International Journal of Cognitive Informatics and Natural Intelligence
Volume 15 • Issue 4

Attention-Based Deep Learning Models for Detection of Fake News in Social Networks

S. P. Ramya, Periyar Maniammai Institute of Science and Technology, India*
https://orcid.org/0000-0003-0677-0826

R. Eswari, National Institute of Technology, Tiruchirappalli, India

ABSTRACT

Automatic fake news detection is a challenging problem in deception detection. While evaluating the performance of deep learning-based models, if all the models are giving higher accuracy on a test dataset, it will make it harder to validate the performance of the deep learning models under consideration. So, one will need a complex problem to validate the performance of a deep learning model. LIAR is one such complex labeled benchmark dataset which is publicly available for doing research on fake news detection to model statistical and machine learning approaches to combating fake news. In this work, a novel fake news detection system is implemented using deep neural network models such as CNN, LSTM, BiLSTM, and the performance of their attention mechanism is evaluated by analyzing their performance in terms of accuracy, precision, recall, and F1-score with training, validation, and test datasets of LIAR.

KEYWORDS

Attention Mechanism, Deep Learning, Fake News Detection, Natural Language Processing, Optimization

1. INTRODUCTION

The online social media websites make a significant role in various dimensions of today's life. Nearly three out of four persons participate in at least any one online social networks namely, Facebook, Twitter, YouTube and LinkedIn and so on. These types of online media provide large amount of information to their users. Such huge amount of information attracts the spammers and fake news producers to spread their vulnerable information to the genuine users.

Fake news has been around from the time news began to be printed and circulated largely after Johannes Gutenberg invented the printing press in 1439. However, there is no agreed definition for the term "fake news". Therefore, here certain broadly used definitions of fake news in the existing research are discussed and compared(Kai Shu et al., 2017).Fake news is a news article that is deliberately and provably false. A narrow definition of fake news is those news articles that are purposely and provably false as well as could deceive the readers. Broader definitions of fake news focus on either the authenticity or intent of the news content.

DOI: 10.4018/IJCINI.295809 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Kai Shu et al.(2017) uses a narrow definition of fake news. There are three reasons for narrowing down on this definition. First, the concealed meaning of fake news gives an in-depth understating of a new dimension to this topic. Second, techniques that are used for truth verification can be applied to both narrow conception of news in addition to a broader definition. Third, this explanation can remove the vagueness that lies between false news and relevant thoughts that are not explicitly (Kai Shu et al., 2017).There are few views which can be reviewed that the news is not false. They include (1) satire news in the right context with no intention to give a wrong impression to the readers (2) rumors that were not initiated from real time happenings (3) hypothesis which are difficult to prove as truth or falsehood (4) the deception that is made accidentally and (5) deceiving content that are only created by play or to cheat the person.

The fake news becomes a real and global problem in the world of widespread instantaneous information. The misinformation can cause confusion and unnecessary stress among the general public. It is one of the greatest threats to democracy and freedom. Fake news is called an infodemic as it spreads the false information purely to gain politically or financially. It disturbs the political path of democracies, spoils not only the individuals but also the society's reliability of upholding the laws and governance, motivating the public for sadism, bullying and inducing violence against innocent people as well as damaging the consumer's views of news media. Some of the real-world examples are as follows(Khan et al.,2020)

- Fake news influenced the voter's decisions in 2016 U.S. presidential elections
- Fake news led to mob lynching of two bikes in India in 2017
- Fake news caused riots in an Indian state whose capital is Kolkata in 2017
- Fake news headed to Dylann Roof to murder nine persons in 2017
- Fake news initiated a claim of Rs.500 crore (RM 286,087,500.00) from Kalyan Jewellers in 2018
- Fake news circulated that after December 31, 2019, the Rs 2,000 currency note would be banned in India in 2019
- Fake news directed that the government is offering internet free of cost (10GB per day) to students for supporting studies due to the spread of Corona virus in India in 2020

Now a days, fake news detection is a big challenge in online social web sites. So, the fake news has to be detected and a safe and fake-news-free environment must be provided to online users.

The semantics of fake news will almost resemble genuine news. So, technically, it is difficult for a deep neural network to "attend to" the "fake only" aspects of any news article. Automatic fake news detection is a hard problem in deception detection. While evaluating the performance of Deep Learning-based models, if all the models give higher accuracy on a test dataset, it will make it harder to validate the performance of the Deep Learning models under consideration. So, a complex problem is needed to validate the performance of a deep leaning model. LIAR is one such complex, most recent, labeled benchmark dataset which is publicly available for carrying out research on fake news detection and to model statistical and machine learning approaches for combating fake news.

In this paper, a novel fake news detection system is implemented using attention based Deep Neural Network models such as CNN, LSTM, BiLSTM, CLSTM and the performance of their attention mechanisms is evaluated by analyzing their performance in terms of Accuracy, Precision, Recall and F1-score with training, validation and test datasets of LIAR. The proposed attention mechanism will overcome the problems such as the high training time, the over-fitting of the network training and the over-thinking problem of Deep Neural Networks.By doing this evaluation, not only validating the performance of different deep learning models is carried out, but also validating the quality of the benchmark dataset itself is done. Hence, while validating a Deep Learning model, the training, validation and test data should be selected with good statistical qualities with which it is possible to realize the improvement in performance of one particular Deep Learning model.

The other sections of this paper are regularized as follows. In Section 2, the various existing deep neural networks models for fake news detection are discussed. In Section 3, an attention based deep learning models are proposed. In Section 4, the experimental setup and performance evaluation is addressed. The experimental results are explored and discussed in Section 5. In Section 6, the observations and findings are discussed briefly. Finally, the Section 7 concludes the work with future enhancement.

2. RELATED WORKS

In this section, various existing deep learning models for detecting fake news in online text is discussed. False news detection has become vital with the ever-increasing usage of social media leading to crucial research.

Kai Shu et al.(2017) have discussed some open issues and future directions in fake news detection. They have briefed the four categories of a research directions in the field that include Data,Features,Models and Application conerned with fake news detection.

Figure 1. Future challenges and outstanding issues in current false news detection on social network

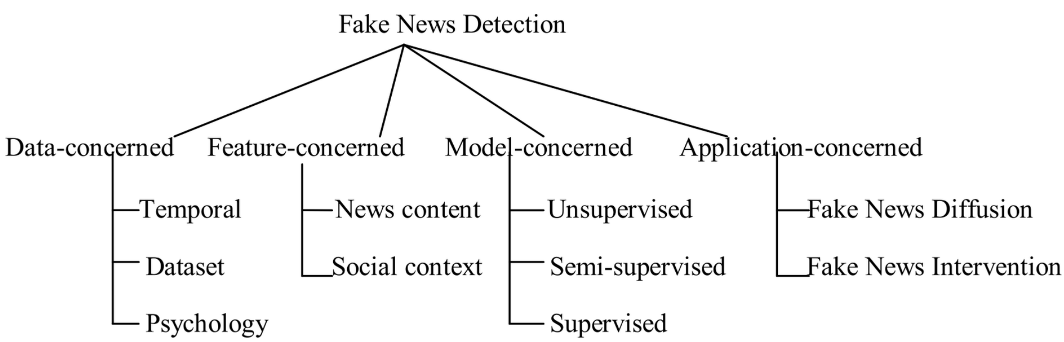


Figure 1 outlines the four categories of the research directions.

Shivam and Pradeep (2018) have presented the different extraction and modeling techniques. The authors categorized these approaches into the following six groups: Linguistic Features based Methods, Deception Modeling based Methods, Clustering based Methods, Clustering based Methods, Predictive Modeling based Methods, Content Cues based Methods and Non-Text Cues based Methods.

Wang (2017) proposed a hybrid CNN framework for combining text and meta-data. The author focused on analytical and arithmetical approaches of fake news detection. Wang (2017) introduced a new LIAR dataset. The author compared his results with various neural networks. LIAR dataset is also used for stance classification, argument mining, rumor detection and Political NLP research. In their paper, the author used word2vec embeddings for word representation and implemented stochastic gradient descent optimization algorithm. Also, (Wang, 2017) evaluated the accuracy of test and validation data. The training data was not evaluated.

Sherry et al. (2018) have presented a general approach to detect the fake news in online text. Their study is based on LIAR dataset collected from POLITIFACT.com. To classify the fake news, they have built various classifiers namely Vanilla RNN, GRU and LSTM. The results proved that GRU produces better results. Only comparison of the accuracy of test data alone was done while other scenarios such as training and validation data were not carried out. The focus was not on any other evaluation metrics like Precision, F1-score and Recall.

Abdullah et al. (2019) proposed a feature engineering-based framework. The feature vectors were extracted from twitter dataset such as count vector, word level vectors, N-gram vectors and character level vectors. The authors used five well known machine learning algorithms, such as SVM, Naïve Bayes, Logistic Regression and RNN models. The results showed that the SVM outperforms the other algorithms. The Chile earthquake 2010 dataset was used which is not a recent dataset. They used TF-IDF for word representation and also RMSProp optimization algorithm for training their models.

