Future Generation Computer Systems 95 (2019) 292–308

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Sentiment analysis through recurrent variants latterly on convolutional neural network of Twitter

Fazeel Abid [a], Muhammad Alam [b], Muhammad Yasir [a], Chen Li [a],*

[a] School of Information Science and Technology, Northwest University, Xian, China
[b] DCSE, Xian Jiaotong-Liverpool University, Suzhou, China

HIGHLIGHTS

- Convolutional Neural Network utilizes many layers learn to extract local features.
- Recurrent Neural Network abled to catch the long-term dependencies in single layer.
- GloVe domain-specific word embedding capably to realize the execution of models.
- It engages one layer RNN and one convolutional layer with global average pooling.
- Joint architecture for the sentiment analysis on small, medium and large datasets.

ARTICLE INFO

Article history:
Received 15 October 2018
Received in revised form 28 November 2018
Accepted 12 December 2018
Available online 8 January 2019

Keywords:
Sentiment analysis
Word embeddings
Recurrent neural network (RNNs)
Convolutional neural network (CNNs)

ABSTRACT

Sentiment analysis has been a hot area in the exploration field of language understanding, however, neural networks used in it are even lacking. Presently, the greater part of the work is proceeding on recognizing sentiments by concentrating on syntax and vocabulary. In addition, the task identified with natural language processing and for computing the exceptional and remarkable outcomes Recurrent neural networks (RNNs) and Convolutional neural networks (CNNs) have been utilized. Keeping in mind the end goal to capture the long-term dependencies CNNs, need to rely on assembling multiple layers. In this Paper for the improvement in understanding the sentiments, we constructed a joint architecture which places of RNN at first for capturing long-term dependencies with CNNs using global average pooling layer while on top a word embedding method using GloVe procured by unsupervised learning in the light of substantial twitter corpora to deal with this problem. Experimentations exhibit better execution when it is compared with the baseline model on the twitter's corpora which tends to perform dependable results for the analysis of sentiment benchmarks by achieving 90.59% on Stanford Twitter Sentiment Corpus, 89.46% on Sentiment Strength Twitter Data and 88.72% on Health Care Reform Dataset respectively. Empirically, our work turned to be an efficient architecture with slight hyperparameter tuning which capable us to reduce the number of parameters with higher performance and not merely relying on convolutional multiple layers by constructing the RNN layer followed by convolutional layer to seizure long-term dependencies.

1. Introduction

Over the most recent years a phenomenal development in online life, for instance, Facebook, Twitter, LinkedIn, Instagram and some more have been utilized among these, Twitter is with 319 million dynamic users.[1] These stages can be a remarkable source of information for some tremendous affiliations that empower them to acquire from this covered information for better achievement, yet due to the immense mass of information abides as customers, posts, comments, messages, marking, minutes, following general customer feelings can be exceedingly intricate. The data on the web as the substance is classified in the form of factual and sentiment where, factual is the target wordings, concerning elements, issues or occasions whereas the sentiment is the subjective phrasings, like assessments, convictions, different preferences. There are various levels (coarse to fine) on which to perform the classification related to sentiment. Sentiment analysis of Twitter information is completed on a sentence level, which is in the middle of the representational level of sentiment, there are two sorts of sentiment present in a sentence are direct and comparative.

Sentiment classification is an errand for natural language processing (NLP) with a large number of applications, for example, techniques of storing and recovering information, grading of the

* Corresponding author.
E-mail addresses: fazeelabid@hotmail.com (F. Abid), m.alam@xjtlu.edu.cn (M. Alam), yasirnwu@outlook.com (M. Yasir), chenli@nwu.edu.cn (C. Li).

[1] https://about.twitter.com/company.

https://doi.org/10.1016/j.future.2018.12.018
0167-739X/© 2019 Elsevier B.V. All rights reserved.

Table 1
Exhibition of preferences and hindrances of relevant literature.

Literature reviews	Advantages	Disadvantages
[3,4]	With the regard of productive sentence structure and its efficient representation, a bag of words or N-grams models is utilized then training through linear classifier on account of many supervised methods likewise in Naïve Byes, Support Vector Machine, Random Forest, Maximum Entropy and logistic regression.	These methodologies, such as a bag of words lack with words order and omit the information about semantics though word N-gram encounters data sparsity besides, on account of features and classes, these classifiers unfit to share parameters.
[5]	Lexicon based techniques which determine sentiment polarity of a tweet or sentence by checking them from pre-established dictionaries with weights to recognize their sentiment orientations, such as SentiWordNet	In the case of Lexicon methods, we need to depend on the nearness lexical features through we can get the sentiment behaviour unequivocally while in social media normally a sentiment in a tweet is implicitly connected with the semantics.
[6]	Rather than just depending on an unlabelled dataset which is vast in size, the training algorithm is capable to discover inward representations which in swings to developed NLP tagger for all the purposes.	Despite the fact that multilayer neural system design that can deal with various NLP assignments with both speed and exactness, yet needs to avoid task-specific architecture however much as could be expected and also exploiting man-made data features.
[7]	Investigated a few expansions of the skip-gram model that improve both the nature of the vectors and the training speed. By subsampling of the incessant words, we acquire critical speedup and furthermore take in more levelled word representations. Additionally, a simple alternative to the hierarchical softmax called negative sampling is portrayed.	It is crucial with the choices that influence the execution which is dependent to model architecture, for instance, vectors and training window measure, the rate of sub-sampling.
[8]	Proposed the system that is skilled to deal with information sentences of differing length and prompts feature(element) chart over the sentence that is prepared to do expressly catching short and long-run relations. Moreover, autonomous to a parse tree and can be connected to any dialect demonstrate.	Use of Multiple convolutional layers alongside the use of a few activities identified with dynamic pooling is required which then able to persuade a feature graph that frames various levelled structure somewhat like to that in a syntactic parse tree while it is not tied to virtuously syntactic in associations.
[9]	The author tended to semantic word spaces in the respect to express the implication of lengthened articulations by catching the compositional impacts proposing another model Titled the Recursive Neural Tensor Network (RNTN) that is outfitted with satisfactorily catching the negation for both positive and negative expressions with an extension at different tree levels.	Polarity discovery in sentences most likely required more extravagant regulated supervised training and evaluation resources and all the more great models of structure towards understanding compositionality in NLP undertakings.
[10]	Introduced a basic design of CNN rather than numerous layers, and the hypermeter with tuning with the utilization of pre-prepared word vectors through un-supervised leaning accumulated remarkable results in deep learning NLP Tasks.	The main downside related with this model it does not perform well on random initialization of words, in any case, can be enhanced to some degree require more layers, further, by sampling each dimension so that randomly initialized vectors have the indistinguishable fluctuation from the pre-prepared vectors, likewise, model performance was bad prepared the vectors as a major aspect CNN Training .
[11,12]	Proposes a 1D structure specifically a word arrange of content information for an exact estimate by applying CNN on textual data with high-dimension, which prompts specifically learning to embed little text regions for the purpose of characterization.	It was a direct adjustment of CNN from picture to content, a basic however requires a new variety which utilizes bag-of-word conversion in the convolution layer. Besides, an extension which consolidates various convolution layers is additionally required for higher precision.
[13]	The arrangement to extricate features from the character- to sentence level utilizing convolutional neural networks in this model including unsupervised pre-training; to identify nullification with the viability and commitment of features identified with a character as well as sentence level.	The model explained with the requirement for features extraction usage of multiple convolutional layers from words and sentences irrespective of its size which requires techniques that go beyond bag-of-words Additionally, to fill the hole of relevant data researched to utilize strategies that can exploit the prior information contained vast arrangements of unlabelled content.
[14]	This work effectively handles the unpredictability in building the dictionary and feature resources by utilizing a distributed representation of words with a model which is composed of Bi-directional LSTM in a new way by training the model with Distributed word representation, furthermore from POS label data to handle the syntactic and semantics in the text.	This work just spotlights on the superiority of sentiment classifier using BiLSTM with an additional weight on the system via preparing in two different ways, such as Distributed and POStag which is then additionally required to distinguish the semantic relationship among lexicon and features.
[15]	This model concentrated on the changing conduct of users towards opinions, refutation and particularly the words with their intensity by using sequential LSTMs not relying on phrase level annotation and structure of parsing tree.	This works as simple as it lacks the modification scope of words with their intensity and refutation, Although utilization of bi-directional LSTM through minimization operator.
[16]	In this work, experimentation through CNNs with LSTMs which are trained through various sorts of distributional word representations, without joining any costly, extensive features for the claim categorization can likewise be effortlessly adjusted to other embeddings to classify the text from any level (Sentence-Document)	Upgrades will just conceivable with the use of more extravagant semantic embeddings, for example, genuine and word sense embeddings ought to be trained from a broad corpus additionally requiring the joining between suggestion predicate associations.

web and so forth. It is necessary to process the text, remembering the end goal to analyze it and information extraction in an unexpected way. Text classification today inquires about to begin by outlining the best feature extractor to picking the idea like in most machine learning classifiers. There are very well-known techniques towards exemplifying a sentence by utilizing Bag of words (BOW), then classifying through the Support Vector

Table 1 (*continued*).

Literature reviews	Advantages	Disadvantages
[17]	Proposed the heterogeneous MSC recurrent random walk network in a distinctive way by exploiting both the users' posted tweets with social relations while training end-to-end through back-propagation method from scratch.	In contrast to the past investigations, Especially with the use of recurrent random walk network learning in this model, Still exist the issue in the regard of microblog sentiment classification through network embedding, which exploiting semantic representation and social circle of the user connection to enhance the precision.
[18]	This paper intends to see steady numerical characteristics which are in the form of multiple dimensions, for example, valence arousal (VA) space insinuated as sentiment analysis, a regional CNN–LSTM demonstrates by means of catching regional information that exists in articulations with long-term dependencies transversely over sentences.	Notwithstanding the way that regional CNN and LSTM anticipate the VA evaluations of texts in a good manner however with some hustles like first-word vectors from vocabulary through embeddings, then regional CNN fabricate text vectors after that information is consecutively coordinated crosswise over regions using LSTM. So a blend is required for Regional CNN and LSTM although still there is a necessity to the utilization of a parser to recognize areas.
[19]	The introduced neural network model that is a combination of CNNs and RNNs alluded as a general encoding architecture that utilizes both neural network layers to proficiently encode character inputs with significantly less convolutional layers contrasted with the typical convolution architecture which is not merely design for document classification tasks or regular dialect inputs.	The proposed model, for the most part, performs better, in any case, when various classes are substantial, the preparation measure is little, and when the quantity of convolutional layers is set to a few (2–3).
[20]	For the consideration of lexical and semantic information identified with emotional expressions HMIO based on bi-directional recurrent neural network is introduced that is composed of two Bi-GRU independent layers for the representation and generation of sentence with parts of Speech respectively than for the consideration of lexical information through attention by the output of softmax instigation in the respect of attention for featuring the word.	Like the common deep learning models, this model additionally comprises two layers which are named as input and output. Moreover, supporting tags are set in the central layer to limit the learning of POS, while the central yields are also viewed as part of a hidden feature for anticipating concluding sentiment tag by consolidating this mentioned embedding layer into the respective layer for updatable training further, has to engage independent Bi-GRU layers by applying for the representation of the content words and parts of speech.
[21]	For the effective expectation stock prices triggered with stock market an inventive work on neural networks dependent on Long–short term memory with embedding layer and automatic encoder respectively to tackle the initial weight of random selection through the specialized stock vector which is equipped with high dimensional historical data of the multi-stock.	In spite of the fact that the models demonstrate by achieving dependable precision for the upgrades towards forecast of stock market prices to a specific degree, yet, a few insufficiencies are found in the part of the contribution of historical data so there ought to be considered to incorporate a factor of textual information augmentation on other stock exchanges.
[22]	With the end goal to watch and foresee user trajectory with the time coordinating the emotions of a human with an intelligent framework, a hybrid intelligence infrastructure-based cloud which use (AV–AT model) for the enthusiastic representation of users information a technique is utilized by which is referred to as "genetically optimized adaptive Fuzzy Logic" be that as it may, conceivable with the conduction analyzes in broad scale.	The model is constrained in the necessity for discrete focuses which ought to be in time for the estimation of the essential components of expectation. We have to demonstrate a cloud computational intelligence infrastructure which is fit to coordinate the feeling demonstrating as recommended in the model which requires extensive scale tests completing information handling, assumption examination, and capacity on information containing one million Facebook clients
[23]	Depp Neural model which includes pre-prepared unsupervised learning of features through GloVe errand of catching the contents recurrently and development by utilizing the convolutional neural network for the text representation by using convolutional and max-pooling operations.	Although the model is effectively retaining the order of information specifically with words alongside the managing the issue of information sparsity, catch more effectively the relevant information of the sentiment errand yet it is conceivable of the hidden layer which comprises of three convolutional and max-pooling layer.

