks Long Short-Term Memory, a deep learning technique to predict the sentiments analysis. Achieved accuracy was found increasing with quality and depth of the datasets.

Keywords: *Sentimental Analysis; Deep Learning; Neural Network; Recurrent Neural Network.*

1. Introduction

Facts and opinions are the two major types of textual information. Expressions that are objective regarding an entity or event are called as Facts. Likewise, expressions that are of subjective nature are called Opinions. Opinions are so important that whenever a decision is to be made others opinions are always taken into consideration. It applies not only to an individual but also to organisations. The word sentiments or opinion is of great importance to humans. "What others think about" is the primary concern of an individual in society. When it comes to business impact is still bigger which intends to significant decisions. The word Opinion Mining, Sentimental Analysis is more relevant to future than the past. It has become the monetary measure for most probable business spanning out areas from product to politics.

E-commerce and social media have raised the level of sophistication of online users. There are varieties of platforms available on the Internet to express the opinions, share the ideas, emotion and interests. Twitter, Blogs, Facebook and Google Plus are most popular social media platform where instant views are broadcasted. Prevailing factors such as availability of opinion mining systems which can automatically classify and summarises users' reviews have also marked the importance of sentimental analysis. Finding opinion orientation in a piece of text concerning a topic is called sentimental analysis. The semantic direction of the content is critical information in the reviews or opinions, but current search engines are not providing it.

Sentimental Analysis being the research topic can be dated back on to 2001. Since then the major work was carried out in English, the language is widely accepted. Natural language processing gained the recent interest of researchers. Taking Malayalam- south Indian language as a case study, according to census 2001, 33,066,392 of Indian population speaks the Malayalam language. In the early days of last decade, Malayalam Unicode has widely adopted user-generated contents like websites, forums and blogs. There is an exponential increase in the amount of user-generated content in Malayalam as Malayalam Unicode key inputs are supported by almost all latest handheld devices currently.

In this paper, a sentimental analyser for the Malayalam Language based on neural network is proposed. The analyser classifies the positive and negative sentiments associated with each sentence. The analyser is based on recurrent neural network and long short-term memory that mimics the learning system in human brains. Layered approach of a neural network is to attain a much higher level of accuracy. The training of the model happens in several folds until target accuracy is achieved.

2. Related works

With the advent of machine learning algorithms, machines were likely to interpret natural languages with much greater accuracy. Statistical machine Translations were widely used in the natural language Processing [1]. Development of the parallel corpus was the main method used in statistical machine translation [2]. Machine Learning algorithms can be best evaluated while using it to parse Garden Path sentences or the sentences having complex structures[3].In the area of Sentimental Analysis, there are basically two basic approaches in sentimental analysis namely lexicon based and machine learning. Algorithms based on lexicon approaches compares sentiment words and seed words [4]. Corpus and Dictionary based are the prominent lexicon based algorithms. These approaches are relative simple to implement since Dictionary is already available for almost all languages. But it more suited for textbook Language rather than Natural Language. Supervised and Unsupervised algorithms [5, 6] categorization showed much differentiation in results and methods used in training and classifying datasets.

Supervised algorithms were based on features that need to be specified beforehand. The probabilistic model of Naïve Bayesian is one of the most popular algorithms [7]. Based on these features identified; rules are formed that classify the results. There are two phases of feature extraction, first, primary extraction of web data

and converting to the standard format and next phase forming the feature vectors [8]. Multinomial Naïve Bayesian is also a trusted model of Bayesian versions. Support vector machines are one of the baseline models for text classification that can be considered for full-length sentimental analysis [9].

Automatic content generation and increased use of social media with the fast-growing internet has made the task of researchers even more difficult. Feature Extraction and feature identification became chaos. The concept of unsupervised learning algorithms paved the way for better machine learning algorithms. Unsupervised algorithms were able to decide upon features in the process of training on the dataset and were able to perform well on exhaustive learning. As a result the new horizon of neural network and deep learning emerged. Basic classification of machine learning algorithms is shown below in Fig. 1.

On the other hand data to analyze can fall into two types. First one is structured and other is the unstructured data[10]. In case of unstructured data, it is important to normalize the data, to provide better understanding to machine [11].While working with language datasets, structural, lexical and semantic ambiguities are to be taken care [12].

3. Deep learning and neural network

Neural networks are used in deep learning in order to stimulate decision making like humans. Neural networks as the name indicate imitates the neuron in human brain. Schematic for a neuron in neural network is shown in the Fig. 1. Neuron cells spike the information to next neuron if activation signals cross certain threshold. Input and its corresponding weight is applied to a transfer function and in turn fed to activation function that generates threshold values.