Pritika et al. (2019) discussed a fake news detection model based on Bi-directional LSTM_ recurrent neural network. They obtained the dataset from open Machine Learning Repository at Kaggle. The authors explored both unidirectional LSTM_RNN and Bidirectional_LSTM_RNN models with various adaptive optimizations training (i.e) RMSProp, AdaGrad and Adam. The proposed model works well for both balanced and imbalanced high dimensional dataset. The exploration of the deep learning model with attention mechanism of the fake news detection system was not carried out. Also, no evaluation metrics like Precision, F1-score and Recall was used.

Tariq et al. (2018) used LIAR/LIAR PLUS human annotated dataset from fact checking websites. They proposed single BiLSTM and Parallel-BiLSTM models with Adam optimization technique. They concluded with notable results of deep neural network models used for a binary categorization tasks (true,false) and also six ways categorization task (pants on fire,false,mostlyfalse,half true,mostly true, true).They explored categorical-cross entropy loss function and also compared the test and validation data only. The training data was not considered.

Junaed et al. (2019) have used three datasets namely LIAR dataset, Fake or real news dataset and combined dataset. They explored on lexical and sentiment feature extraction, n-gram feature extraction, feature extraction using empathy tool and pretrained word embedding techniques. Also they implemented and compared deep neural network models such as CNN, LSTM, CLSTM and HAN (Hierarchical Attention Network). They compared their work with the various traditional machine learning algorithms like SVM, Linear Regression, Naïve Bayes, KNN. They proved that Naïve Bayes outperforms for smaller dataset and proposed Conv-HAN works well for all three datasets. They did not include any attention mechanism in their approach.

In this paper, the main focus is on an attention based deep neural networks for fake news detection in social networks. And also did an extensive analysis with respect to Precision, F1-score and Recall, Accuracy on the training, testing and validation performance with the LIAR dataset.

3. MODELING A FAKE NEWS DETECTION SYSTEM

3.1 Fake News Detection

An automatic fake news detection problem is more demanding than the deception detection on online reviews or opinions due to majority of communicative media messages like Facebook posts, Tweets, TV/radio conversations, speeches and dialogues only contain precise comments (Wang,2017). Anyhow, the absence of human annotated fake news datasets is a major barrier for implementing the comprehensive and in-depth deep models towards this direction(Wang,2017).

Kai Shu et al. (2017) states that the fake news detection function F can be used to conclude if the news article a is a false news or not. If there is Awhich represents the number of social news activities among n users for a news article, then it is defined as F: A® {0, 1} such that,

$$F(A) = \begin{cases} 1, \ if \ \ a \ \ is \ \ a \ \ piece \ \ of \ \ false \ \ news, \\ 0, \ if \ \ otherwise \end{cases} \tag{1}$$

The false news detection is commonly specified as a binary categorization problem since fake news is basically a misrepresentation on information according to media bias detection theory. So, this misrepresentation is generally modeled as a binary classification problem.

3.2 Proposed Work

The following Figure 2 outlines the idea of overall process involved in fake news detection.

3.2.1 Data Preparation

The tsv Format LIAR Dataset is used for Training, Validation and Testing. For doing binary classifications, the classes belonging to 'false', 'pants-fire' and 'barely-true' are labeled as '0' (zero). The classes belonging to 'half-true', 'mostly-true' and 'true' are labeled as '1' (one).

Tokenizing Training Data: Using Keras Preprocessing, the news text (statements) is tokenized.

3.2.2 GloVe: Global Vectors for Word Representation

Word embeddings: Word embedding is the most widely used best practice in Natural Language Processing (NLP) (n.d.). Retrieved from https://ruder.io/deep-learning-nlp-best-practices/index.html#optimization.Using the pre-trained embeddings content, the overall work of Word embedding dimensionality which is a task-oriented is carried out. A smaller dimensionality works better for more syntactic tasks like Entity Identification (Kai Shu et al., 2017) or Part-Of-Speech (POS) tagging (Shivam& Pradeep, 2018) however for higher dimensionality is more suitable for linguistic tasks such as Recommendation Systems and Opinion Mining (Andrea et al., 2020).

GloVe is a unique word embedding method for attaining vector representations for words in NLP. This model is trained on non-zero vectors of word-word co-occurrence probabilities in a corpus. It provides the result of linear substructure of word vector space representation. The cosine similarity or Euclidean distance between two different words brings an efficient method for calculating the lexical relationship of the analogous words.

The nearest neighbor's algorithm rarely addresses this issue but identifies the closest words that are available outside of human's dictionary. For example, the related words such as computer, computers, computing, electronic, machine, software will occupy closest location in GloVe vector space (n.d.).Retrieved from https://nlp.stanford.edu/projects/glove/.

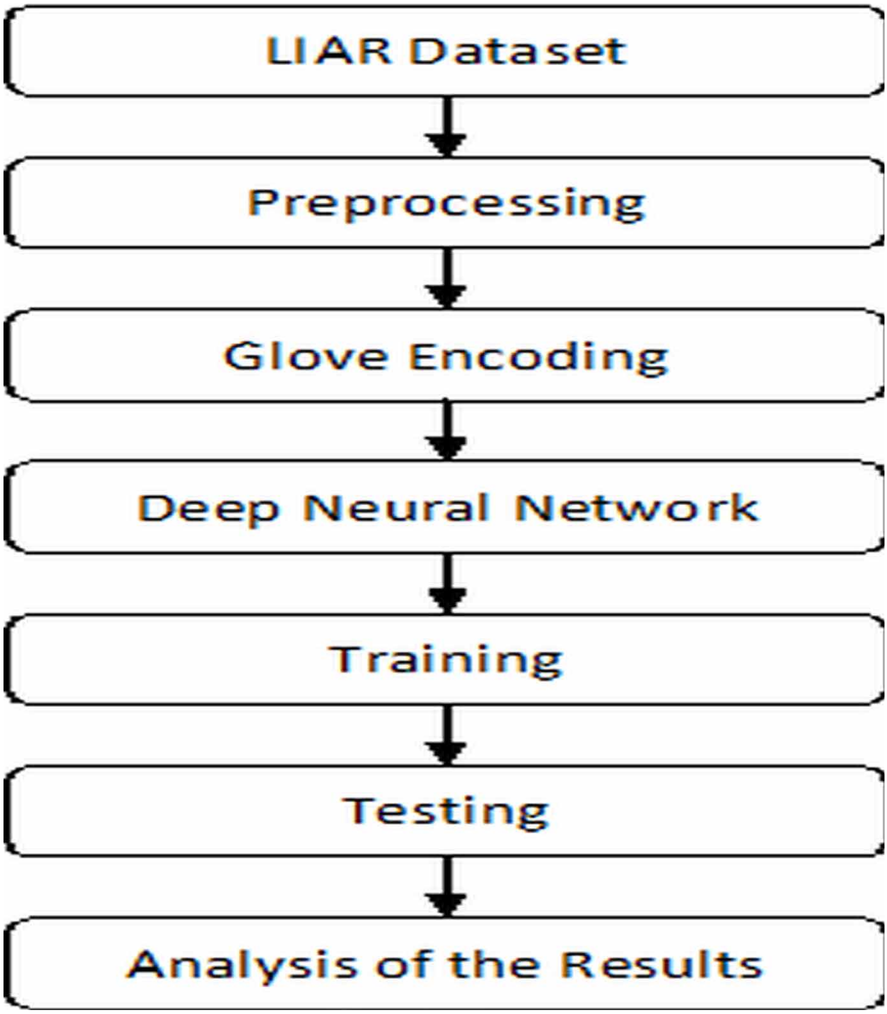
Pre-trained word vectors are usable for solving other related tasks. In this work, a 100-dimensional Pre-trained GloVe word vectors is used to form the embedding matrix that can be used in deep neural networks.Then an embedding matrix is prepared where stored Vectors for Glove Embedding is loaded from the saved Glove file glove.6B.100d.txt. The embedding of the training data using loaded glove vectors will result in an embedding matrix that can be used in a deep neural network at the embedding layer.

3.2.3 Attention Based Deep Learning Models

The term "dropout" refers to dropping out units (both hidden and visible) in a neural network [wiki]. It refers to avoiding some neurons/units and these rejected units are not treated at the time of training. This technique resolves the co-adaption issues in deep neural networks. Dropout is a regularization technique for deep neural networks in NLP (Srivastava et al., 2014). It's mainly used for any dense neural networks i.e. CNN, RNN, LSTM and etc., The dropout is only applied for training phase. While using the dropout, the layer weights are much larger than normal network. So, we scaled the weights before used. In most scenarios the good value of dropout rate is 0.5. accessed 25 July 2017, <https://ruder.io/deep-learning-nlp-bestpractices/index.html#optimization>. Dropout is the transparent way to avoid Deep Neural Networks from Overfitting used on classification, machine vision system and voice recognition (Srivastava et al., 2014).