Machine (SVM) or Logistic Regression (LR) [1,2]. Few of experiments in view of joining neural networks have progressed towards becoming progressively prominent recently, it has turned out to be imaginable to prepare more complex models on the considerably extensive datasets. Deep neural network techniques mutually actualize a feature extraction along with classification. Overall, neural network based methodologies get a document as an input which is an exemplification of the sequence of words as the representation of hot vector by multiplying with weight. After feeding into the neural network, which forms numerous layers in a sequence bringing about the prediction of childhood [24].

Neural networks, such as convolutional neural networks, recursive neural network and recurrent neural network have achieved comparable results when applied to text sentiment. These models expect experts to determine the correct model and set going with hyper-parameters, having a size of the filter. Some experimental works used recursive neural networks to perform sentiment analysis on the sentence level, while some proposed a structure based on the tree by utilizing the capabilities of long–short-term memory networks (LSTMs) to enhance the semantics [25]. Contemporary work done by [26] is comprised of a multi-layer of the convolutional neural network alongside with max-pooling operation, observed that utilizing convolutional to catch the additional dependencies need numerous layers [27] explored the mix of the neural networks of the convolutional neural network with the recurrent neural network with an objective to encode character input, executed to learn the input sequence of the high-level feature. Among the many models, an effective model associated recurrent neural network for NLP was proposed by [28]; it affirmed that it is prepared to catch the long-term dependencies even on account of a single layer. The enhanced execution of these described methods essential benefits by the accompanying angles: The distributional vectors with high dimension supply comparative semantic-situated words with high comparability. The convolutional neural network is ready to take local features from words or expressions in the best places of the text. It takes words in a sentence in a successive order and can take long-term dependencies of text instead of local features.

Our work embraces of carrying out sentiment analysis by using deep learning techniques with the intent to classify tweets using RNNs at initial because it recurrently and sequentially managing each word in a sentence which is useful to locate a dense and low dimensional semantic representation on account of sentence embedding and after that feeds to CNNs which could be very helpful for who wants to rapidly scale the sentiment by employing these two networks together accordingly. To perform sentiment analysis identified with a deep neural network specifically in the language model, we found that RNNs as an option for pooling layers because pooling only extracts important features while ignoring the others which is in result compromises on evaluation measures. Convolutional layers extract local features since they bound the receptive fields of the hidden layer. In addition, it has special spatially local correlation by enforcing a local connectivity pattern which is in between neurons of adjacent layers, which is useful to locate local indications with respect to the class. Bi-LSTM, GRU and Bi-GRU are the variants of the recurrent neural network, which is obliging in sequential data processing in texts due to its capacity to make use of long-term information, while the convolutional layer is able to locate local indicators, regardless their position.

This paper acquaints a join framework distinctively which is in between the recurrent neural network variants such as Bi-LSTM, GRU and Bi-GRU and afterwards connected with the convolutional neural network as shown in Fig. 1. In our work, aiming for the mapping of features, pre-trained word embedding using GloVe take as input, which contains windows of various lengths and weights feeds to the recurrent neural network. Having the abilities of RNNs to learn long-term dependencies, it can be seen as a representation of the phrase. The yield of the recurrent neural network viewed as a contribution to the convolutional neural network and global average pooling layer. The deep learning technique that we implement also comparatively differs from many existing methods based on first applying various weights and length windows while pooling moves left to right rather than in the whole phrase resultant features in our convolutional model has sequential information. Besides taking the favourable position of the encoded features which is classified from the convolutional neural network and capabilities owned by the recurrent neural network. Keeping in view, we present the sentiment analysis of twitter because it is an interpersonal interaction site that is growing with the use of short forms, and diversity of languages makes it a more challenging task. Following are the main contributions of this work:

1. Initialization of the domain-specific word embeddings by unsupervised learning using GloVe [29] which is trained on large twitter corpus.
2. We take the word embedding firstly as the input to our deep neural architecture in order to capture the long-term dependencies initially with RNN model utilizing its variants Bi-LSTM, GRU, Bi-GRU.
3. In the previous stages, all parameters related to the network along with the word embeddings by utilizing the same architecture, additionally fed to the convolutional neural network with global average pooling, which is capable of containing the windows various length and weight for the generation of the number of feature maps. Results showed that our architecture with few of parameters attains modest results.

The rest of this paper is organized as follows. Section 2 proceeded with the related work. The architecture of neural networks for the analysis sentiments is presented in Section 3. While Twitter dataset description is in Section 4. Section 5 contains the procedure of experiments. The discussion is in Section 6. The paper is concluded in Section 7.

2. Related work

The majority of the current work on Twitter sentiment classification can be clarified in the respect of supervised methods [3,4] which are based on training through linear classifier likewise, in Naïve Byes, Support Vector Machine, Random Forest, Maximum Entropy and logistic regression utilizing feature mixtures such as Word N-grams, bag of words, Parts of Speech Tagging and Twitter specific features which are capitalized words, emoticon, and hashtags, however, these methodologies, such as a bag of words lack with words order and omit the information about semantics though word N-gram experiences information sparsity besides, on account of features and classes, these classifiers unfit to share parameters. While lexicon based techniques which determine sentiment polarity of a tweet or sentence by checking them from pre-established dictionaries with weights to recognize their sentiment orientations, such as SentiWordNet [5] whereas in many cases a tweet is verifiably connected with the semantics instead of expressly.

Deep learning strategies are currently settled in machine learning for the image and voice recognition accomplishing great exhibitions. Many techniques have started to update models for natural language processing. It has turned out to be more typical in various natural language processing applications; awesome work included word vector representation through neural networks and perform classification by avoiding task-specific engineering [6]. To represent the word and real vector and to catch words closeness of two vectors by utilizing the distance between them word embedding by utilizing subsampling of frequent words but critical to handle different parameters like size of training window, vector and rate identified with sub-sampling, as a part of [7]. Various procedures have been effectively proposed for the task related to sentiment classification specifically for the semantic modelling, for instance, the convolutional neural network has multiple layers, positioning latent, dense and low dimensional word vector as input prescribed in [8]. Recursive neural tensor, which represents a sentence as word vector and parses the tree for computation of the vector for the purpose of high nodes in the tree utilizing similar tensor capacity for only the capturing of compositional effects endorsed in [9].

Another examination presented in [10] on the classification of the sentence obtains remarkable results using pre-trained word vector representation involving the convolutional neural network but low with random initialization [11,12] investigated text classification and obtains the best outcomes for the sentiment analysis, yet the model is intricate and costly to prepare in light of the fact that they apply the convolutional neural network to high dimensional text. Later on, another model which is the same just the replacement of high dimensional vector representation as input to convolutional neural network concentrating on the extensive text rather on sentences. For the execution of sentiment classification based on short texts a new deep convolutional neural network, which deal with limited contextual data from character level to sentence level by using two convolutional layers proposed in [13]. For the representation of the distribution of words a Bi-LSTM network model for the Japanese language performing sentiment analysis which is capable of handling syntactic as well as the semantics of text and also requires the training of POS-tag data proposed by [14].

A model in light of regularization of Long-term-short memory comprising of linguistics like intensity, polarity, and negativity in words to deal with the sentiment exhibited by focusing on phrase level explanation and not relying on parsing tree structure in sentence explained by [15]. Classification of sentences that are factual or related to sentiments made out of a blend of deep neural networks like long–short-term memory and CNNs employs different embedding of linguistics and word2vec but without intertwining expensive and deep feature set introduced in [16]. For the reason identified by the polarity detection in respect of semantic representation of tweets, a random walk layer (RWL) approach in view of recurrent concerning tweets that were the post and their respective network proposed by [17]. To anticipate the valence arousal ratings for dimensional polarity classification of text, comprised of long–short-term memory and convolutional neural network regional, a model known as regional CNN–LSTM excluding the use of parser for the identification of regions discussed in [18]. A methodology by [30] in the regard twitter based sentiment characterization in boosting way utilizing an alternate mix of measurements as meta-level features in numerous angles, for example, strength, emotion and factors indicating the polarity identified along-with opinion, which is generated by several ongoing sentiment classification architectures and resources yet it is restricted to emotion-oriented features which will in general increase low estimations of data resultant a poor conclusion characterization since emotions are with multidimensional objects while for the classifying tweets is one single dimensional task which is category based. Numerous Tasks, for example, characterization of the picture [31] and Recognition of Speech [32] included the combination of CNNs and RNNs. A model builds on CNN–RNN for capturing sub word information, a high-level feature input arrangement related to character level executed by [19] functions admirably only in the case of accessibility of several numbers of classes. Recently a model which uses independents BD-GRU (Bi-Directional Gated Recurrent Unit) layers for phrase Level demonstration and POS with the intention for polarity classification reply on reviews of the customers collected from www.jd.com with emotional expression namely hierarchical multi-input and output model (HMIO) presented by [20]. For the effective expectation stock prices triggered with stock market an inventive work on neural networks dependent on Long–short term memory with embedding layer and automatic encoder respectively to tackle the initial weight of random selection through the specialized stock vector which is equipped with high dimensional historical data of the multi-stock by [21]. With the end goal to watch and foresee user trajectory with the time coordinating the emotions of a human with an intelligent framework, a hybrid intelligence infrastructure-based cloud which use (AV–AT model) for the enthusiastic representation of user's information a technique is utilized by [22] which is referred to as "genetically optimized adaptive Fuzzy Logic" be that as it may, conceivable with the conduction analyzes in broad scale. Additionally, for the training of many complex models applied to extensive datasets which have been used unsupervised learning through word2vec proposed in [33]. For the enhancement in the proficiency of representation, a model called differential state framework is employed by [34] to hold long-term memory by learning among moderate and quick evolving information-driven representation. To propel the exactness of the deep learning method for sentiment analysis by fusing domain knowledge a methodology by utilizing regression and weighted cross-entropy as upgraded loss function explored by [35]. Presently, there are many techniques however, we are motivated by work effectively done on neural networks in the respect of natural language processing proposed in [23] in which pre-prepared unsupervised learning of features through GloVe errand of catching the contents abled with the recurrent structure and development utilizing the convolutional neural network for the text representation by using a numerous convolutional and max-pooling layer.

It is important that the convolutional neural network utilizing numerous layers can merely learn to extract local features that resided well inside the content, yet it fails to catch the long-term dependencies due to the locality of the layers; Convolutional and pooling which composes a very deep neural network serving with many convolutional layers likewise with the length of input that is straightforwardly corresponding to layers, so the recurrent neural network has the capacities to address this issue through sequential modelling which in turns performs better outcomes when the network is trained through domain-specific word embedding which incredibly influence on the general execution of neural networks in numerous viewpoints. For the stipulation of this paper which is concomitant with the necessity of our work bring into line with the aforementioned literature with some more inconspicuous components are exhibited in the underneath table.

3. The architecture of neural network for the sentiment classification

In this section, we will examine the architecture of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) for the purpose of classification of sentiments on twitter dataset. We intend to build a single model of recurrent neural networks with the convolutional neural network.