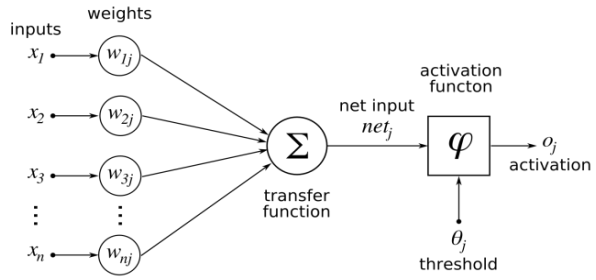

Fig. 1: Schematic for a Neuron in Neural Network.

Table 1: Classification of Algorithms Used in Sentimental Analysis

Sentimental Analysis	Lexicon Based	Dictionary Based	
		Corpus Based	Statistical
			Semantic
		Decision Tree	
		Rule Based	
	Machine Learning	Linear	SVM
			Neural Network
			Naïve Bayesian
	Supervised	Probabilistic	Bayesian Network
			Maximum Entropy
			CNN
	Unsupervised	Deep Learning	LSTM
			RNN

Programmed Activation neuron is a classifier called as perceptron which is capable of linearly separating the datasets. Neural networks can be applied to linearly separable data. The neuron can accept input vector V=[x_1, x_2.....,x_n] and their corresponding weights that decide the prominence of feature W= [w_1, w_2, w_n] respectively. The transfer function (a_i) is evaluated as the summation of the product of input element, and corresponding weights are given as $a_i = x_1w_1 + x_2w_2 + \ldots\ldots\ldots\ldots + x_nw_m + b$, Where 'b' is the bias. The activation function f (a_i) decides threshold attained, which is given below

$y = f(a_i) = f(\sum x_i w_i + b)$, where i=1 ton.

This model of neuron is to be trained with all possible datasets. Weights and corresponding bias is adjusted through learning to obtain minimum error, relative to target data.

4. Deep learning model for natural language processing

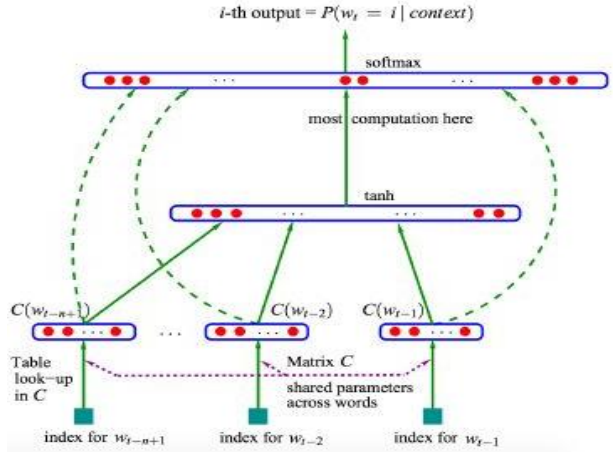

Fig. 2: Bagnio's Deep Learning Model for Natural Language Processing.

Bengio's primary model for Natural Language Processing is shown above in Figure 2. Representation of words in a distributed environment can be learned by using language modeling [13]. A sequence of words forms corpus which is then associated with a feature vector for its real value and probability function and its parameters depicting sequence. The network is then trained with this corpus.

Sentence Level Architecture of Natural Language Processing [14] is shown in Fig. 3 that takes the whole sentence as input. The sequence layer followed by pooling layer creates feature map of a fixed size. The feature map goes to neuron layers producing one tag for each sentence. The Network model is characterized by a dimension of word vectors, activation function and neural layers which are tuned to achieve the best performance for a particular task.

An end to end system that leads to output in just one step is basically called deep learning whose mathematical computation is relatively high taking huge time for training [15]. Deep learning paradigm can be categorized predominantly into convolution neural network and recurrent neural network. Convolution neural networks (CNN) are based on word embeddings. Single convolution networks with a language model cater semi-supervised learning of generalized shared task improving the performance of analysis [16]. Other versions of CNN include CNN- rand that use word vectors that are initialized randomly, CNN-static utilize vectors that are pre-trained, that are never updated, CNN-non-static is same as static, but word vectors are updated during training. CNN-multichannel makes use of two set of word vectors out of which one is updated, and other is not updated. A hybrid version of CNN and RNN framework is also used for better accuracy [17].