Attention Mechanism in Neural Networks: Attention mechanism is a significant improvement in deep learning exclusively for Natural Language Processing tasks such as summarization, image generation, language modeling and translation. It is used to improve the performance of the sequence-to-sequence model by Bahdanau. The ability of this model is to focus on the subset of its features. This mechanism which accepts the entire input sequences and selectively chooses the specific elements from

Figure 2. Fake News Detection Process



that sequence to yield the output. Attention is a prominent mechanism used in a broad range of neural architectures. This encoder-decoder model is broadly used in NLP functions and also other domains like computer vision, speech processing, recommender systems, healthcare and self-driving cars.

Attention processes are as follows: 1. Input 2. Encoder 3. Compute attention vectors 4. Compute context vector 5. Decoder 6. Output. The hidden states of every element in the input sequence are generated by the encoder. Generally, these states are called as values or keys. By using the last decoder hidden state and all encoder hidden states the attention vectors are calculated. These attention vectors are multiplied with the encoder hidden states called context vectors. This context vector is combined with previous decoder output and give into the Decoder produces the output. Attention can be explained as alignment/attention vectors of weights: so, to predict a word in a sentence, it is estimated how fully it is connected with (or attend to) other words and add their values weighted by the attention vector as the similarity of the output accessed 24 June2018, <https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html>.

Using attention, a context vector c_i is obtained based on decoder past hidden states ($s_1 \ldots s_m$) that can be used together with the current hidden states of encoder h_i for prediction. The context vector c_i is calculated as an average of the past hidden states weighted with the alignment scores a_i:

$$c_i = \sum_j a_i s_j \tag{2}$$

$$a_i = soft\max(f_{att}(h_i, s_j)) \tag{3}$$

The attention function $f_{att}(h_i, s_j)$ calculates an alignment score between the current hidden state h_i and the past hidden state s_j. The weights of a_i is calculated by using Softmax function. There are different kinds of attention mechanisms such as additive attention, multiplicative attention, self-attention, and key-value attention accessed 25 July 2017, <https://ruder.io/deep-learning-nlp-best-practices/index.html#optimization>.

3.2.4 Training Optimization

The optimization algorithm is the main part of the deep neural network model used to reduce the training error or loss function and achieve the better performance. The loss function is calculated by the difference between the ground truth and the predicted output.

The commonly available optimization algorithms are Gradient Descent, Momentum, Adagrad, RMSprop and Adam. The Adam algorithm does a good practice and compares fairly to other stochastic optimization methods. This algorithm is recommended as the default optimization method for deep learning applications. In this work, Adam optimization is implemented because of the mentioned reason.

Adaptive Moment(Adam) Estimation algorithm is an association of RMS(Root Mean Square)prop and Stochastic gradient descent with momentum.It is a fast and a popular optimization method for training the deep neural networks.It is also well suited for large-scale datasets and models. It estimates the adaptive learning rates for every parameter i.e., the exponential weighted average of the gradient and squared of the calculated gradient. To be more specific, Adam algorithm operates as follows:

$$\begin{aligned} v_t &= p_1 v_{t-1} + (1 - p_1) dw \\ s_t &= p_2 s_{t-1} + (1 - p_2) dw^2 \\ \frown{v}_t &= \frac{v_t}{1 - p_2^t} \\ \frown{s}_t &= \frac{s_t}{1 - p_2^t} \\ w_{new} &= w_{old} - \frac{\alpha \frown{v}_t}{\sqrt{\frown{s}_t} + \varepsilon} \end{aligned} \tag{4}$$

3.2.5 Attention Based Deep Learning Models for Fake News Detection

3.2.5.1. Attention Based CNN Model

The Figure 3 shows the architecture of the proposed CNN based Fake News Detection Network.

CNN is multilayer feed-forward neural networks. First, the sentences of the text are tokenized into words. Then,the words are converted into a word embedding matrix. This is especially used in 100- dimensional pretrained GloVe embeddings. The default dropout rate of 0.5 is used. The CNN includes 128 filters with size of 3 and also max pooling layer of pool size 4 is chosen. The proposedCNN model is streamlined by Adam optimizer with learning rate of 0.001 to diminish the binary-cross entropy loss. The sigmoid activation function is used in the final layer. Finally, the dense layer is used to classify the fake news. The proposed CNN model is trained with 3 epochs.

3.2.5.2. Attention Based LSTM Model

Figure 4 shows the architecture of proposed LSTM based Fake news Detection Network.

LSTM units are core components for the layers of a Recurrent Neural Network (RNN). The LSTM comprised of memory blocks called cell and three gates. The proposed LSTM model is pretrained with 100 dimensional GloVe embeddings. The output of embeddings is passed to LSTM and dense layer over sigmoid activation. The model is optimized using Adam optimizer and learned through 10 epochs.

3.2.5.3. Attention Based BiLSTM Model

Figure 5 shows the architecture of the proposed BiLSTM based Fake news Detection Network.

The fake news does not fully contain false information. It is mixed with the true information. So, to identify the variation of news, BiLSTM model (i.e.) this model maintains the information from both the past and future, is implemented. It is pretrained with Glove embeddings. The default dropout rate of 0.5 is used and is also optimized by Adam optimizer with a learning rate of 0.001. The proposed BiLSTM model is trained with 1 epoch. The output dense layer is activated by sigmoid function.

3.2.5.4. Attention based CLSTM Model:

Figure 6 shows the architecture of the proposed CLSTM based Fake news Detection which is a Hybrid of a CNN and LSTM Networks.

The CLTM is a combination of CNN and LSTM model. 100 dimensional GloVe word embeddings is used for word representation. The CNN layer output is fed into the max pooling layer and then to the LSTM. A batch of 128 filters with size of 3is used.Finally, sigmoid as the activation function in the dense layer is used.

4. EXPERIMENTAL SETUP AND EVALUATION METRICS

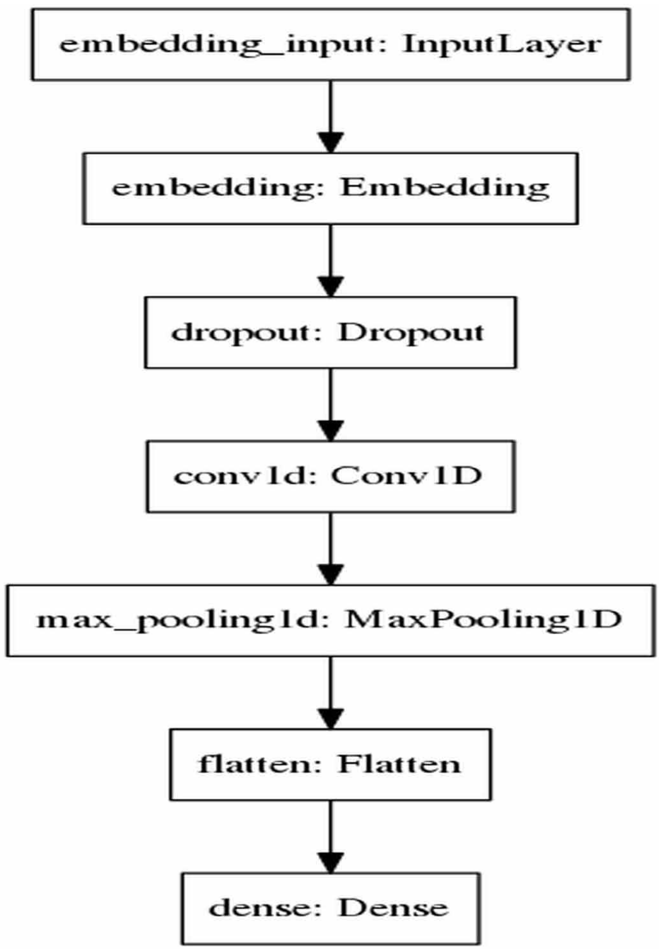
Python version 3.5 has been used to implement the proposed deep learning based fake news detection models. Python 3.5 and other necessary packages and libraries under Lubuntu 16.04 LST operating system were installed to implement the proposed system. Using a laptop with Intel Core I7 processor and 16GB RAM this evaluation was carried out.

4.1 Datasets

Wang(2017) proposed a publicly available larger benchmark dataset called LIAR, that includes more than 12k human annotated brief statements extracted from POLITIFACT.com. These labels are represented for truthfulness, subject, context/venue, speaker, state, party and history. For trueness, the LIAR dataset has six close-grained multi class labels such as pants-fire, false, mostly-false, half-true, mostly-true, and true. These six label sets are almost balanced in size. The statements were collected from a variety of broad-casting media, such as TV interviews, speeches, tweets as well as debates and they cover a large range of areas such as the economy, health service, taxation and elections.

LIAR-plus is a Politifact based extended version of LIAR dataset(Tariq et al.,2018). In this dataset, automatically obtaining the claims have associated with the justification given by the humans

Figure 3. The CNN Model used for Fake News Detection



in the fact-checking. At the end, this article has a summary section named 'our ruling'. The summary contains a certain justification sentences set correlated to the statements. These filtered sentences are supposed to improve the accuracy of the classification process.

4.2 Libraries and Packages Used

The following are some of the important packages or libraries that were used to implement the proposed fake news detection system (i.e) TensorFlow,Keras,Sklearn,NLTK,Numpy,Pandas and Matplotlib.