The recurrent neural network has a capability of its own memory cell which can be utilized to process the input in a sequential manner. Being equipped with its cell, this neural network, for sequential information, a word in a sentence from left to right can be observed. In a convolutional neural network, several layers are used to extract the local features which have significant characteristics related to many text classification tasks, especially in natural language processing. A model with the combination of recurrent neural network and convolutional neural network performs better when diverged from past models in which the output of the convolutional neural network over the characters is managed as input to the recurrent neural network proposed in [36]. For the best of our review, recurrent neural networks with the ability of its own memory can handle the long-term dependencies and work extremely well on sequential input, On the other hand, convolutional neural networks replacing pooling layer with global average pooling used for the extraction of local features.

The core idea behind is that the output of tokens will not only entirely able to store information on the initial token, additionally, stores the information of the previous token. RNN layer which will get sentence matrix by word embedding as input for each token. RNN layer is generating a new encoding for the original input. The output of the RNN layer is then taken into a convolutional layer which will abstract local features. More precisely, the convolutional and global average pooling layer's output will be then pooled to a smaller dimension and ultimately output is produced by the classification layer; softmax.

Our proposed model has the succeeding portions: Word embedding and Representation of words as features, RNN Layer (Bi-LSTM, GRU, Bi-GRU), Convolutional and (GAP) Global average pooling layer and Softmax Output as shown in Fig. 1.

3.1. Word's representation as feature

The most recent approach used for the word representation which contains all words in a given vocabulary are plotted and described as an n-dimensional vector. There are multiple examples of neural networks that employ the use of word embedding to solve problems related to text classification [37–39]. Unsupervised learning means pre-trained like word2vec, GloVe and random initialization of word embedding are typically drilled. Our experiments performed word embedding using random initialization and also by using unsupervised learning GloVe; which is a global log-bi-linear regression model used in various models for named entity recognition, analogies of words etc. Unsupervised learning like GloVe utilizing the words by finding similar words for the purpose

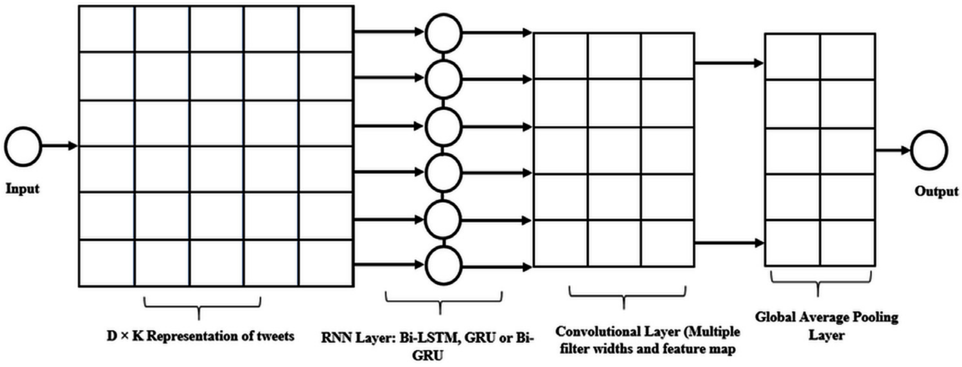
Input
Output
D × K Representation of tweets
RNN Layer: Bi-LSTM, GRU or Bi-GRU
Convolutional Layer (Multiple filter widths and feature map
Global Average Pooling Layer

Fig. 1. Model architecture based on GloVe-RNN–CNN for classification of sentiments.

of representation of the vector. Representation of the tweets as semantic features can be accomplished by utilizing the unsupervised learning such as GloVe. For the purpose of the training vector specialized tools related to Glove presented in [29] and concerned model prescribed in [40].

We prepared the GloVe using an extensive pre-trained corpus containing 2 billion tweets, 27 billion tokens and 1.2 million vocabularies. Further, we decide to interpret the words in a tweet containing in our dataset as a 200-dimensional vector.

Let "T" for the tweet and for the tokens "m", GloVe produced a table of word vector $W_v \in R^{D\times V}$ for mapping purposes, which is to be used for comparison related to each token that presents in the tweets where "D" represents dimensions, W_v for word vector and for vocabulary "V" is used. Subsequent to mapping each word "w" as $w_1 \in R^D$.

The feature vector F_v for a tweet "T" concatenated "$\oplus$" with word embedding is as per the following:

$$F_v = w_1 \oplus w_1 \oplus w_1 \ldots \oplus w_m \tag{1}$$

3.2. Recurrent neural network

Recurrent neural networks with the vital idea to utilize the information in a sequential manner. In these Networks, a word with a specific time step in the sequence is connected. The maximum sequence of the input length in the resultant is proportionate to the time steps amount. From its name, recurrent clearly delineates that for a comparable task identified with the component in the grouping, there is a requirement for the output of the past calculation as it stores all the calculations in its own particular memory cell. The formal architecture of the recurrent neural network shown in Fig. 2.

So, tweets containing the set either negative or positive polarity, as feature vectors subsequent to changing feed as an input which is given to the recurrent neural network (RNN) layer is given below:

$$hs_t = f(X.I_t + W_m hs_{t-1} + b) \tag{2}$$

$$O_t = softmax(Y.hs_t) \tag{3}$$

where,

- I_t For time step t, is an input, for the respective word in the sentence that can function as a hot word vector.
- O_t For time step t, it is an output.
- f: A non-linearity function such as hyperbolic tangent (tanh) or Rectifier Linear Unit (ReLU), however for the first hidden state, typically initialized to all zeros which are represented by hs_{-1}.

From the above figure, outputs can be seen by the respective time step, however, relying upon the errand this may not be fundamental. In the case of sentiment detection, our goal is to focus on the final output, not with the concern to recognize the polarity in every word identified with a specific sentence. In a recurrent neural network layer, the same case in which there is no requirement of input with the respective time step moreover, with hidden state, it is able to accumulate sequential information. However, researchers found that greater improvement in the recurrent neural network can be refined by presenting LSTM (Long–Short-Term Memory Network) Bi-LSTM (Bi-directional LSTM).

3.3. Long–short term memory

It is usually known as "LSTMs" and designed to deal with the long-term dependency problems. Long–short term memory network can address the vanishing gradient problem [39] with this layer when it is trained to utilize backpropagation with the time. It "LSTM' is composed of three internal Gates that control the stream to and from the memory blocks as shown in Fig. 3, how to optimized hidden states hs_t and result R_t in the following equations:

- **Gate Vectors**:

$$i_t = \sigma(C_i.[hs_{t-1}, I_t] + P_i) \tag{4}$$

$$o_t = \sigma(C_O.[hs_{t-1}, I_t] + P_O) \tag{5}$$

$$f_t = \sigma(C_f.[hs_{t-1}, I_t] + P_f) \tag{6}$$

- **Adaption**:

$$\tilde{R}_t = tanh(C_R.[hs_{t-1}, I_t] + P_R) \tag{7}$$

- **Update State**:

$$R_t = f_t * (R_{t-1} + i_t * \tilde{R}_i) \tag{8}$$

$$hs_t = o_t * \tanh(R_t) \tag{9}$$

From the above C and P represents cell parameters; i_t, O_t, f_t are gate vectors while the sigmoid function is σ and input vector is I_t.

3.4. Bidirectional long–short term memory

In spite of the fact that to deal with sentences which have syntactic as well as semantic information, LSTMs is thought to be an appropriate methodology. However, its architecture merely relies on the previous context which can be a cause of not functioning well in some cases especially in the text classification tasks where the coming context is additionally imperative in order to understand the structure. Bidirectional LSTMs has contributed a vital job in numerous zones particularly identified with content grouping issues [41,42]. Accordingly, bi-directional LSTM has the ability to capture both contexts; previous and upcoming with its

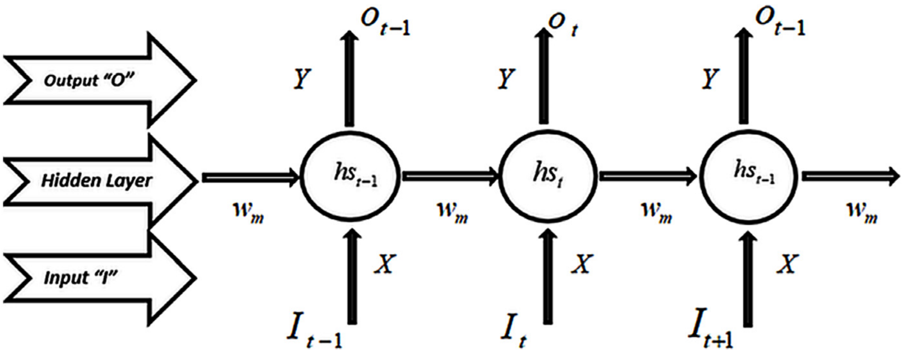

Fig. 2. Recurrent neural network architecture.

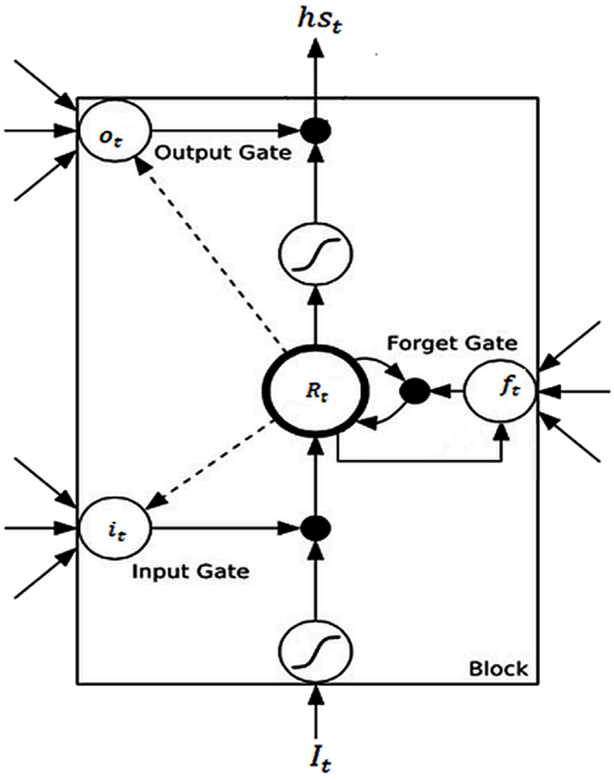

Fig. 3. Long–short term memory block.

two hidden states forward $\overrightarrow{hs}_t$ and backward $\overrightarrow{hs_t}$ from the LSTMs result, it finally produces z by LSTMs by iterating as follows.

- **Forward Layer**:
  $s_t = 1$ to ST
- **Backward Layer**
  $s_t = ST$ to 1

$$z = W_{\overrightarrow{hsz}} \overrightarrow{h}_{st} + W_{\overleftarrow{hsz}} \overleftarrow{h}_{st} + b_z \tag{10}$$

3.5. Gated recurrent unit

Gated Recurrent Unit (GRU) is another variant of Long–short term memory which is proved to be better than a model which is purely based on regular LSTMs. Gated Recurrent Unit (GRU) has only two gates by consolidating the input gates and the forget gates presented in [43,44] when contrasted with LSTMs which contains three gates namely input, output and forget gates respectively. Furthermore, it combines the hidden state with the cell state as shown in Fig. 4 and the following equations are used to update the parameters.