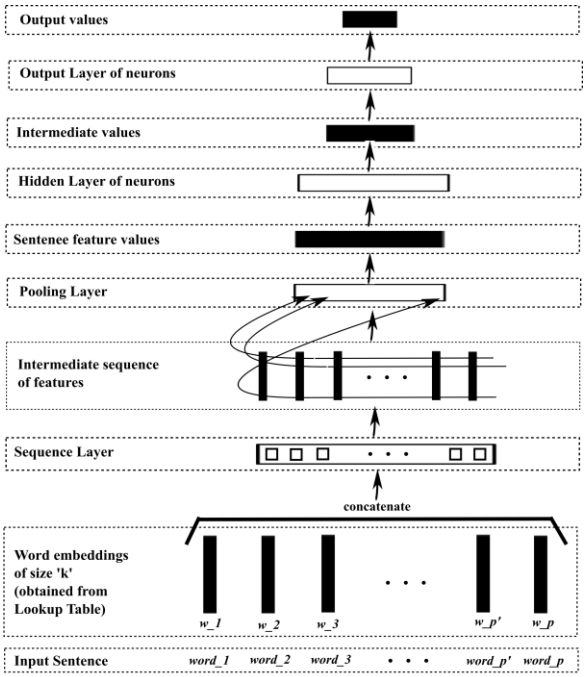

Fig. 3: Sentence Level Architecture of Natural Language Processing.

5. Recurrent neural networks (RNN)

Recurrent Neural Network Models [18] is shown in the figure 4 is one of the models of neural networks that do not depend on size about a window when natural language processing scenario is considered. RNN can keep the record of all the input values that are viewed by the network and also current input, which is value hidden at each layer network depend on all the previously seen inputs.

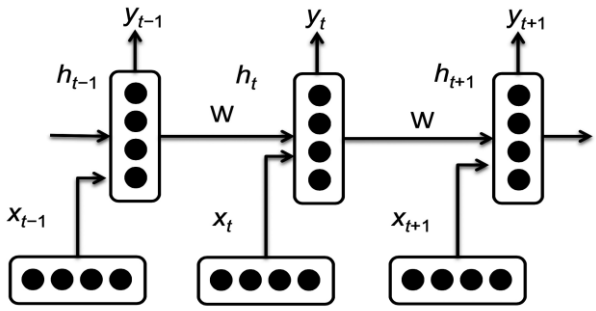

Fig. 4: Recurrent Neural Network with Three Time Steps.

Exhaustive training is to be done to this model and weights are to be adjusted so that higher accuracy is obtained [19]. Recurrent neural network in the scenario of sentimental analysis can also be modeled based on Sentiment Treebank [20]. Deep Recursive Neural Network is one of the newest architecture found by stacking multiple Neural Networks which can be used for fine-grained classification [21].

6. Long short term memory (LSTM)

Long Short-Term Memory networks [22] are the new version of the recurrent neural networks, but activation units are highly complicated. The memory cell is the major component of the unit. Two ways in which information can be stored are Short-term Memory that holds the recent history represented by activation function of the neuron cell, based on back propagation, weights that are modified is stored in Long-term Memory. Information can be stored for a longer term in LSTM model mostly from 10-12 steps of natural Language Processing Task. Better results are obtained if linear chain structure of LSTM is replaced by tree structure of the LSTM [23]. LSTM cell is shown below in figure 5.

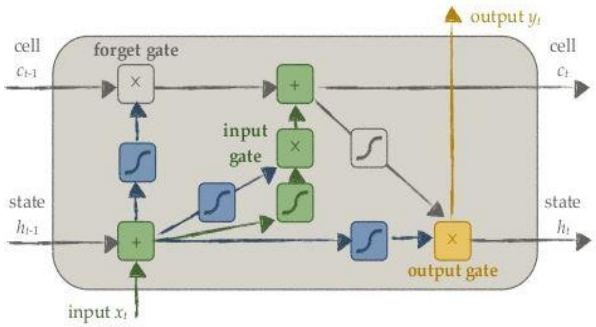

Fig. 5: LSTM Cell.

7. Methodology

The sentimental analysis system can address classifications of sentiments in to binary as well as fine-grained over user reviews and natural opinions in Malayalam. The slang of the language is also to be understood by the system that is used at various places. Better capturing of local syntax with the slang of the regions with meaningful collocations should be identified to improve the predictability of the sentiment analyser. The crucial component of the proposed system is the sentiment analyser. The paper shows the implementation of sentimental analyser using neural networks and LSTM for Malayalam datasets. The algorithm used is mentioned below.