4.3 Parameter Settings

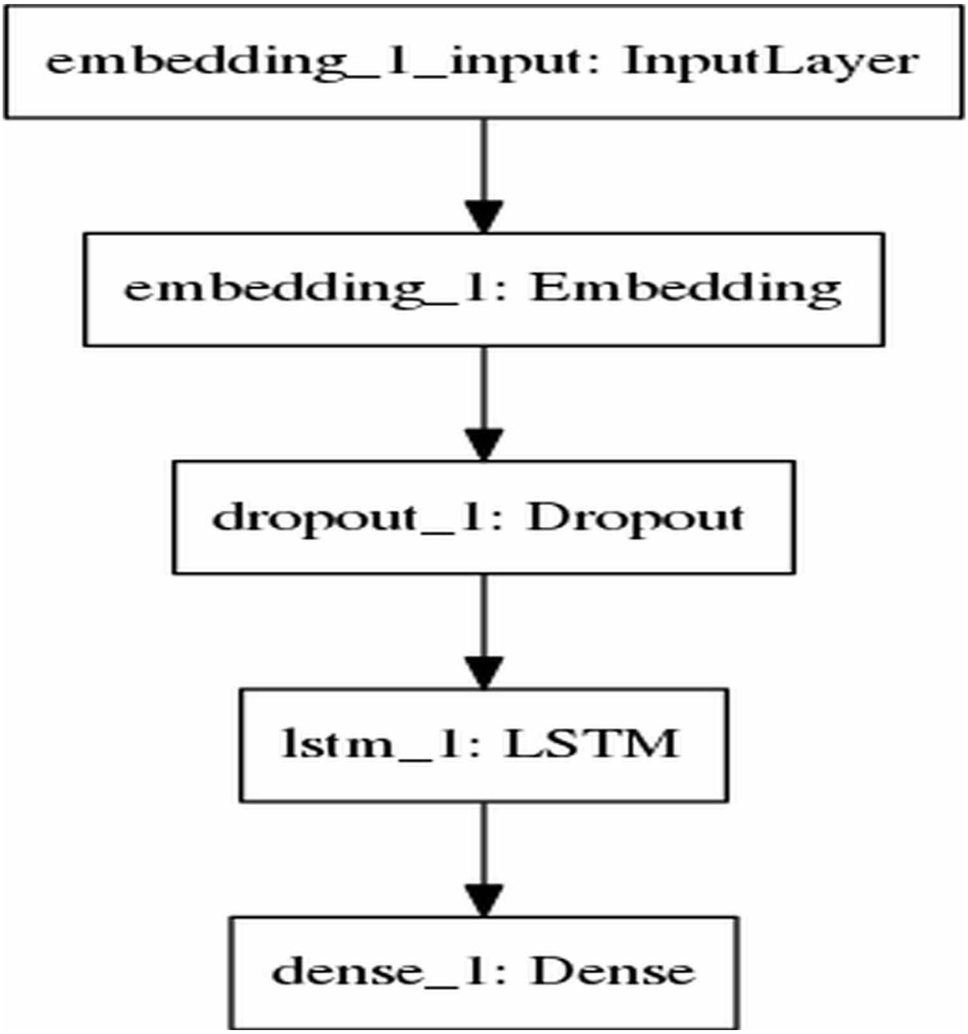
In Table 1, the measured batch-wise average of loss and accuracy during each epochs of training with different deep learning network model is presented.

The following Table 2 shows the parameters of the different layers of the proposed CNN based Fake news Detection Network.

The following Table 3 displays the parameters of the different layers of the proposed LSTM based Fake news Detection Network.

The following Table 4 displays the parameters of the various layers of the proposed BiLSTM based Fake news Detection Network.

Figure 4. The LSTM Model used for Fake News Detection



The following Table 5 shows the parameters of the various layers of the proposed CLSTM based Fake news Detection Network.

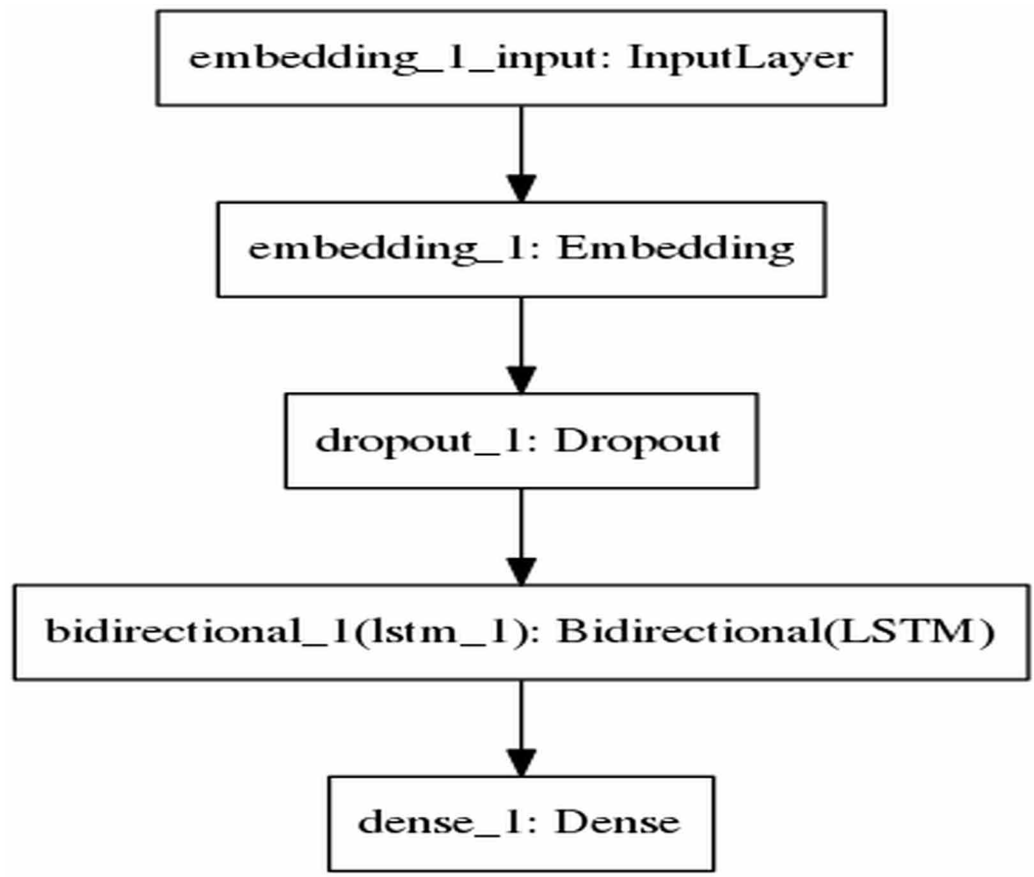
4.4 Evaluation Metrics

To evaluate the performance of the implemented deep learning based fake news detection systems, their performance with suitable evaluation metrics has to be first evaluated that have been proposed in previous works such as (Kai Shu et al.,2017). Here, the metrics that have been used for evaluating out fake news detection models is discussed.

Generally, the fake news detection problem is a binary classification issue. According to them, to conclude if a news article is false or not, the true positive counts, true negative counts, false positive counts and false negative counts from the results are considered while deriving these metrics.

The following four counts are important in these metrics

Figure 5. The BiLSTM Model used for Fake News Detection



- True Positive (TP): The portion of predicted fake news samples that are actually annotated/marked as fake news in the dataset.
- True Negative (TN): The portion of predicted true news samples that are actually annotated/marked as fake news in the dataset.
- False Negative (FN): The portion of predicted true news samples that are actually annotated/marked as fake news in the dataset;
- False Positive (FP): The portion of predicted fake news samples that are actually annotated/marked as fake news in the dataset.

By using the above counts/metrics, the following metrics of a typical classification problem are formulated:

4.4.1 Accuracy

Accuracy measures the similarity between predicted false information and true false information. Higher accuracy signifies the overall better performance of the classifier.