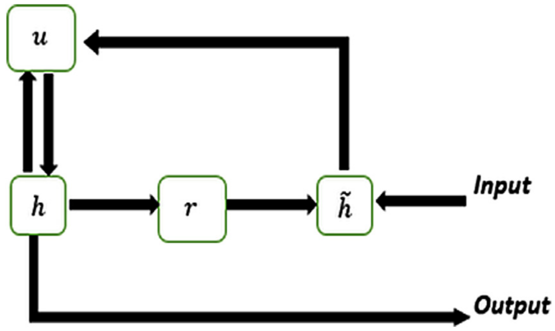

Fig. 4. Gated recurrent unit.

$$r_g = \sigma(W_r I_t + Xrhs_{t-1}) \tag{11}$$

$$u_g = \sigma(W_u I_t + Xrhs_{t-1}) \tag{12}$$

$$\tilde{h}_t = tanh(WI_t + X(r_g \odot hs_{t-1}, x_t)) \tag{13}$$

$$hs_t = \left((1 - u_g)hs_{t-1}\right) + u_g \tilde{h}_t \tag{14}$$

where σ Logistic Sigmoid function, Candidate hidden layer represented by $\tilde{h}_t$, $\odot$ Element Wise Multiplication, r_g shows reset gate while update gate representation in u_g.

3.6. Bidirectional gated recurrent unit

In the regard of recurrent neural networks, for Instance, GRU, a bi-directional gated recurrent unit is another form of long–short-term memory architecture, however, quicker than LSTMs. With the capacity of the long–short-term memory for capturing long-term dependencies, the Bi-directional gated recurrent unit is accused with the ability to store the sequence of information in both directions such as previous and upcoming. Numerous achievements have been effective in regard to the sequence of information related to both directions in many regions [45]. Bi-directional recurrent unit process the sequence not in the same direction which is then concatenated as the output for instance hidden state generated as:

$$[\overrightarrow{hs_1} \parallel \overleftarrow{hs_1}, \overrightarrow{hs_2} \parallel \overleftarrow{hs_2}, \ldots, \overrightarrow{hs_n} \parallel \overleftarrow{hs_n}] \tag{15}$$

where arrows demonstrate the directions of processing.

For the representation of the word vector of the specific tweet, the input sequence of final while working not in the same direction. The bi-directional Gated recurrent model is as per the following:

$$T = \overrightarrow{hs_n} \parallel \overleftarrow{hs_1} \tag{16}$$

3.7. Back-propagation through time

In order to train the recurrent neural networks, we employed backpropagation algorithm's expansion which is Backpropagation through time in which error is dealt through repetitive connection back specifically particularly when we need to work with time steps utilizing chain rule and error is additionally back-propagated by [46]. In this manner, backpropagation through time will be helpful for recalling data which reside in hidden layer for many time steps that will be at last equipped for training deep model computationally controllable in the respect of obtaining the gradients by [47] However, One Critical Issue related to this is the necessity of an expansive memory utilization to ingest and recalls data for the various time steps in the hidden layer which causes computationally an inefficient algorithm for backpropagation of errors on RNNs because of settled memory. This triggers the algorithm to re-evaluate the state by forward-propagating inputs beginning from the earliest starting point till time t when it *(t)* has to be reinstated since backpropagation occurs in the reverse time-based direction, in resulting the past forward steps cannot be reprocessed since the shortage of memory. Additionally, it requiring *(t)* forward steps to repeat before back-propagating gradients one-step reverse which at last requiring $t(t+1) = 2$ forward passes to back-propagate errors with time *(t)* steps, the resultant complexity is $O(t2)$ in time and $O(1)$ in space which can be lessened by storing only hidden states of time points of RNNs for the errors which need back-propagating from time t to $t-1$, with the usage of internal state of RNN re-evaluated by executing forward operation by taking the contribution of the recently hidden state while the backward operation can pursue instantly as described by [48]. More insights about the usage portrayed can be found in [49].

Recurrent neural networks with repeating modules consistently made out of the simple architecture. In any case, on account of Long–short term memory is not as simple as RNNs because it comprises various layers which are interacting in a particular manner with two states, namely cell state and hidden state, however, detailed description in the above sections. In our work, tweet in the form of the token will be treated as input to an underlying layer of RNN Layer BI-LSTM as word embedding. It has the capability to capture the information about the (initial) current token yet, in addition, the previous and upcoming tokens. After the output produced by this initial layer, for the extraction of local features will act as an input to the convolutional layer. Furthermore, pooled to global average pooling layer resultant the sentiment characterization of tweets as positive or negative with softmax output.

The long-term dependencies that exist in the sentence specifically in our work in the tweets, bi-directional gated recurrent (Bi-GRU) unit attempt to process whole information which is present in the form of tweets having two hidden state vectors which are fixed. Bi-directional gated recurrent unit with all the hidden states to the arrangement of vector representation, first the outputs of this unit given to convolutional neural network and a filter is utilized to link hidden states which are in a consecutive way. By global average-pooling layer, an activity which capitalizes the sentiment features. The features created by these filters with different sliding window sizes are then linked to the vector representation of the tweet. To put it plainly, the output of the convolutional neural network based on the output produced by the recurrent gated unit will then give to output layer for the classification of tweets correspondingly.

3.8. Convolutional neural network

At first, intended for image recognition, the convolutional neural network was designed which is clarified in the above sections. While in the similar case on account of deep neural network an enhanced architecture composed with 9 – layers of convolutional neural network for the identification abnormal breast (Analyze Macroscale Image) proposed by [50] on small-scale MIAS dataset using the learning of cost-sensitive for balancing while to deal with extensive training set by data augmentation a deepen comparison in three phases first in respect of numerous layers, furthermore in the correlation between the enactment capacities as rectified linear unit (ReLU) with its variants for instance, leaky ReLU and parametric ReLU additionally empowered the comparison also with pooling techniques such as, average, max, stochastic including these pooling operations based on rank, empirically found that combination of parametric ReLU with the usage of rank-based stochastic pooling operations admirably on all previously mentioned methods henceforth to show signs of improvement execution of CNN, enhances towards the initiation capacities as leaky rectified linear the supplanting of common pooling with rank-based stochastic pooling.

However, it has turned into a versatile model for the wide range of NLP tasks achieving interesting outcomes [26,27]. CNNs comprises of various layers in the terms of convolutions which are associated along the actuation capacities as Rectified Linear Unit or tanh, that are connected to the results. In a conventional Networks dependent on Neural, the commitment of every neuron associated with each yield in the accompanying layer which is known as an affine layer. Conversely, today's convolutional neural network works with various methodologies in which utilization of convolutions over the input layer is utilized to figure the outcome with neighbourhood associations, furthermore each layer with many numerous filters along with applied different kernel for the purpose to combination the overall output. While in the case of pooling operations and training, neural networks specifically CNNs filter size rely on the task for which the values to learn. In NLP undertakings, matrix of full words on which a filter normally slides it implies there is a closeness between the widths of both the input matrix and with a respective filter yet area size may not be the equivalent, despite the fact that it comprises of two to five words in a sliding window at some random time. Convolutional neural networks are associated with the assumptions related to independence between the concerned inputs and their respective outputs. Essentially, word embedding given to a convolutional layer which will then create different filters resulting will learn multiple features then applied these features sequentially to the different divisions of input. Afterwards, the output is normally pooled to smaller dimensions and later fed into a coupled layer as explained in [10]. The typical structure of the convolutional neural network explained well in [51] is beneath in Fig. 5.

3.8.1. Convolutional layer

In this layer for the extraction of local features, multiple features are utilized. The following equation will demonstrate how a filter F_t learns feature Map m_i^t as below:

$$m_i^t = f(V_{i;i+w-1} \odot W^t + b^t) \tag{17}$$

where *W represents the matrix weight*, $V_{i;i+w-1}$ are token vectors, b for bias, while the convolutional operation is $\odot$. Furthermore, in a sentence, for the extraction of local features identified with each window use of same weight matrices. However, to get sufficient and various features, additionally filters with multiple lengths along with different weight matrices are utilized.

3.8.2. Max-pooling layer, dropout and output layer

Max pooling layer performs two significant tasks, first discard the values that are non-maximal and pick a maximal value, this is the reason with a specific end goal to deal with the characteristics

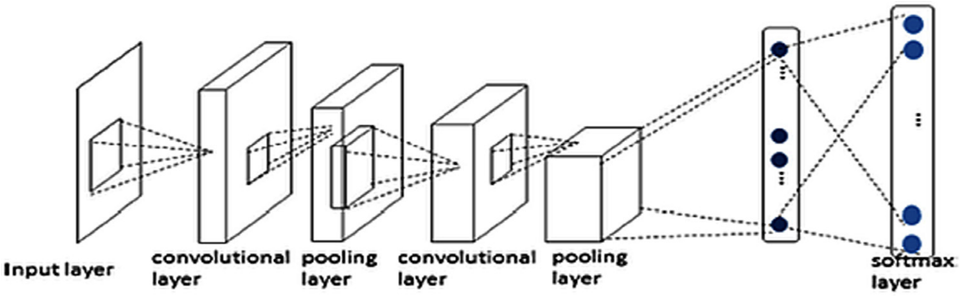

Fig. 5. CNN distinctive architecture as depicted in [51].

of the tweet. There is one certainty which is about the convolutional neural network is over-fitting so a concept of dropout layer regularization is essential which is subsequently produce a dense connection in both max-pooling and convolutional layer proposed in [52]. For the ampleness of the convolutional neural network a novel model in remote sensing image segmentation in the respect of polarimetric synthetic aperture radar [53] comprising of multiple layers; 11 a deep convolutional neural network Furthermore, comparing the pooling operations such as average and Max, broadly exhibits to superior to Alex Net on account of validation. Later on another model with the correlation between different techniques such as average, max and stochastic pooling on pooling layer for the detection of alcoholism utilizing data augmentation based on conventional convolutional neural network led with consequence of five convolution layers and two fully connected layers along with stochastic pooling is better than individual pooling consist of max and average as well in light of the fact that such pooling picks the pooled map reaction by selection after a multinomial dissemination moulded as of the initiations of particular pooling area [54].

However, with a definite objective identified in this paper is to interrogate the computational complexity whether it less or not, a global average pooling (GAP) layer while ignore the pad tokens in the case of implementing this operation identified by replacing traditional max-pooling which is applied to the output created by a convolutional layer which then additionally replaces the fully connected layer and dropout due to itself a regularizer further, it specifically coordinates between the mapping of features and classifications. The one issue with the fully connected layer that hampers the speculation capacity of the overall network is over-fitting on the other hand dropout by [55] inclined to haphazardly set portion of the initiations to the fully connected layers to zero during training when performed as a regularizer, so Global Average pooling is persuaded to manage the speculation capacity and to a great extent forestalls over-fitting [56].

In addition, in the global average pooling to stay away from the above-mentioned problem does not need any optimization of parameters. In the same case, we concentrated on the average value related to each feature map, rather than including fully connected layers of the specific feature map. The average value of each feature map generates a particular value; confidence for every class and afterwards, the subsequent vector is then inputted straightforwardly into the softmax function to get the likelihood distribution. The accompanying expression depicts the probability appropriation over analysis of sentiment.

$$P(i_j \,|k, \theta\,) = \frac{\exp\left(x_j(k,\theta)\right)}{\sum_{1\le i\le |X|} \exp\left(x_i(k,\theta)\right)} 1 \le j \le |X| \tag{18}$$

where, $x_j(k, \theta)$ as average pooling result with a parameter set θ corresponding to the class j, and class space is represented as X. For the minimization of negative log-probability, a stochastic gradient descent is exploited.

3.8.3. *Computationally convolutional neural network's cost*

As of late numerous CNNs involved on expending additional time when contrasted with as in the recent literature aforementioned in [56] as far as training and testing. The explanation for the broad cost of computational can be tended to with the system by taking into contemplations of number factors, for example, the depth [57,58] and width [59,60] of the network which is alluded with the extent of layers and filters and the amalgamation of these two factors in like manner. Then again the computational power can be handle halfway with hardware prompts additional expense, as far concern memory issue can likewise be settled by utilizing the littler size of mini-batch. In our work the architectural plan, the time cost of the fully connected layer is not involved resultant enhancement in 5%–8% in the computational time, in any case, with convolution layer, the computational cost in term of time can be settled by the accompanying articulation:

$$o\left(\sum_{I_c}^{d_c} w_{I_c-1}.f^2_{w_{I_c}}.w_{I_c}.l^2_{w_{I_c}}\right) \tag{19}$$

where, convolutional layer index and the number is depicted by I_c, d_c, while w_{I_c}, w_{I_c-1} is width and input channels with corresponding indexed layer and $f_{w_{I_c}}$ filter length lastly $l_{w_{I_c}}$ is a feature map for output.