Algorithm for RNN-LSTM

1) Open the file and read sentences and its tagged sentiments
2) Remove all punctuation and extra characters from the text
3) Split the sentences into individual words
4) Build a dictionary of words and map each word to integers starting at 1
5) Convert sentences in to array of integers
6) Convert the sentiment labels (Positive/negative), to integers 1 & 0 respectively
7) Truncate reviews whose length is more than 250 to first 250 characters.
8) Remove zero length reviews and left pad 0s to reviews shorter than 250.
9) Split the data in to Training (80%) & test (20%) data sets
10) Build the LSTM Graph with number of units in hidden layers in LSTM cells as 256, with 1 layer, batch size of 500 and learning rate of 0.01
11) Add an embedding layer with number of units in layer as 200
12) Create a LSTM cell for the graph with size of 256
13) Add dropouts to the inputs/outputs to the cell which wraps the cell within another cell
14) Create multiple LSTM layers with multi RNN Cell
15) Add forward pass through RNN by passing vectors from embedding layer
16) Convert training data to batches and feed in to model
17) Make test data in to batches and run through, and store accuracy percentage of each batch
18) Calculate Test Accuracy by taking average of batch accuracy values.

An output of the program is the sentiment of the sentence. The basic Program block is depicted in the Fig. 7 given below. Extensive training dataset was collected, and weights were associated with each sentence in the dataset.

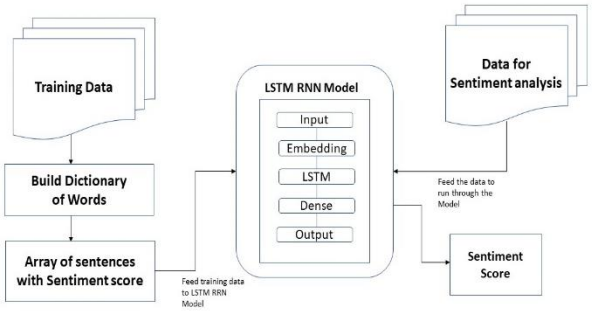

Fig. 7: Basic Program Blocks.

Extensive training dataset was collected, and weights were associated with each sentence in the dataset. Dataset is then fed to neural model. The dataset has training data and test data. Training data is textfile consisting of Malayalam language sentences that are positive and negative. It is always desirable that some positive and negative sentences in the training datasets are equal. Once training of the model is done, and desirable Batch Accuracy is obtained, Then the model can be used to predict the sentiments associated with the Malayalam sentence fed as an input with greater accuracy.

8. Experimental results and discussions

Input is fed in batches and accuracy of the model is evaluated each time in batches. The average of batch accuracy will be the actual accuracy of the model. Fig. 8 shows the snapshot of the test result. Analysis of the Result collected and used to train and test the model is highly significant. Fig. 9 shows the accuracy level attained. Initial batch Iterations have given low accuracy. Automatic selection of features from data makes deep learning techniques much better algorithm since it helps the programmer from the task of identifying the features. This is very relevant in the case when datasets grow and when numbers of features to be considered grow exponentially.

```
C:\Users\merin\AppData\Local\Programs\Python\Python36\python.exe C:/Users/merin/Py
25001
25000
Zero-length reviews: 1
Maximum review length: 2514
Zero-length reviews: 0
Maximum review length: 200
25000
<class 'numpy.ndarray'>
[    0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     0     0     0     0
     1   103    26  6060     2     1   226     6   367     8    13     3
   433     4     1  1931     4 19878     2 21040]
200
[1, 103, 26, 6060, 2, 1, 226, 6, 367, 8, 13, 3, 433, 4, 1, 1931, 4, 19878, 2, 2104
20
                     Feature Shapes:
Train set:           (20000, 200)
Validation set:      (2500, 200)
Test set:            (2500, 200)
label set:           (20000,)
Validation label set:  (2500,)
Test label set:              (2500,)
```

Fig. 8: Results Snapshots.

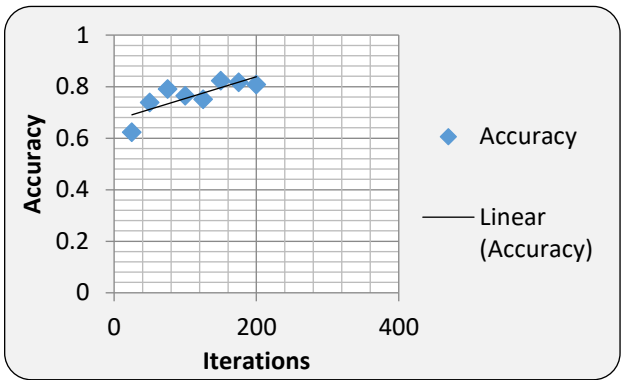

Fig. 9: Test Accuracy.