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (5)$$

Figure 6. The CLSTM Model used for Fake News Detection

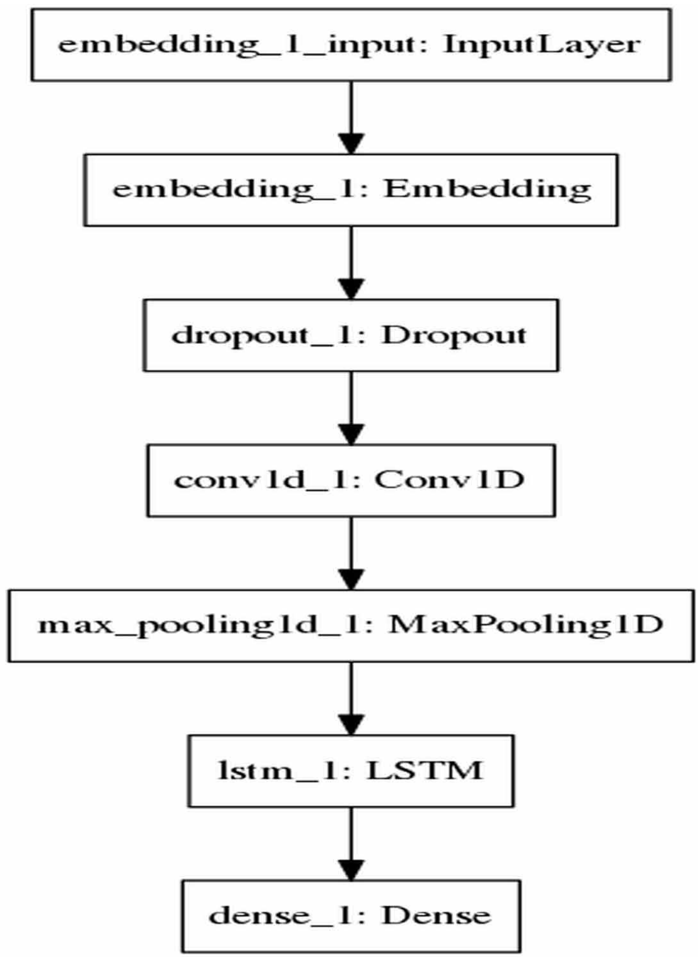


Table 1. The Parameters and Metrics Used

The Network Models	CNN, LSTM, CLSTM & BiLSTM
The Optimization Algorithm	Adam
The Learning Rate	0.001
Loss Function	binary_crossentropy
Embedding_Dimention	100
Vocabulary_Size	40000
Other Parameter	Keras Defaults
The Epochs of Training	10
The Training Batch Size	128
Metrics used During Training	**'MSE, Accuracy, Precision**
Validation and Testing	**Recall and F1-Score**

Table 2. The Parameters of the proposed CNN Network. Model: "sequential 1"

Layer (Type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 300, 100)	924900
dropout_1 (Dropout)	(None, 300, 100)	0
conv1d_1 (Conv1D)	(None, 298, 128)	38528
max_pooling1d_1 (MaxPooling1)	(None, 74, 128)	0
flatten_1 (Flatten)	(None, 9472)	0
dense_1 (Dense)	(None, 1)	9473
Total params: 972,901 Trainable params: 48,001 Non-trainable params: 924,900		

Table 3. The Parameters of the proposed LSTM Network. Model: "sequential 1"

Layer (Type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 300, 100)	924900
dropout_1 (Dropout)	(None, 300, 100)	0
Lstm_1 (LSTM)	(None, 128)	117248
dense_1 (Dense)	(None, 1)	129
Total params: 1,042,277 Trainable params: 117,377 Non-trainable params: 924,900		

Table 4. The Parameters of the proposed BiLSTM Network. Model: "sequential 1"

Layer (Type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 300, 100)	924900
dropout_1 (Dropout)	(None, 300, 100)	0
bidirectional_1 (Bidirectional)	(None, 256)	234496
dense_1 (Dense)	(None, 1)	257
Total params: 1,159,653 Trainable params: 234,753 Non-trainable params: 924,900		

Table 5. The Parameters of the proposed CLSTM Network. Model: "sequential 1"

Layer (Type)	Output Shape	Param #
embedding_1 (Embedding)	(None, None, 100)	924900
dropout_1 (Dropout)	(None, None, 100)	0
conv1d_1 (Conv1D)	(None, None, 128)	38528
max_pooling1d_1 (MaxPooling1)	(None, None, 128)	0
Lstm_1 (LSTM)	(None, 128)	131584
dense_1 (Dense)	(None, 1)	129
Total params: 972,901 Trainable params: 48,001 Non-trainable params: 924,900		

4.4.2 Precision

Precision measures the fraction of all identified false news that is noticed as fake news. This measure is used to identify which news is fake. The higher precision can be easily obtained by creating fewer positive predictions when the dataset is unbalanced. The test dataset of LIAR has 553 fake news samples and 714 real news samples (if it is considered as a binary classification problem). So, LIAR data set is almost a balanced one. So, in the proposed case, higher precision only will signify better performance of the fake news classifier.

$$\Pr ecision = \frac{\left|TP\right|}{\left|TP\right| + \left|FP\right|} \tag{6}$$

4.4.3 Recall

$$\operatorname{Re} call = \frac{\left|TP\right|}{\left|TP\right| + \left|FN\right|} \tag{7}$$

Recall is used to measure the perception. If the precision and recall value are high then the classifier is producing accurate high precision results along with high recall positive results.

4.4.4 F1-Score

Thus, F1 is used to balance between the precision and recall that can give an overall prediction performance of false news detection.

$$F1 - score = 2\ x\ \frac{precision \times recall}{precision + recall} \tag{8}$$

These metrics are generally used in the machine learning approaches and thus enabled to evaluate the performance of a classifier from different perspectives. If a classifier gives High values of the Precision, Recall, F1, and Accuracy, then it signifies better performance of the classifier.

5. RESULTS AND DISCUSSIONS

This implementation with four different deep learning models namely, CNN, LSTM, CLSTM, BiLSTM were trained with the LIAR training dataset and validated with the validation dataset at each epoch of training.

The batch-wise calculation of metrics is quite different from the calculations of the same metric with the whole dataset. The training and validation graphs only show the batch-wise calculated metrics. During training, each batch of data is trained and the metrics will be calculated for that batch along with the average of already calculated batches. But, while testing, if the same metric for the whole applied data is measured at once, then practically, those values will be a little bit different from the ones that have been computed batch-wise during training.

As highlighted in the training and validation graphs in the following part,during training phase, the performance, regarding accuracy is increased in each epoch and loss is decreasing with reference to the increase of training in the event of CNN and BiLSTM.

But, if it is carefully noticed in the validation performance graphs, then it can be decided that in the event of CNN and BiLSTM, the performance of validation is not at all increasing with respect to the increase of training.

The graphs clearly indicate that the training and validation performance of LSTM and CLSTM are almost equal and very worse. They are not at all getting any training with respect to the increase in training epochs.

The training and validation performance with respect to precision, recall and F1-score are presented in the following part. They also show that the CNN and BiLSTM based network models are getting progressive training over epochs. Even though the training performance of LSTM and CLSTM is constant and high with respect to F1-Score and recall, they are actually not at all getting any training over the epochs and give the worst performance with respect to accuracy and precision.

The two tables namely, Table 6 and Table 7 show the performance of 4 different network models (6 different trained networks).

Table 6 lists the performance using LIAR training dataset.

Table 7 lists the performance using LIAR validation dataset.

Table 8 lists the performance using LIAR Test dataset. The performance with validation dataset and test dataset is important and they are used to validate the performance of these trained networks.

The various charts in this section display the performance of CNN, LSTM, CLSTM and BiLSTM deep learning models with Training, Testing and Validation datasets.

Even though the network was experienced with a huge training dataset of LIAR, only after 10 epochs of training, the trained networks were able to identify around 75% of its trained records. This is realized by observing the accuracy, precision, and F1 scores of these three different data sets.

The three bars in each of the five set of bars clearly shows that the validation dataset is little bit complex than the testing dataset.

Figure 7 is the chart that shows the comparison of the models in terms of Accuracy with three different datasets and six different models.

Figure 8 is the chart that shows the comparison of six different models in terms of Accuracy with test datasets.

Figure 9 is the chart that shows the comparison of the modelsin terms of Precision with three different datasets and six different network models.

With respect to precision, the CNN and BiLSTM models performed better than all other models, while classifying the test dataset. The chart in Figure 10 clearly shows that the Performance of CNN and BiLSTM are comparatively good in terms of precision which is important in validating the prediction performance of a binary classification based fake news detection system.

Table 6. Performance of the models for LIAR Training Dataset

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.73	0.72	0.84	0.77
CNN*	0.63	0.63	0.83	0.71
LSTM	0.56	0.56	1.00	0.72
CLSTM	0.56	0.56	1.00	0.72
BiLSTM	0.67	0.70	0.73	0.71
BiLSTM*	0.59	0.59	0.89	0.71

CNN* (Result with 3 Epochs Training), BiLSTM* (Result with 1 Epoch Training)

Table 7. Performance of the models for LIAR Validation Dataset

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.54	0.54	0.75	0.63
CNN*	0.53	0.53	0.85	0.66
LSTM	0.52	0.52	1.00	0.68
CLSTM	0.52	0.52	1.00	0.68
BiLSTM	0.54	0.55	0.65	0.60
BiLSTM*	0.52	0.52	0.94	0.67

CNN* (Result with 3 Epochs Training), BiLSTM* (Result with 1 Epoch Training)

Table 8. Performance of the models for LIAR Test Dataset

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.54	0.57	0.73	0.64
CNN*	0.56	0.58	0.83	0.68
LSTM	0.56	0.56	1.00	0.72
CLSTM	0.56	0.56	1.00	0.72
BiLSTM	0.55	0.59	0.64	0.61
BiLSTM*	0.56	0.57	0.93	0.70

CNN* (Result with 3 Epochs Training), BiLSTM* (Result with 1 Epoch Training)

The chart in Figure 11 shows the performance of six different models in terms of Recall with test datasets. Even though, the performance of LSTM and CLSTM are 100% (indicated as 1), they are not good in precision.