However, computational cost with time complexity in the aforementioned expression instead of the actual running time relies upon the presented architecture, due to its affectability which is depended on the elements of cost of implementations and particularly with the hardware by [61].

However, all the works discussed in this section is on consuming many layers thus far our work inclines to perform of single convolution layer taking the global average pooling operation in ignorance with the padding, although it is considered as deep neural network due to the combination of two neural networks; RNN and CNN but comparatively with less overhead and in computation.

4. Twitter data set description

Tweets normally composed of scarce meaning, incomplete expressions, ineffectively organized sentences due to the presence of frequently used acronym, irregular grammar, and words that are not present in the lexicon that will clearly affect the performance of sentiment classification as explained in [62,63]. In our case, the tweets containing the most extreme length in the dataset is considered as the settled length for frameworks with the craving to unify the tweets of multiple lengths while padding of zero on account of the shorter tweet. However, for the optimal evaluation of our situation we choose to pick the different combination of words identified with negations, "#", POS, Emoji and user mention "@" for the features. A series of pre-processing is required on tweets so as to reduce redundancy, anomalies in the miniaturized scale blog content.

4.1. Preprocessing of tweets

Even though word embedding is very popular to work with the textual frame of information, there are very generic techniques that do not assume anything from the given text. Nonetheless, Twitter is not quite the same as other text structures like it contains Mentions (@user) and Hashtags (#topic) that can have an imperative sentiment in just 140 characters in length, which led numerous users to utilize shortenings in a very casual manner. There are various pre-processing techniques identified with tweets some of them we use with the similar prescribed order presented in Table 1, which empowers confrontational possessions clarified in [64] while the selected datasets along polarity as well are explained in Table 2.

Handling Non-English words and unicode

There must be a standard form of the tweets so, to clean the dataset at the initial step we assured that the tweets that are not in English and Unicode like "\u018e" be evacuated which were caused miscellanies in the process of crawling.

Handling user mentions and URLs

In general, normally people tweets by mentioning a user with the hashtag or/and @ and numerous tweets are seen with the URL. In our work, we removed the URLs, hashtags and @ however, we assume with the concerned of this paper, tweets with # and @ have significance related to sentiment. We used a dictionary with the most frequently used English words and tried to guess the correct split by using the Viterbi algorithm.

Abbreviation replacement

People used to compose ordinarily casual and self-made short words via web-based networking media, particularly on Twitter where the tweet length is restricted because of this reason tweets are complex to understand. In order to interpret the right implications, it is, therefore, necessary to expand the short informal words. For Instance, tc, brb and so forth instead of "take care" and "be right back".

Slang word replacement

These are the words which take rapidly active part in a particular context or in a specific network in social media correspondence. Due to different structures, classification algorithms often cannot recognize them. We created a compressive list of slang words and their formal replacements, for example, "F2F" for Face to Face and "SP" for Sponsored. We utilized the acronym dictionary Internet Slang Dict presented in [65].

Removal & replacement of numbers

Numbers do not contain any sentiment so we need to remove and replace the numbers in order to refine our dataset, For example, exceln8 to "Excellent" and 2good to "too good" etc. This is the inspiration driving why this pre-processing technique inspected after slang word replacement.

Emoji & emoticons replacement

Likewise, short and informal words, people used to express their attention by using emoji, that could contain some polarize information for sentiment analysis. We use a regular expression to find the emoji and replace them with positive and negative words respectively by using an emoticon dictionary [66].

Contraction replacement

It is obvious to use common contractions in English by users on social media like I'm, We've, Weren't, isn't and so forth, so there is an emergency to replace the contractions resultant prescribed contractions as "I am", "we have", "were not" and "is not" respectively. Moreover, if it does not in this way, each token for "I" and "m" created for I'm and So on. We handle this by utilizing Tweet-NLP [67].

Reappearances handling/replacement

Users use lots of punctuation to express their feelings, interest in social media, so most likely to contain sentiment information, For Instance, we replace '" with continuing and !!! with excitement and ??? with urgency. This replacement is quite necessary before to deal with punctuation

Handling negations with antonyms replacement

This pre-processing technique identified with the unusual approach in which we look up in every tweet for the negation words and thereafter change it into its antonym. For Instance, "not good enough" will be replaced with the word "Satisfactory". In Addition, we use wordnet for this task.

Dealing with punctuation

Due to a limitation in tweets length, the importance of punctuation marks, plays a quite interesting role in the sentiment analysis. For the most part, an exclamation mark represents something to focus while inverted commas depict something new or important, so a cautious thought ought to be taken keeping in mind the end goal to acknowledge while removing the punctuation this can prompt to the accuracy of the tweets characterization.

Remove stopwords

There is a chance of Function words in a tweet with a high recurrence which does not derive any sentiment and subsequently is not valuable features. By utilizing the natural language toolkit (NLTK) stopwords list for removal of these sorts of words.

Handling lowercasing

People use to express their feelings in the way they want, for example, usage of capital words which sometimes infer the importance or intensity related to the topic, like OHH, HA HA etc. Moreover, we detect these words up to the certain limit of characters because to handle the lower case first we need to pre-process the capitalized words, then lowercasing the words which are merged being the same words eventually diminish the dimensionality.

Prolonged words replacement

These words identified with the casual composition style which is utilized as a part of web-based life these days such us okaaaaaaay, byeeee, muccch etc. To decrease these words with their unique words with the end goal of not treating them diverse words which will enhance the quality of the dataset, however, these types of words are typically dismissed due to the low occurrence.

Spelling correction

Today, every user, tweets by using cell phones, it is possible to type the words with wrong or incorrectly spelt, we utilized "Typo Corpus" which lists the most common typo to recover the words.

Part of speech tagging

It clarifies how the word used in a sentence and encourages us to understand the structure of the sentence, ordinarily tweets are in informal grammar, so the explanation of this is to disregard the parts in the tweets that do not contain any sentiment, however, in our case we just engaged verbs, adverbs and noun.

Lemmatization and stemming

Lemmatization is the process of grouping together words which are in modified form to analyze as a single item that is recognized by lemma whereas stemming, is the process of removing the endings of the words to perceive their root. By doing this, comprehend numerous words that are merged and their dimensionality is additionally lessened.

5. Experimentation

Keeping in mind the end goal to precisely assess our proposed techniques, we test the system on the accompanying datasets.

Table 2
Twitter's pre-processing techniques in the prescribed manner [64].

Sequence	Techniques
1	Handling non-English words and unicode
2	Handling user mentions and URLs
3	Abbreviation replacement
4	Slang word replacement
5	Removal and replacement of numbers
6	Emoji translation
7	Contraction replacement
8	Reappearance handling/replacement
9	Handling negations and antonyms replacement
10	Dealing with punctuations
11	Stopwords replacement
12	Handling lowercasing
13	Orthographic words replacements
14	Spelling correction
15	Parts of speech tagging
16	Lemmatization and stemming

Table 3
Twitter's dataset along Tweet's Polarity.

Datasets	Total tweets	Positive tweets	Negative tweets
STSC-4K	4000	2000	2000
STSC-1.5K	1500	750	750
SS-TDS	4242	1252	1037
HCR-TDS	839	163	536

5.1. Selected datasets

(1) Stanford Twitter Sentiment Corpus[2] with 1.6 Million tweets taken from [68] set apart as positive and negative. To procure two test datasets, unified modelling is utilized, STSC-4K and STSC-1.5K having 4000 and 1500 tweets individually. Although the dataset is not substantial, despite this has been pragmatic in numerous fields illuminated in [3,30,61,62,69,70].

(2) Sentiment strength twitter test[3] dataset developed by [71] contains the aggregate number of tweets is 4242 out of which 1252 are positive and 1037 are negative, which we pick, whatever is left of the tweets are named as neutral or irrelevant as done in [72].

(3) In March 2010, this dataset containing tweets with "#hcr" constructed by crawling as explained in [73]. This dataset contains a third type of label that is neutral other than Positive and Negative by manual annotation. However, in our paper, we utilized 163 positives and 536 negatives labelled out of total 839 tweets in (HCR) Health Care reform test dataset,[4] despite the fact associated with the time of its crawling it is being into many researcher's recent consideration in [74,75], as shown in Table 3 with respective polarities.

5.2. Baseline

In this section, only for reference identified with the recurrent neural network and the convolutional neural network evaluation, we at first depict baseline models, which are long–short-term memory (Bi-LSTM), gated recurrent unit (GRU), Bi-directional GRU and convolutional neural network model tested with random initialization and also trained with GloVe respectively. By then we show how we expand these baseline models, by joining RNN layer on its variants with CNN (refers to Bi-LSTM-CNN, GRU-CNN, Bi-GRU-CNN) exploring different avenues regarding random initialization and also trained with GloVe each model individually as shown in Tables 4–7.

Table 4
Baseline methods with random initialization.

Baseline	Narrative
Rand-Bi-LSTM Rand-GRU Rand-Bi-GRU Rand-CNN	Methods with random initialization

Table 5
Baseline methods using pre-trained word vector GloVe.

Baseline	Narrative
GloVe- Bi-LSTM GloVe-GRU GloVe-Bi-GRU GloVe-CNN	Methods using pre-trained vectors from GloVe

Table 6
Proposed methods with random initialization.

Proposed	Narrative
Rand- Bi-LSTM -CNN Rand-GRU-CNN Rand-Bi-GRU-CNN	Methods with random initialization, Bi-LSTM, GRU, Bi-GRU with CNN

Table 7
Proposed methods using pre-trained word vector GloVe.

Proposed	Narrative
GloVe- Bi-LSTM -CNN GloVe-GRU-CNN GloVe-Bi-GRU-CNN	Methods using pre-trained vectors from GloVe, Bi-LSTM, GRU, Bi-GRU with CNN

5.3. Experimental setup

When we built up our model, we have to dissect our selected datasets into word embedding and feature vectors that will be continued to test different situations to locate the best parameter. Our early experimentation entailed in assessing the best parameter incorporates the embedding dimension, epoch, batch size, filter windows and learning rate.

In a Neural Network, we can train with the mini batches (also called batches) of multiple training cases at once, which permit to diminish the substantial dataset into smaller portions. Sometimes accuracy and number of layers are straightforwardly relative which implies with the expansion of the number of layers tends to expand the precision, much of the time, with the expansion of the number layers, accuracy descent drastically which referred as "Vanishing Gradient Problem" [38]. Surely the computational execution tending to diminish with an expansion in the extent of the single hidden layer, it does not have any impact on the accuracy of the neural network model. In case if we have too little-hidden layer will be a cause of damaging the performance when feature expected to be as input. In the respect of our structure, one layer with respect of convolution layers is utilized also a field size of (3,3,5) while the dimensions related to hidden state is considered to be 128. Over each batch, global average pooling is connected, for the purpose of keeping the average output of vectors concerning each word (time step) Furthermore, the output identified with previous operations are additionally concatenated.

In our Experiments, same pre-processing steps performed to the selected datasets and for the best appraisal of our model and in the wake of testing various mixes of windows filter, seven is

[2] http://help.sentiment140.com/for-students.

[3] http://sentistrength.wlv.ac.uk.

[4] http://bitbucket.org/speriosu/updown.

Table 8
Baseline methods accuracy with random initialization on datasets.

Methods	Accuracy (%)				
	STSC-4K	STSC-1.5K	SS-TDS	HCR-TDS	Average
Rand- Bi-LSTM	72.93	74.81	81.95	76.40	76.52
Rand-GRU	74.26	75.69	71.97	81.11	75.75
Rand-Bi-GRU	79.81	75.11	80.69	77.24	78.21
Rand-CNN	66.71	65.30	71.07	69.87	68.23

Table 9
Baseline methods accuracy using pre-trained word vector GloVe on datasets.