9. Conclusion

Sentimental Analysis is modeled by the methods of RNN-LSTM. The average accuracy of 80 % is acquired by this model. The accuracy of the results can be still improved by having deeper datasets. Comparative approach can also be done with various Deep Learning techniques. Deep Learning using RNN can also be used for the sentimental analysis of other language domains and also to deal with cross-linguistic problems.

Acknowledgement

I thank CHRIST Deemed to be University, Bangalore, for providing favorable environment for carrying out my Research work.

References

[1] Ali, Aasim, ShahidSiddiq, and Muhammad Kamran Malik. "Development of parallel corpus and english to urdu statistical machine translation." *Int. J. of Engineering & Technology IJET-IJENS* 10 (2010): 31-33.

[2] DEEP, Kamal; KUMAR, Ajit; GOYAL, Vishal. Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System. International Journal of Engineering & Technology**,** [S.l.], v. seven, n. two, p. 690-693, may 2018. ISSN 2227-524X.

[3] Du, Jiali, Pingfang Yu, and Minglin Li. "Machine Learning from Garden Path Sentences: The Application of Computational Linguistics." *International Journal of Emerging Technologies in learning (iJET)* 9.6 (2014): 58-62https://doi.org/10.3991/ijet.v9i6.4109.

[4] PrabhuPalanisamy, Vineet Yadav and HarshaElchuri, (2013). "Serendio Simple and Practical Lexicon Based approach to Sentiment analysis," Volume 2: Seventh International Workshop on Semantic Evaluation, Atlanta, Georgia, pages 543–548.

[5] I.Hemalatha, Dr.G.PSaradhi Varma, Dr.A.Govardhan(2013), "Sentiment analysis tool using machine learning algorithm", volume 2, Issue 2, International journal of emerging Trends and Technology in computer.

[6] Peter D.Turney, (2002), "Thumbs up and thumbs down? Semantic Orientation Applied to Unsupervised Classification", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, 417-424.

[7] Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. 2002. Thumbs up: sentiment classification using machine-learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86.https://doi.org/10.3115/1118693.1118704.

[8] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." *International Journal of Engineering and Technology* 7.6 (2016): 1-7.

[9] Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pages 90–94.

[10] Asha, T., et al. "Building machine learning algorithms on Hadoop for bigdata." *International Journal of Engineering and Technology* 3.2 (2013).

[11] Hanafiah, Novita, et al. "Text Normalization Algorithm on Twitter in Complaint Category." *Procedia Computer Science*116 (2017): 20-26.https://doi.org/10.1016/j.procs.2017.10.004.
[12] Sreelekha, S., Pushpak Bhattacharyya, and D. Malathi. "A case study on englishmalayalam machine translation." *Proceedings of the iDravidian* (2015).
[13] YoshuaBengio, RejeanDucharme, Pascal Vincent, and ´ Christian Janvin. 2003. A neural probabilistic language model. The Journal of Machine Learning Research, 3:1137–1155
[14] Ronan Collobert, Jason Weston, Leon Bottou, Michael ´ Karlen, KorayKavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 12:2493–2537.
[15] Satyanarayana P, Charishma Devi, Sowjanya P, Satish Babu, Syam Kumar, Implementation of conventional communication system in deep learning.International Journal of Engineering & Technology v.7 (1.1) p. 696-698, 2018. ISSN 2227-524X.
[16] Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.https://doi.org/10.1145/1390156.1390177.
[17] Hassan, Abdalraouf, and Ausif Mahmood. "Convolutional Recurrent Deep Learning Model for Sentence Classification." *IEEE Access* 6 (2018): 13949-13957.https://doi.org/10.1109/ACCESS.2018.2814818.
[18] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan ´ Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, pages 1045–1048.
[19] EPhzibah, E. P., and R. Sujatha. "Big data management with machine learning inscribed by domain knowledge for health care." *International Journal of Engineering & Technology* 6.4 (2017): 98-102.https://doi.org/10.14419/ijet.v6i4.8214.
[20] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP), volume 1631, and pages 1631–1642.
[21] OzanIrsoy and Claire Cardie. 2014. Deep recursive neural networks for compositionality in language. In Advances in Neural Information Processing Systems, pages 2096–2104.
[22] Sepp Hochreiter and JurgenSchmidhuber. 1997. ¨ long short-term memory. Neural computation, 9(8):1735–1780.https://doi.org/10.1162/neco.1997.9.8.1735.
[23] Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075.