The chart in Figure 12 clearly shows the differences in Recall with the test dataset.

The chart in Figure 13 clearly shows the performance in terms of F1-score with three different datasets and six different models.

Figure 7. Comparison of the models in terms of Accuracy

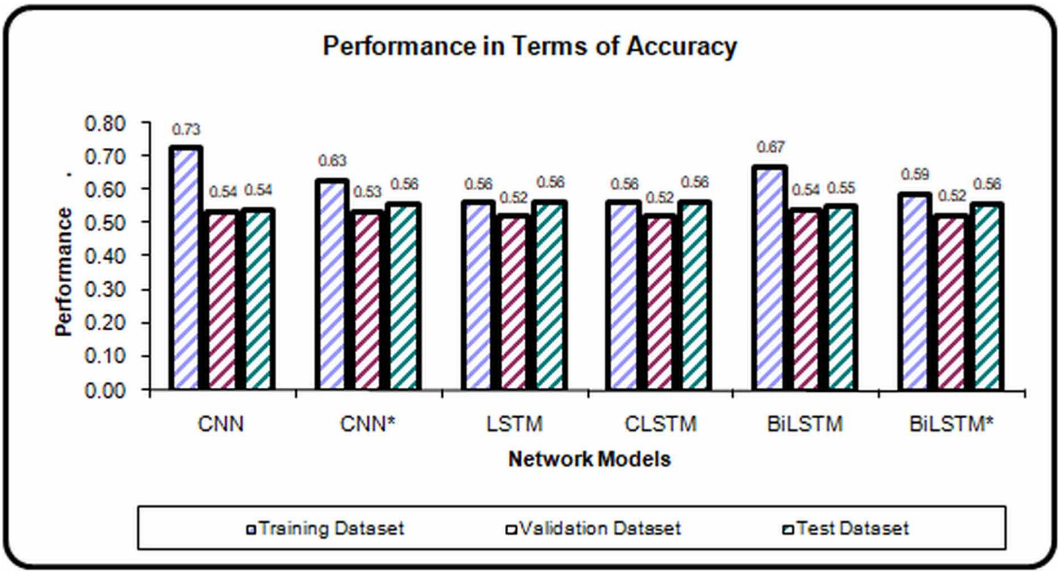


Figure 8. Comparison of the models in terms of Accuracy (with Test Data)

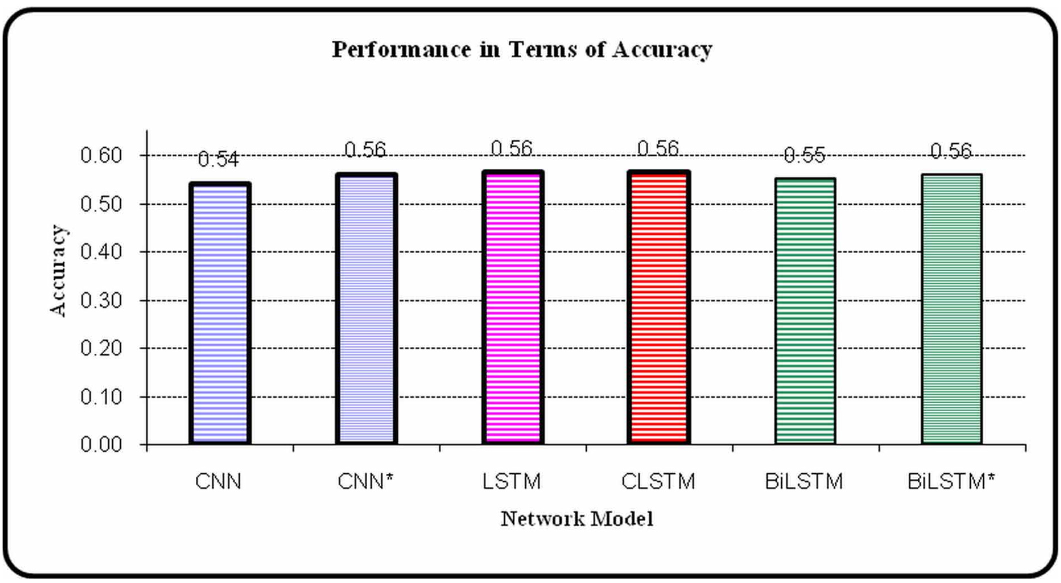


Figure 9. Comparison of the models in terms of Precision

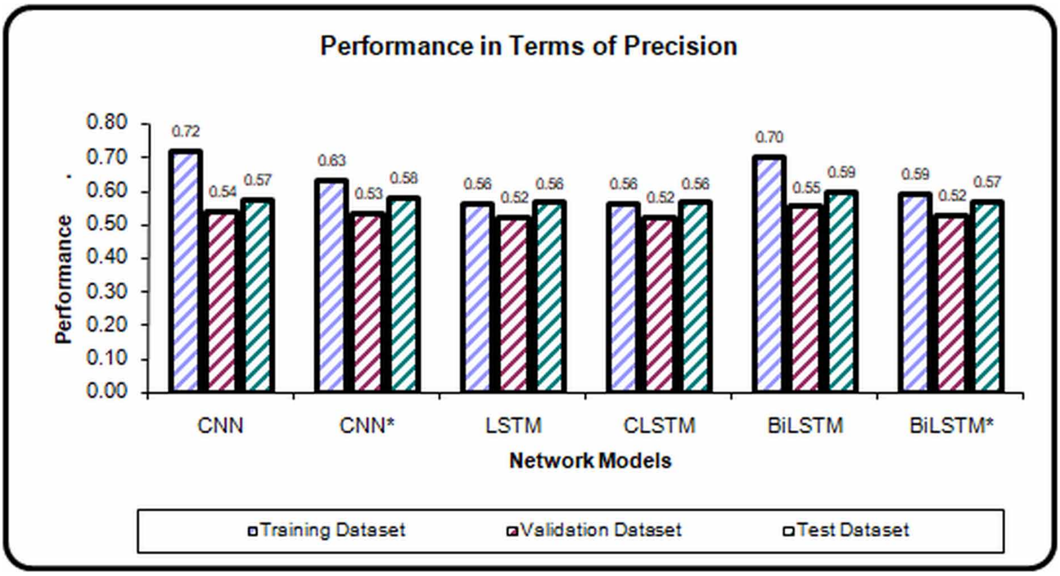
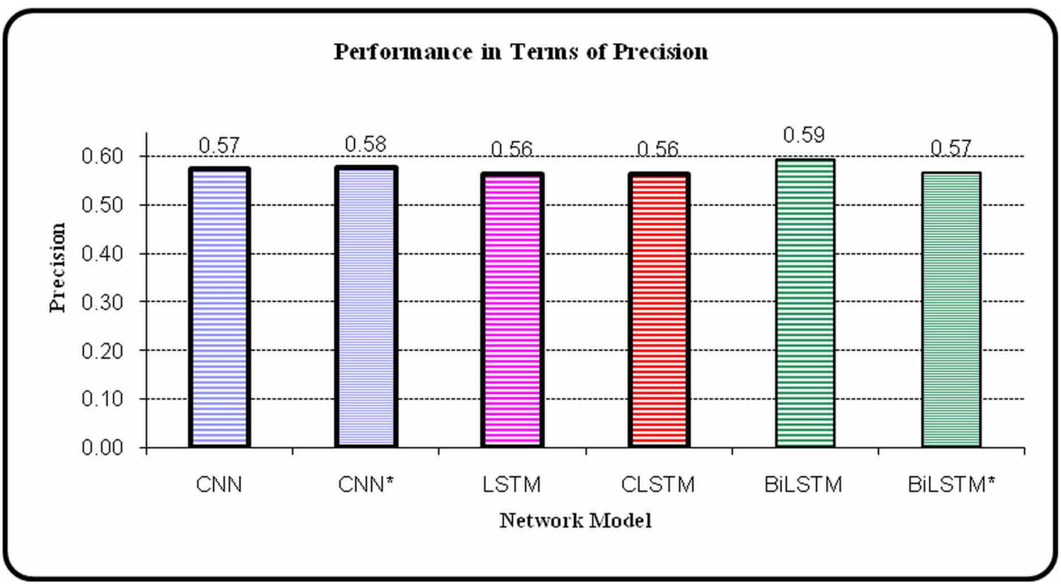


Figure 10. Precision (with Test Data)



Some of the published papers discuss the performance in terms of only the F1-score in their papers. This implementation of CNN and BiLSTM has achieved good performance in terms of F1-score (while considering their high Precision) (Figure 14).