Methods	Accuracy (%)				
	STSC-4K	STSC-1.5K	SS-TDS	HCR-TDS	Average
GloVe- Bi-LSTM	78.21	74.79	80.24	81.63	78.71
GloVe-GRU	80.06	77.10	79.21	81.39	79.44
GloVe-Bi-GRU	79.10	81.97	81.28	80.11	80.61
GloVe-CNN	70.84	72.42	71.69	69.35	71.07

Table 10
Comparison based on proposed method's accuracy using random initialization and using pre-trained word vector GloVe with the joint architecture of recurrent neural network and convolutional neural network on datasets.

Methods	Accuracy (%)				
	STSC-4K	STSC-1.5K	SS-TDS	HCR-TDS	Average
Rand- Bi-LSTM -CNN	80.41	82.14	85.52	79.91	81.99
Rand-GRU-CNN	84.67	85.60	82.61	80.11	83.24
Rand-Bi-GRU-CNN	85.29	86.63	83.46	81.55	84.23
GloVe- Bi-LSTM-CNN	85.92	87.60	82.79	88.72	86.25
GloVe-GRU-CNN	86.51	83.55	89.46	84.91	86.10
GloVe-Bi-GRU-CNN	90.59	87.18	85.59	86.98	87.58
GloVe-DCNN [76]	87.62	–	81.36	–	–
SCNN-CHAR [13]	86.4	–	–	–	–
RNN [9]	82.4	–	–	–	–

observed to be the ideal decision. Moreover, a batch size as 256, epochs remain between 10 to 50 for all the datasets and the learning rate to 0.01, however, for the baseline models dropout was set at 0.5 to regularize them. While going on a series of experiments, it is likewise seen that in the convolutional layer for the better performance the activated function hyperbolic ReLu is found to be the best under our nature of work.

5.4. Empirical result and analysis

The assessment metric of our experimental work for the purpose of sentiment classification is the accuracy as appeared in Table 8 which incorporates the accuracy of our datasets. We tested a progression of work in view of baseline methods which comprise each model with random initialization and also with pre-trained word vector for our reference in order to make a competitive comparison between them. However, we did our best to the proposed model that consists of the joint framework of baseline methods also tested with random initialization and pre-trained word vector using GloVe with the intent to compare the best accuracy.

The average accuracy of GloVe-Bi-GRU-CNN is 87.58% as compared with Rand-Bi-GRU-CNN, which is 84.23%, while the other method's accuracy by using GloVe-GRU-CNN, GloVe- Bi-LSTM-CNN is at 86.10% and 86.25% as compared with Rand-GRU-CNN, Rand- Bi-LSTM-CNN which are 83.24% and 81.99% respectively as shown in Fig. 6 and in Table 10. With the end goal of accuracy in view of the individual dataset, GloVe-Bi-GRU-CNN accomplishes the most elevated change in the precision is 90.59% on a dataset STSC-4K, while GloVe-GRU-CNN and GloVe- Bi-LSTM-CNN with 89.46% on SS-TDS and 88.72% on HCR-TDS as shown in Fig. 7. While in the baseline method for the best accuracy is found in 81.97% by GloVe-BiGRU on STSC-1.5K as compared with Rand-Bi-GRU which is 80.69% on SS-TDS can be seen in Tables 8 and 9 respectively also depicted in Fig. 8.

6. Discussion

From the above results, we are presently ready to assess the models that are set up by utilizing GloVe effectively to accomplish the preferred execution over the models that are not pre-trained for the sentiment task on selected corpora. Further, pre-training utilizing GloVe learning on the dataset, particularly when managing with large corpus related to tweets comparatively effective for finding the words which are of the same sort which is then used for the semantic and the information in regards to the sentiment that dwell in the middle of words of the tweets Likewise, the issue of Semantic sparsity with expansive corpora can be unravelled to a specific degree by utilizing pre-trained vectors GloVe. In this paper, our exploratory work performs precisely when our architecture gets a sentence level feature representation using pre-trained GloVe, RNN Layer being (Bi-LSTM, GRU, Bi-GRU) and CNNs along with GAP by maintaining error propagation extracts local features of the input given by the RNN layer so as to choose features from the tweets which have the sentiments for enhancing the performance of the classification. Furthermore, the assessment grids, such as accuracy or precision of the model in the respect of deep learning not simply relying on classifiers but rather several elements like vanishing the gradients, feature extractors, especially dataset and its size because still, it is an open research to locate the model that fits on all sort of datasets.

Moreover, numerous procedures for the development of the representation of sentences vigorously depended upon tree structure using recursive neural networks as explained in relevant literature [9,25] shows time complexity of $O(n^2)$; n denominates as text length, which is tedious on account of long sentences, so these systems are not reasonable. In Contrast, a model on text based on words using recurrent neural Network with a $O(n)$ time complexity in a fixed size hidden layer by [77] of all previous text, which is advantageous for the contextual information henceforth towards long sentences yet biased to concentrate later words rather to prior words that can at last trade off on viability.

In the case of numerous NLP undertakings, to handle previously mentioned issue related to biased and with time complexity convolutional neural networks with max pooling layer, turns out to be successful to oversee discriminative in contextual as well as incarceration the semantic information as compared to recurrent and recursive neural network yet needs to use modest convolutional kernels which is vital to the span of window that may take the model to relinquishment of some basic data when window is little, though huge parameter space as the window measure which is huge to train as introduced by [6,8].

Intentionally, to manage the computational cost in term of time complexity in the aforementioned studies problem in a manner by relating RNN variants such as Bi-LSTM, GRU, Bi-GRU triggered activated with less noise accordingly for contextual information capturing and hold a larger scope of the word ordering in noteworthy degree when learning with GloVe word representation Secondly, we employ convolutional neural network in a global average-pooling operation which unavoidably chooses the features with a key role for the classification of text to internment the text which is generally significant. By combining two neural networks, utilizes the advantage of both neural models; recurrent and convolutional and exhibits a time complexity $O(n)$ since the overall model is a cascade of the two neural networks, in this manner, the time complexity of our model is as examined which is straightly connected with the degree of the text dimension as compared with [9] where the computational cost is $O(n^2)$ and reported training is 2 to 5 h though for the situation is going to a few minutes by utilizing single string machine.

The Hybrid neural model with the regard of this paper, a few specialists endeavour to consolidate the upsides of CNN and

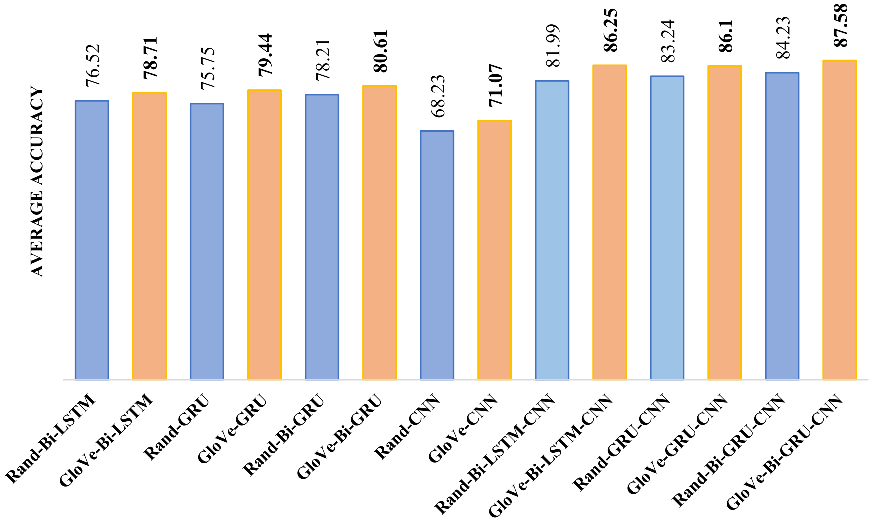

Fig. 6. Comparative analysis of average accuracy using random initialization and Glove on the baseline and proposed methods.

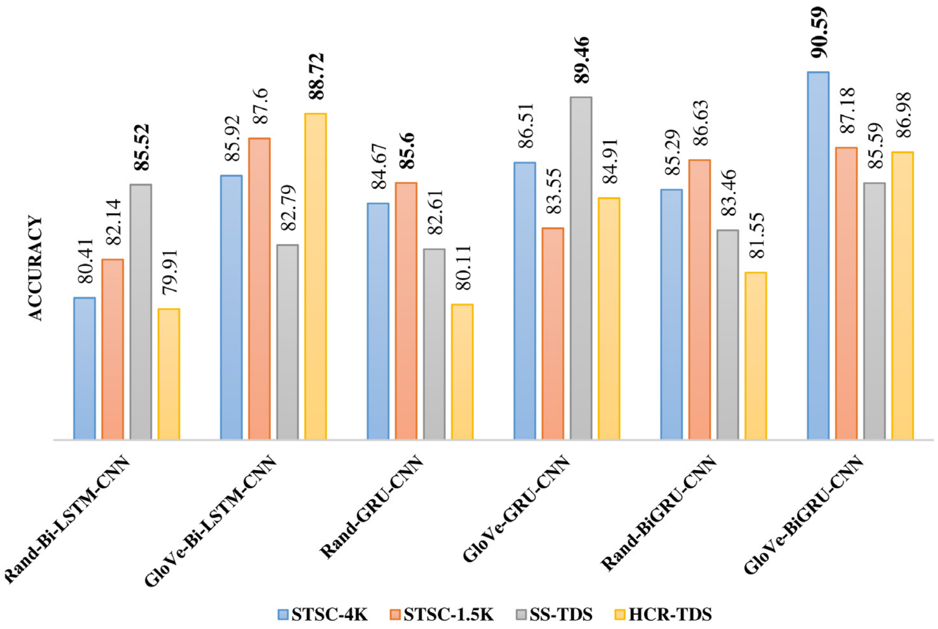

Fig. 7. Proposed Method's accuracy on individual data sets.

RNN [27] extricate local and global features by utilizing described neural networks independently. At firstly modelling of sentences by RNN [78], and after that utilization CNN to get the representation of sentences while [79] supplant convolution filters with profound RNN variants; LSTM. The fundamental contrasts in our work with them contained in the way like, they see their models as CNN and set a little window size of 3, while we propose to utilize an alternate window estimate as little window estimate influences the model to lose the capacity to capture long-term dependencies Furthermore, we use global average pooling, however, not mean pooling, in light of the fact that worldwide this pooling is better in position-invariance maintaining [80]. Strangely, combining RNN and CNN with this particular pooling into a solitary system likewise seemed to enhance precision. In spite of the fact that the contention was made that the CNN models accomplish the best precision for their number of parameters, the joined structures outflanked every single other model on unadulterated exactness scores. Further, we can observe our model execution if there should arise an occurrence of expansive datasets on the accompanying situations particularly with this work as the word embeddings are pre-prepared on a lot bigger unannotated corpora [29] to accomplish better speculation given the small amount of training data since these are fine-tuned during training to enhance the performance of classification by breaking down its semantic properties as it identifies with larger dimensions can give a predominant execution, utilizing it as a feature and to initialize neural networks in various perspectives like corpus domain which could really compare to corpus size after that, training on a huge corpus, for the most part, enhances the nature of word embeddings, and training on an in-domain corpus can fundamentally enhance the nature of word embeddings for a particular errand whereas faster models give an adequate performance much of the time, on the other hand on the bases training corpus which is of large extent more unpredictable models can be utilized, while the early stopping metric for repeating ought to depend on the enlargement set of the anticipated undertaking as divergent to the validation loss related to training embedding. In neural language models, training on large corpora depicts the training time, rather