Even though the performance of the two LSTM models are good in terms of F1-score, their performance in terms of Accuracy and Precision is not good compared to CNN and BiLSTM. So,

Figure 11. Performance of the models in terms of Recall

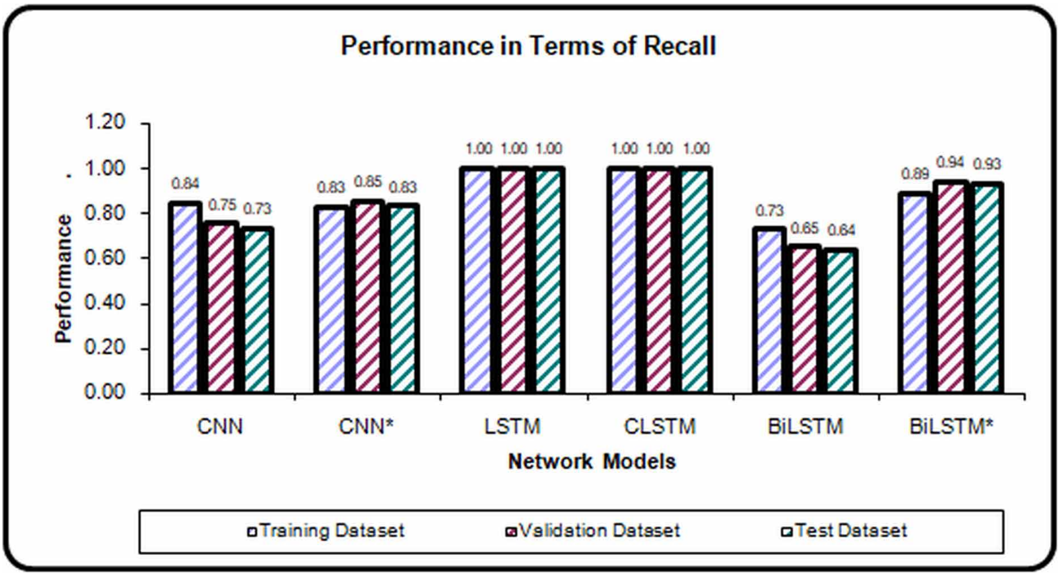


Figure 12. Recall (with Test Data)

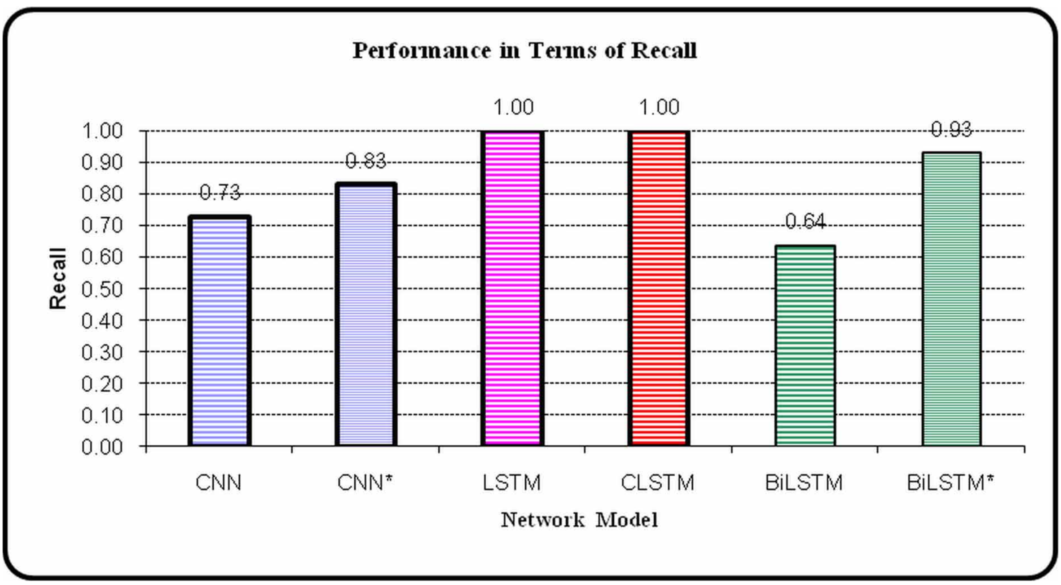


Figure 13. Performance in terms of F1-Score

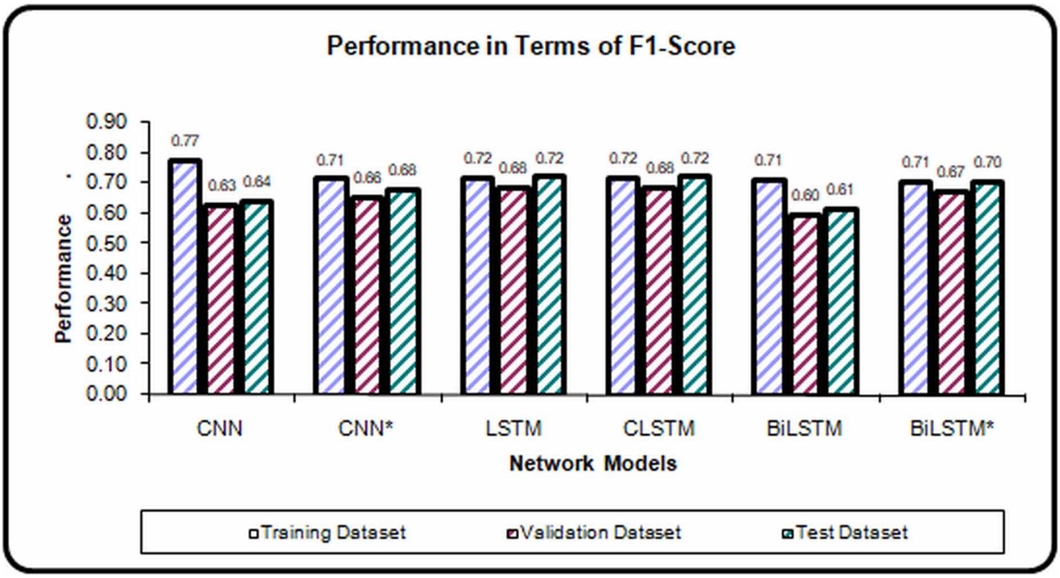
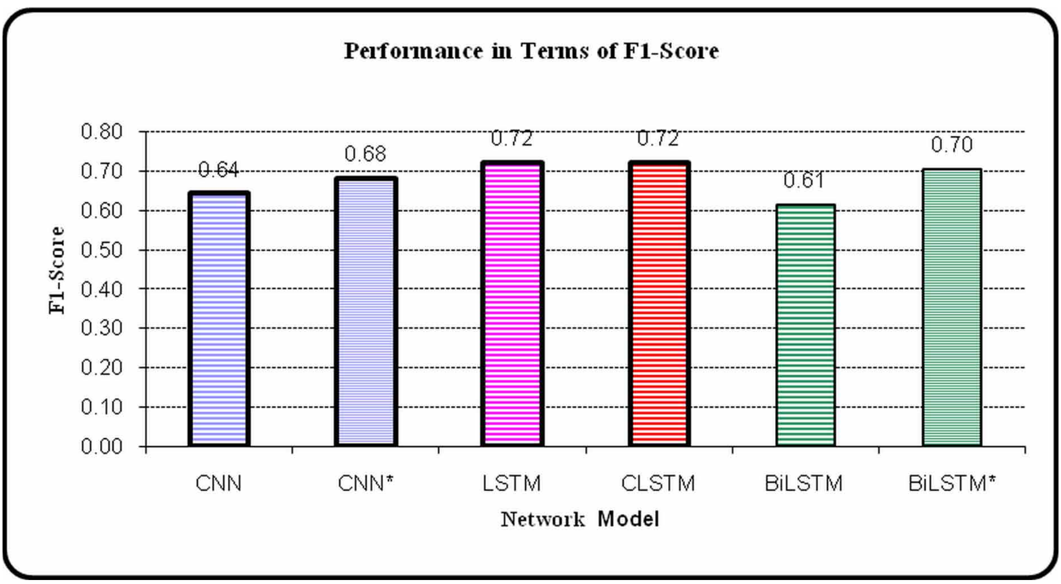


Figure 14. F1-Score (with Test Data)



the both LSTM and CLSTM are not good in overall classification performance with the LIAR fake news dataset.

5.1 Comparison of Results with Previous Works

Since the referred works used only F1-score in common, here, comparison of the performance of these implemented models has been done with the reference works. The following Table 9 and graphs in Figure 15 show the comparison of the results.