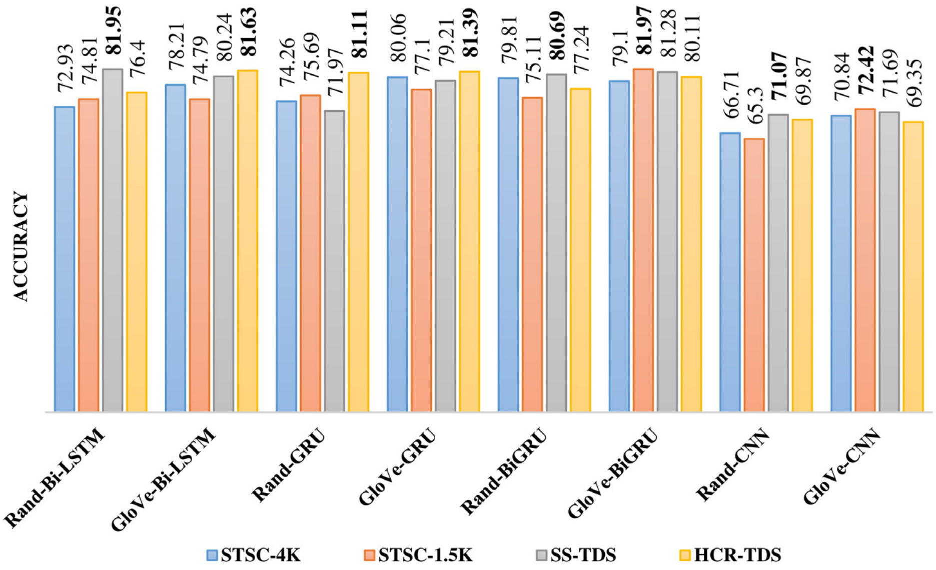

Fig. 8. Baseline method's accuracy on individual data set.

than the training data which is the principal factor that is for the constraining the performance. Training time in this way to be considered while expanding the limit so large training sets and a fixed training time introduce competition between slower models with greater limit and observing more training data. Scaling the training to expansive datasets can significantly affect perplexity, notwithstanding when data from the appropriation of intrigue is small. Our target here is to choose the association as far as neural networks size in term of training time, hidden layers, region size, filter selection, word error rate, and last but not least, the size of the training set when there are large datasets. In our case, the network design is relatively modest as a single layer of RNN is enough to hold long-term dependencies especially with the regard of this paper its variants Bi-LSTM, GRU, Bi-GRU along with the CNNs with one convolutional layer using global average pooling replacing the fully connected layer and dropout loads on network, up to this point, synchronous increments to the size of the training set tend to improve the performance of the neural network; as such, more data helps, requires more parameters to be trained. We keep on being surprised even a simple architecture works fills in and additionally for the complex undertakings. Another part of research can likewise centre around the examination of large data by applying two neural networks due to the number of neurons and the number of hidden layers in the network that can also directly influence execution in light of the fact that few layers can process quicker than a major one. By and large, the number of the hidden layer can expand the exactness of learning. Be that as it may, it will influence the learning time considerably more than a little layer. Consequently, our proposed strategy enables the neural network to learn bigger informational collections the precision is practically identical to the network trained by the large data corpus while the training time is drastically diminished.

Our experiments support numerous key discoveries for the viable execution of neural networks in the event of little, medium and also substantial(Large) datasets. (i) Evaluation of Domain Specific Word Embeddings (ii) RNNs with CNNs utilizing Global Average pooling give corresponding data to content characterization errands. Which design performs better relies upon the fact that it is so vital to semantically comprehend the entire task. (iii) Learning rate changes execution generally easily, while changes to the hidden size and batch size result in extensive variances.

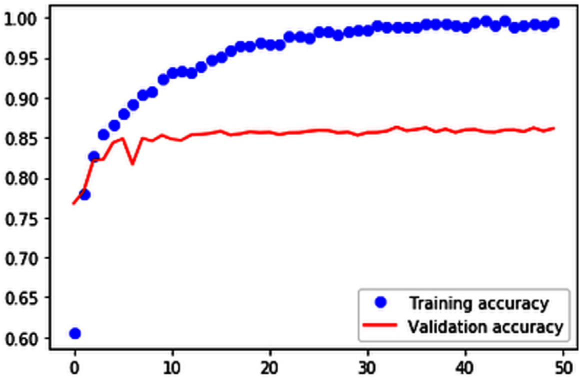

Fig. 9. Rand-Bi-LSTM-CNN on SS-TDS.

The performed outcomes in Table 10 obviously depicts comparison with the presently proposed methodologies based on neural networks that our proposed architecture is brought about in accumulating the effective accuracy which is either better or equivalent for the task identified with sentiment analysis in numerous perspectives, such as, joint architecture yet simple and prosperous by diminishing the dense connection created by convolutional layer in respect number of layers and parameters as far as, in terms of parameter tuning relying on the capabilities of recurrent neural networks. Moreover, there is a considerable measure of tunable hyper-parameters associated with in deep learning, tuning numerous hyper-parameters in the meantime can be exceptionally costly, particularly when the datasets are expansive. Then again, settling hyper-parameters over all datasets and models would not be a reasonable methodology in our examination in light of the fact that the datasets and model structures are altogether different and hence may require altogether different hyper-parameter settings. Following figures in this section exhibit the convergence of the accuracy in the part of comparing the curves representing to the exactness on a particular dataset which attains the maximum accuracy of both training and testing set in our deep neural network which prescribes that the model is not overfitting exorbitantly. Further, we can likewise find in Figs. 9–14 that our proposed network converges after about the recommended number of epochs, and its accuracy is then consistent on the test corpus.

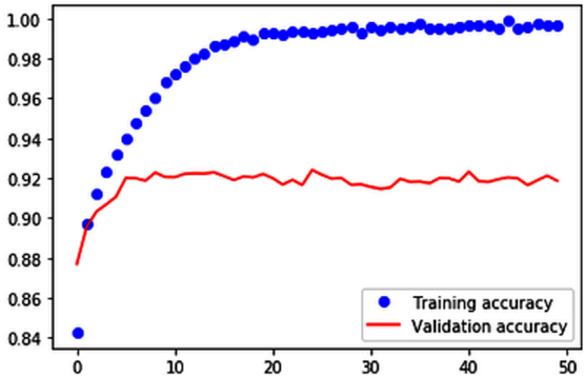

Fig. 10. GloVe-Bi-GRU-CNN on 4K.

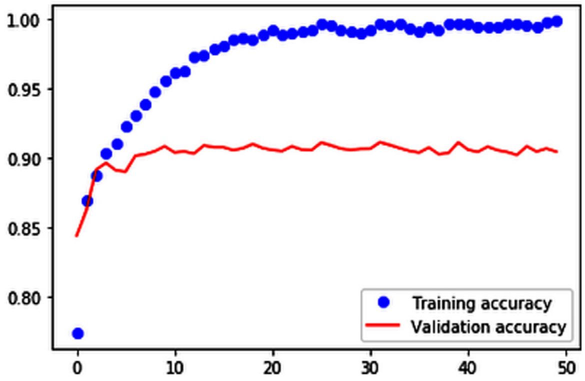

Fig. 11. GloVe-Bi-LSTM-CNN on HCR-TDS.

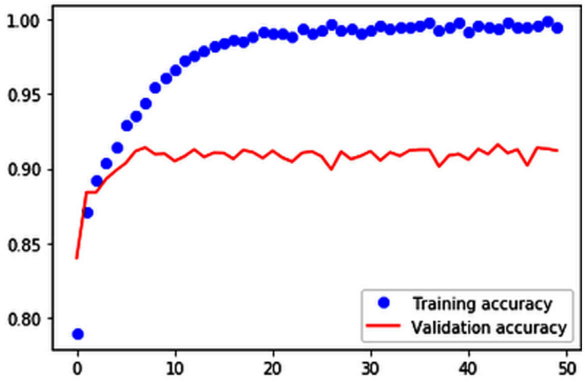

Fig. 12. GloVe-GRU-CNN on SS-TDS.

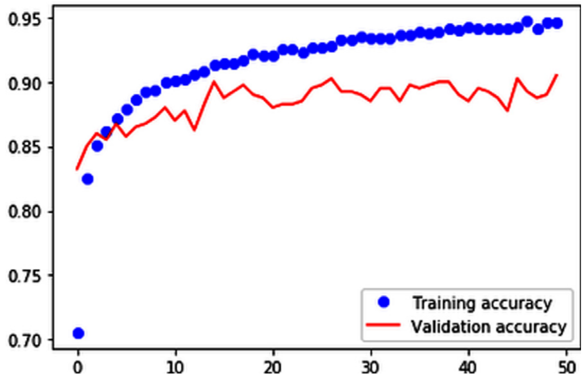

Fig. 13. Rand-Bi-GRU-CNN on 1.5K.

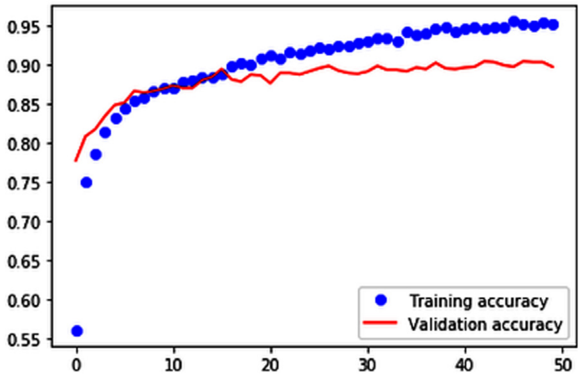

Fig. 14. Rand-GRU-CNN on 1.5K.

7. Conclusion

Although CNNs extracts higher level features, so as to capture long-term dependencies, it requires numerous layers. This will a cause of very deep network that contains the excessive number of convolutional layers. To overcome this problem, in this paper, we abridged our work with a new, simple and efficient framework that is a joint architecture between the RNNs variants which will capture long-term dependencies and reduces the loss of local information with CNNs utilizing global average pooling that replaces the fully connected layer. Briefly, we use pre-trained word vectors using GloVe to represent a feature vector of the tweet, then fed it to the recurrent neural network which learns the long-term dependencies, afterwards inputs to convolutional and global average pooling layer for the purpose of retaining the relational sequence and local features in the tweet to perform sentiment analysis.

Our methodology shows reliable classification accuracy on various benchmark datasets and illustrated that it is conceivable to utilize small architecture despite the fact of joint architecture for better classification accuracy. This empowers approaching researchers to further explore the proposed methods to the highest degrees in many areas such as machine translation, information retrieval and many other applications, particularly in natural language processing. Finally, we conclude that the architecture composed with RNN layer and CNN on the top of pre-trained word vectors in a mentioned manner reveals reliable outcomes as compared with random initialization for the classification of sentiments on Twitter.

References

[1] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: Machine Learning: ECML-98, in: Lecture Notes in Computer Science, 1998, pp. 137–142.

[2] Y. Mu, Y. Fan, L. Mao, S. Han, Event-related theta and alpha oscillations mediate empathy for pain, Brain Res. 1234 (2008) 128–136.

[3] S. Kiritchenko, X. Zhu, S.M. Mohammad, Sentiment analysis of short informal texts, J. Artificial Intelligence Res. 50 (October) (2014) 723–762.

[4] N.F.F. da Silva, E.R. Hruschka, Tweet sentiment analysis with classifier ensembles, Decis. Support Syst. 66 (2014) 170–179.

[5] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining sentiwordnet, Analysis 10 (2010) 1–12.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, andj.ean, Distributed Representations of Words and phrases and their Compositionality, 16-Oct-2013. [Online]. Available: http://arxiv.org/abs/1310.4546 [Accessed: 23.02.18].

[8] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, Proc. ACL (2014) 655–665.

[9] R. Socher, et al., Recursive deep models for semantic compositionality over a sentiment treebank, in: Proc. Conf. Empirical Methods Natural Lang. Process, Seattle, DC, USA, 2013, pp. 1631–1642.

[10] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[11] R. Johnson, T. Zhang, Effective Use of Word Order for Text Categorization with Convolutional Neural Networks, Vol. 2011, 2014.