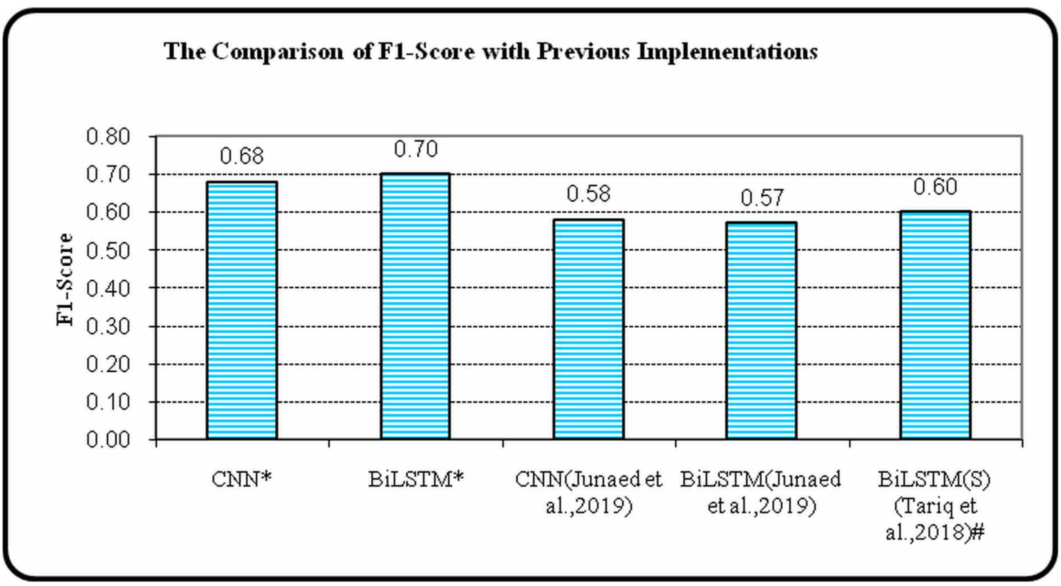
Table 9 lists the comparison of the previous works.

Table 9. Comparison of Performance with Previous works

Model	Accuracy	Precision	Recall	F1_Score
CNN*	0.56	0.58	0.83	0.68
BiLSTM*	0.56	0.57	0.93	0.70
CNN((Junaed et al.,2019)	0.58	0.58	0.58	0.58
BiLSTM(Junaed et al.,2019)	0.58	0.58	0.58	0.57

CNN* (Result with 3 Epochs Training), BiLSTM* (Result with 1 Epoch Training)

Figure 15. Comparison of F1-Score with Previous works



As shown in the following graph of Figure 15, the proposed implementation of CNN and BiLSTM produced good results in terms of F1-Score.

6. OBSERVATIONS AND FINDINGS

As far as the LIAR fake news benchmark dataset is concerned, almost all the models suffer in finding an optimal solution space during training. Because of the very complexity of the fake news, the training always tends to over-fit in to a subspace and that leads to poor performance in terms of all the metrics. The significant findings are listed below

- Even though the performance of the two LSTM models are good in terms of F1-score, their performance in terms of Accuracy and Precision is not good compared to CNN and BiLSTM.
- The Reason for high F1-Score in the case of LSTM is because of its ability to only predict the normal news with almost 100% accuracy.
- Even though the training was done up to 10 epochs, the better network model was not always found at 10th epoch. This signifies that the validation and testing performance are not increasing with respect to the increase in training epochs.
- Even though the performance of CNN and BiLSTM are much lower with respect to Recall and F1-Score, they are superior because, they only provided good results in terms of precision which is much important in validating the prediction performance of a binary classification based fake news detection system.
- In the previous works (Tariq et al., 2018) and (Junaed et al.,2019) they only presented the performance of their model in terms of F1-score. With respect F1-score, the proposed implementation of CNN and BiLSTM models performed better. But it can be realized and concluded that F1-score only is not enough to validate the performance of the fake news detection system.
- Implementation results of this model CNN, LSTM, CLSTM and BiLSTM are much higher than the previous implementation in (Junaed et al.,2019).
- In these tests, F1-score of 0.7 has been achieved which is the most achievable result by previous models. Even these LSTM models achieved F1 score around 0.72 but those are not considered as an achievement because of their poor performance in terms of Accuracy and Precision.
- In almost all trials, the LSTM and CLSTM models suffered with overfitting issues while training with LIAR dataset.
- Even though the training in the case of CNN and BiLSTM were progressive with respect to the training data, it is realized and concluded that they also struck at bad solutions in the huge problem space. This is very clearly indicated by the results of validation and tests.
- Even with the CNN, always the model belonging to the highest epoch of training didn't give better validation and test performance.

7. CONCLUSION

In this work, successful implementation of four deep neural networks based fake news detection models for binary classification and an extensive analysis on their training, validation and test performance with the LIAR dataset has been carried out. The training performance graphs shown in the previous section clearly shows the nature of training and validation.

The batch-wise calculated metrics clearly shows the network performance is not improving with respect to the increase in training epochs.

The LIAR bench mark dataset is really a good bench mark dataset for validating a fake news detection system because achieving better results with the validation data and the test data is not so easy for any deep learning model. So, if a state-of-the-art deep learning model has been able to achieve a little more improvement with respect to precision and accuracy, while testing it with validation data or test data, then certainly that particular model is definitely a superior one for fake news detection.

As mentioned earlier, unlike other natural language processing and machine learning tasks, fake news detection is more complex from the perspective of the deep learning-based attention mechanisms involved in their design. These results clearly prove the complexity in attention towards fake aspects of the news. The reason for getting poor performance with the network attention mechanisms is that the lexicons of fake news will almost resemble to that of genuine news. Hence, technically, it is significantly difficult for a deep neural network to "attend to" the fake only views of the news article.

As far as the complication of the LIAR dataset is concerned, the existing deep neural networks are not able to "attend to" the fake aspects of the news and hence only giving a marginal performance.

To make a deep neural network to "attend to", the fake aspects of the news in a better way, the news/data must be presented in a better way with some enhanced distinguishable features of fakeness. For that, in future works, the more advanced NLP based techniques, preprocessing and feature selection techniques and better encoding methods for enhancing the performance of training, can be explored. Also, the possibility of hybrid network models and sophisticated learning optimization techniques to improve the performance, can be explored and future works will address these issues.

REFERENCES

Aldwairi & Alwahedi. (2018). Detecting Fake News in Social Media Networks. *The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks(EUSPN 2018).*

Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your Evidence: Improving Fact-checking by Justification Modeling. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85-90. doi:10.18653/v1/W18-5513

Ba, J., & Frey, B. (2013). Adaptive dropout for training deep neural networks. *Advances in Neural Information Processing Systems*. https://papers.nips.cc/

Bahad, P. A., Preeti Saxena, A., & Raj Kamal, B. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *International Conference on Recent Trends in Advanced Computing (ICRTAC 2019).* doi:10.1016/j.procs.2020.01.072

Dozat, T., & Manning, C. D. (2017). *Deep Biaffine Attention for Neural Dependency Parsing.* A conference paper at ICLR 2017. https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/

Galassi, A., Lippi, M., & Torroni, P. (2020). *Attention in Natural Language Processing*. arXiv:1902.02181v2[cs.CL].

Girgis, S., Amer, E., & Gadallah, M. (2018). Deep learning Algorithms for Detecting Fake News in Online Text. *13th International Conference on Computer Engineering and Systems (ICCES)*. doi:10.1109/ICCES.2018.8639198

Khan, K. Iqbal, & Afroz. (2019). *A Benchmark Study on Machine Learning Methods for Fake News Detection.* arXiv:1905.04749 [cs.CL].

Khan, S. A., Alkawaz, M. H., & Zangana, H. M. (2020). The Use and Abuse of Social Media for Spreading Fake News. *IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2019).*

Li, Z., Gong, B., & Yang, T. (2016). Improved Dropout for Shallow and Deep Learning. Advances in Neural Information Processing Systems(NIPS 2016).

Lopez, M. M., & Kalita, J. (2017). *Deep Learning applied to NLP*. arXiv: 1703.03091.

Melamud, O., McClosky, D., Patwardhan, S., & Bansal, M. (2016). The Role of Context Types and Dimensionality in Learning Word Embeddings. *Proceedings of NAACL-HLT*. doi:10.18653/v1/N16-1118

Oshikawa, Qian, & Wang. (2018). *A Survey on Natural Language Processing for Fake News Detection*. Academic Press.

Parikh, S. B., & Atrey, P. K. (2018). Media-Rich Fake News Detection: A Survey. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. doi:10.1109/MIPR.2018.00093

Pérez-Rosas. (2017). *Automatic Detection of Fake News*. arXiv:1708.07104.

Plank, B., Søgaard, A., & Goldberg, Y. (2016). Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. doi:10.18653/v1/P16-2067

Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*. doi:10.18653/v1/D16-1103

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations*, *19*(1), 22–36. doi:10.1145/3137597.3137600

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*.

Tanvir, Mahir, Akhter, & RezwanulHuq. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. *7th International Conference on Smart Computing & Communications (ICSCC).*

Wang. (2017). *Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection*. arXiv:1705.00648v1[cs.CL].

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). *Recent Trends In Deep Learning Based Natural Language Processing*. arXiv:1708.02709v8 [cs.CL].

S. P. Ramya currently pursuing her Ph.D. degree at the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu, India. She has obtained her M.Tech. degree and a B.Tech., degree. Currently, she is working as an Assistant Professor in the Department of Computer Science and Engineering at Periyar Maniammai Institute of Science and Technology, Vallam, Thanjavur. Her research interest includes data mining, Web technologies, deep learning, and artificial intelligence.

R. Eswari is an Assistant Professor in the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu, India. She has obtained a Ph.D. degree from the Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu. She authored and co-authored more than 25 research papers and published papers having good citations. Her research interest includes Computational Intelligence, Multi-objective Optimization, Cloud Computing, Distributed Scheduling, Wireless Sensor Networks.