[12] R. Johnson, T. Zhang, Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding, 2015, pp. 1–12.

[13] C.N. dos Santos, M. Gatti, Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, Coling-2014, 2014, pp. 69–78.

[14] L. Nio, K. Murakami, Japanese Sentiment Classification Using Bidirectional Long Short-Term Memory, Recurrent Neural Network, Vol. C, 2018, pp. 1119–1122.

[15] Q. Qian, M. Huang, J. Lei, X. Zhu, Linguistically regularized LSTMs for sentiment classification, in: Proc. 55th Annu. Meet. Assoc. for Comput. Linguist. Volume 1 Long Pap, 2016, pp. 1679–1689.

[16] C. Guggilla, T. Miller, I. Gurevych, CNN- andLSTM-based claim classification in online usercomments, in: Proc. COLING 2016, 26th Int. Conf.Comput. Linguist, 2016, pp. 2740–2751.

[17] Z. Zhao, H. Lu, D. Cai, X. He, Y. Zhuang, Microblog sentiment classification via recurrent random walk network learning, in: IJCAI Int. Jt. Conf. Artif. Intell, 2017, pp. 3532–3538.

[18] J. Wang, L.-C. Yu, K.R. Lai, X. Zhang, Dimensional sentiment analysis using a regional CNN-LSTM model, in: Proc. 54th Annu. Meet. Assoc. Comput. Linguist. Volume 2 Short Pap, 2016, pp. 225–230.

[19] Y. Xiao, K. Cho, Efficient character-level document Classification by combining convolution and recurrent layers, (Feb. 2016). [Online]. Available: https://arxiv.org/abs/1602.00367.

[20] L. Zhang, Y. Zhou, X. Duan, R. Chen, A Hierarchical multi-input and outpbelut Bi-GRU model for sentiment analysis on customer reviews, in: IOP Conf.Ser. Mater. Sci. Eng., Vol. 322, 2018, p. 062007.

[21] X. Pang, Y. Zhou, P. Wang, W. Lin, V. Chang, An innovative neural network approach for stock market prediction, J. Supercomput. (2018) 1–21.

[22] C. Karyotis, F. Doctor, R. Iqbal, A. James, V. Chang, A fuzzy computational model of emotion for cloud-based sentiment analysis, Inf. Sci. (Ny) 433–434 (2018) 1339–1351.

[23] Z. Jianqiang, G. Xiaolin, Deep convolution neural networks for twitter sentiment analysis, IEEE Access 6 (2018) 1–1.

[24] Y. Shen, X. He, J. Gao, L. Deng, G. Mesnil, Learning semantic representations using convolutional neural networks for web search, in: Proceedings of the 23rd International Conference on World Wide Web - WWW 14 Companion, 2014.

[25] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, in: 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.

[26] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Adv. Neural Inf. Process. Syst. (2015).

[27] Y. Xiao, K. Cho, Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers, 2016.

[28] Machine Translation by Jointly Learning to Align and Translate, 01-Sep-2014. [Online]. Available: http://arxiv.org/abs/1409.0473 [Accessed: 22.01.18].

[29] J. Pennington, R. Socher, D. Christopher, Manning. GloVe: Global Vectors for Word Representation. [Online], 2014. Available: https://nlp.stanford.edu./projects/glove/ https://nlp.stanford [Accessed:16.01.18].

[30] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining. strengths, Combining strengths emotions and polarities for boosting Twitter sentiment analysis, in: Proc. Second Int. Work. Issues Sentim. Discov. Opin. Min. - WISDOM '13, 2013, pp. 1–9.

[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. Adv. Neural Inf. Process. Syst, 2012, pp. 1097–1105.

[32] T.N. Sainath, et al., Deep convolutional neural networks for large-scale speech tasks, Neural Netw. 64 (2015) 39–48.

[33] T. Mikolov, G. Corrado, K. Chen, J. Dean, Efficient estimation of word representations in vector space, in: Proc. Int. Conf. Learn. Represent. (ICLR 2013), 2013, pp. 1–12.

[34] A.G. Ororbia, II, T. Mikolov, D. Reitter, Learning simpler language models with the differential state framework, Neural Comput. 29 (12) (2017) 3327–3352.

[35] K. Vo, D. Pham, M. Nguyen, T. Mai, T. Quan, Combination of domain knowledge and deep learning for sentiment analysis, in: Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell, 2017, pp. 162–173.

[36] Y. Kim, Y. Jernite, D. Sontag, A.M. and Rush, Character-Aware Neural Language Models, 2015.

[37] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol, 2011, pp. 142–150.

[38] D. Tang, F. Wei, B. Qin, T. Liu, M. Zhou, Coooolll: A deep learning system for Twitter sentiment classification, Semeval-2014, Vol. SemEval, 2014, pp. 208–212.

[39] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertainty Fuzziness Knowl.-Based Syst. 06 (02) (1998) 107–116.

[40] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, [Online]. Available: http://aclweb.org/anthology/D14-1162 [Accessed:16.01.18].

[41] B.Y. Lin, F.F. Xu, Z. Luo, K.Q. Zhu, Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media, in: Proc. 3rd Work. Noisy User-Generated Text, 2017, pp. 160–165.

[42] S. Misawa, M. Taniguchi, Y. Miura, T. Ohkuma, Character-based bidirectional LSTM-CRF with words and characters for japanese named entity recognition, in: Proc. 1st Work. Subword Character Lev. Model. NLP, 2017, pp. 97–102.

[43] K. Cho, et al., Learning Phrase Representations using RNN Encoder- A decoder for Statistical Machine Translation, 2014, pp. 1724–1734.

[44] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling 11-Dec-2014. [Online]. Available: http://arxiv.org/abs/1412.3555 [Accessed: 22.03.18].

[45] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proc. 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol, 2016, pp. 1480–1489.

[46] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Cognit. Model. 5 (3) (1988) 1.

[47] Paul J. Werbos, Backpropagation through time: what it does and how to do it, in: Proceedings of the IEEE, 1990.

[48] A. Gruslys, R. Munos, I. Danihelka, M. Lanctot, A. Graves, Memory-Efficient Backpropagation Through Time, 2016.

[49] M. Bodén, A guide to recurrent neural networks and backpropagation, Dallas Project, Dept. Comput. Sci. Univ. Skövde, Skövde, Sweden, Tech. Rep. 2002.

[50] Y. Zhang, C. Pan, X. Chen, F. Wang, Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling, J. Comput. Sci. 27 (2018) 57–68.

[51] Q. Shen, Z. Wang, Y. Sun, Sentiment analysis of movie reviews based on CNN-BLSTM, in: IFIP Advances in Information and Communication Technology, Vol. 510, 2017, pp. 164–171.

[52] R. Tobergte, S. Curtis, Improving neural networks with dropout, J. Chem. Inf. Model. (2013) 1689–1699.

[53] S. Wang, J. Sun, P. Phillips, G. Zhao, Polarimetric synthetic aperture radar image segmentation by convolutional neural network using graphical processing units, J. Real-Time Image Process. (2017).

[54] S. Wang, Y. Lv, Y. Sui, S. Liu, S. Wang, Alcoholism Detection by Data Augmentation and Convolutional Neural Network with Stochastic Pooling, 2018.

[55] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R. Salakhutdinov, Improving neural networks by preventing the co-adaptation of feature detectors. arxiv preprint arXiv:1207.0580, 2012.

[56] Alex Krizhevsky, Ilya Sutskever, Geoff Hinton, Image net classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, Vol. 25, 2012, pp. 1106–1114.

[57] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions. arXiv:1409.4842, 2014.

[59] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Over feat: Integrated Recognition, Localization and Detection using Convolutional Networks, 2014.

[60] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional neural networks, in: ECCV, 2014.

[61] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 5353–5360.

[62] Z. Jianqiang, Pre-processing boosting twitter sentiment analysis? in: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), 2015, pp. 748–753.

[63] Z. Jianqiang, G. Xiaolin, Comparison research on text pre-processing methods on twitter sentiment analysis, IEEE Access 5 (2017) 2870–2879.

[64] S. Symeonidis, D. Effrosynidis, A. Arampatzis, A comparative evaluation of pre-processing techniques and their interactions for Twitter sentiment analysis, Expert Syst. Appl. 110 (2018) 298–310.

[65] Internet & Text Slang Dictionary. [Online]. Available: https://www.noslang.com/dictionary (Accessed: 02.02.18).

[66] Wikipedia. List of Emoticons. [Online]. Available: https://en.wikipedia.org/wiki/Listofemoticons (Accessed: 02.02.18).

[67] Brendano GitHub.Com. [Online]. Available: https://github.com/brendano/ark-tweetnlp/tree/master/src/cmu/arktweetnlp (Accessed: 06.01.18).

[68] Stanford Twitter Sentiment Corpus. [Online]. Available: http://help.sentiment140.com/for-students (Accessed: 10.01.18.).

[69] H. Saif, Y. He, H. Alani, Semantic sentiment analysis of Twitter, in: Proc. Semantic Web-ISWC, 2012, 2012, pp. 508–524.

[70] A. Bakliwal, P. Arora, S. Madhappan, N. Kapre, M. Singh, V. Varma, Mining sentiments from tweets, in: Proc. 3rd Workshop Comput. Approaches Subjectivity Sentiment Anal. Assoc. Comput. Linguistics, Jeju, Korea, 2012, pp. 11–18.

[71] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol. 63 (1) (2012) 163–173.

[72] H. Saif, M. Fernandez, Y. He, H. Alani, Evaluation datasets for Twitter sentiment analysis a survey and a new dataset, in: the STS-Gold, CEUR Workshop Proc. 1096, 2013, pp. 9–21.

[73] M. Speriosu, N. Sudan, S. Upadhyay, J. Baldridge, Twitter polarity classification with label propagation over lexical links and the follower graph, in: Proc. Conf. Empir. Methods Nat. Lang. Process, 2011, pp. 53–56.

[74] X. Zou, J. Yang, J. Zhang, Microblog Sentiment Analysis Using Social and Topic Context, 2018, pp. 1–24.

[75] M.M. Fouad, T.F. Gharib, A.S. Mashat, Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble, no. February 2018.

[76] Z. Jianqiang, G. Xiaolin, Deep convolutional neural networks for twitter sentiment analysis, IEEE Access 6 (2018) http://dx.doi.org/10.1109/ACCESS.2017.27769, 1–1.

[77] J.L. Elman, Finding structure in time, Cognit. Sci. 14 (2) (1990) 179–211.

[78] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao, Recurrent convolutional neural networks for text classification, in: AAAI, Vol. 333, 2015, pp. 2267–2273.

[79] Yangyang. Shi, Kaisheng. Yao, Le. Tian, Daxin. Jiang, Deep lstm based feature mapping for query classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1501–1511.

[80] Dominik Scherer, Andreas Muller, Sven Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: International Conference on Artificial Neural Networks, Springer, 2010, pp. 92–101.

Fazeel Abid received his MS-Degree in Information Technology specialized with data Mining and Big Data correspondence. He is currently pursuing PhD degree and associated with lab of Intelligent processing and computing for big Data in Northwest University Xian, China. His Research interest includes social network and web mining in the aspect of deep learning.

Muhammad Alam holds a PhD degree in computer science from University of Aveiro, Portugal. He is currently working as an assistant professor in Xian Jiaotong-Liverpool University, Xian China. His research interests are Machine learning application, sensor networks, Internet of things, smart cities, smart homes and Mac protocol.

Muhammad Yasir is currently pursuing PhD degree with Northwest University Xian, China. His Research interest is currently including language identification and Data Mining.

Chen Li is working as a Professor in school of information science and technology in Northwest University Xian, China. Her main research areas are intelligent information processing, Data mining and network security. She is deputy director of National Discrete Mathematics committee. She is executive director of Shaanxi provincial computer society, executive director of Shaanxi signal processing society, senior member of computer society and deputy secretary general of IET Xian branch.