I.J. Information Technology and Computer Science, **2022, 5, 58-69**
Published Online on October 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2022.05.05

Multi-stage Transfer Learning for Fake News Detection Using AWD-LSTM Network

Sirra Kanthi Kiran
Dept. of IT, GVP College of Engineering (A), Visakhapatnam, Andhra Pradesh, 530048, India
E-mail: sirrakanthikiran@gmail.com

M. Shashi
Dept. of CS&SE, A.U. College of Engineering, Visakhapatnam, 530003, India
E-mail: smogalla2000@yahoo.com

K. B. Madhuri
Dept. of IT, GVP College of Engineering (A), Visakhapatnam, Andhra Pradesh, 530048, India
E-mail: drkbmadhuri@gvpce.ac.in

Received: 13 March 2022; Revised: 13 June 2022; Accepted: 17 August 2022; Published: 8 October 2022

Abstract: In the recent decades, the automatic veracity verification of rumors is essential, since online social media platforms allow users to post news item or express opinion towards a circulating piece of information without much restriction. The intention of fake news is to make the readers believe in inaccurate information, where the detection of fake news by using content is a difficult task. So, the auxiliary information: user profile, social engagement of the users, and other user's comments are useful in the detection of fake news. In this manuscript, a novel multi-stage transfer learning approach is introduced for an effective fake news detection, where it utilizes user's comments as auxiliary information to detect whether the given tweet is true or false. The stances of the response tweets contain opinions on news/rumors are often used for verifying the veracity of the circulating information. In order to devastate the effects of the specific rumors at the earliest, the multi-stage transfer learning approach automatically predict veracity of rumors jointly with the stances of their response tweets. The proposed multi-stage transfer learning is an inductive transfer learning variation that is used to forecast the stance of responses, then to identify fake news. The proposed model's effectiveness is evaluated on the two-benchmark datasets: semEval-2017 task 8 and PHEME. The proposed model outperformed the existing approaches by obtaining a classification accuracy of 64.30% and 65.30%, an F-measure of 65.95% and 63.90% on semEval-2017 task 8, and PHEME on event-wise datasets.

Index Terms: Fake news classification, Inductive transfer learning, Language model, Long short-term memory network, Multi stage transfer learning.

1. Introduction

False information is purposefully put online and disseminated as actual news. Nowadays, people rely heavily on online social networks for both social contact and news, so news spreads quickly through these networks. The use of social media websites for news consumption has both benefits and drawbacks. One benefit is quick and cost-free access to local information on mobile devices at any time, wherever. The main drawback, however, is the inability to confirm the accuracy of the news after it has been published, which has a negative effect on society. The detection of fake news is a difficult issue that calls for the application of the most recent developments in computer science research, such as deep learning. Existing fake news detection techniques require large amounts of labelled datasets to train the models, which in real time may not be possible. Hence, this study aims to identify methods that would perform well even if the dataset size is small.

In recent times, the online social media platforms have become an integral part of human life, where it provides societal identity and information on all kinds of events from all over the world [1,2]. However, the majority (around 57%) of the people relied on social media for news in that most of the news on social media is inaccurate. The impact of fake news is far more devastating than the expectation [3,4]. In the recent decades, the social media websites are empowered by any person, who having a web-associated gadget. The rumor stances are in the form of comments and the rumors are pivotal signs to identify drifting bits of gossip and to indicate its veracity [5]. Hence, it is important to consider the crowd opinions related to a trending message on social media before classifying it as a true or false [6,7]. However, it is un-likely that the stances are available in the early propagation of a particular message [8], where extra

time has to be elapsed before a sufficient number of comments emerge. Therefore, the automated framework for detecting rumor's veracity without confining to the availability of stances is desirable [9.10]. The recent advances in machine learning in the form of transfer learning for language modelling are explored to propose a superior framework for early detection of fake news.

The inductive transfer learning is a revolutionized computer vision task, due to the availability of pre-trained models. The effectiveness of inductive transfer learning is proved in various natural language processing tasks: question/answering, sentiment analysis, and recommendation systems [11,12]. Language modelling in natural language processing is started with generating conventional word embedding's, Glove, and word2Vec that are shallow representations of words in a text without context. Traditionally, such static word embedding is used at the first layer of the model and then hidden layers of the neural network are needed to be trained for a target task from scratch thereby requiring large dataset for training. This static word embedding's fail to capture contextual information. Contextualized word embedding is generated from language models like Universal Language Model Fine tuning (ULMFIT) [13], Embedding's from Language Model (ELMo) [14], and Bidirectional Encoder Representations from Transformers (BERT) [15] which are useful to understand higher-level concepts such as anaphora involving long-term dependencies, agreement, and text negation.

In this article, a multi-stage transfer learning approach is introduced for the fake news detection. The multi-stage transfer learning approach is a variant of the inductive transfer learning framework that contains four stages. Initially, a pre-trained language model is developed that learns the general language syntax and semantics. Further, the pre-trained language model is fine-tuned based on the domain specific language to capture task specific syntax and semantics in the target domain. In the third stage, the supervisory learning is employed to further refine the task specific language model to generate task specific encoder in the process of building a classifier for stance detection. Finally, the rumor classification model: Averaged SGD (Averaged Stochastic Gradient Descent) Weight Dropped Long Short Term Memory (AWD-LSTM) [16] network is developed on top of the task specific encoder that extracts essential features to find the veracity of the tweets. In the resulting phase, the AWD-LSTM architecture's effectiveness is tested by means of f-measure, accuracy, recall, Matthews Correlation Coefficient (MCC), and precision.

This manuscript is designed as follows: a few articles related to fake news detection are surveyed in section 2 and the problem statement is presented in section 3. The theoretical background and experimental evaluation of the AWD-LSTM architecture are specified in the sections 4 and 5. The conclusion of the paper is given in section 6.

2. Related Work

In the fake news detection application, the existing models used numerous techniques such as machine learning, deep learning and natural language processing. Hunt, Allcott, et al. [17] explored societal beliefs to analyse the impact of fake news on the 2016 US Presidential elections. Some of the existing studies adopted non-neural models such as support vector machines, decision tree, Naïve Bayes classifier, etc. Comparatively, larger number of existing studies adopted neural network models for fake news detection. R.K. Kaliyar, *et al*, [18] presented a new neural network named FNDNet for an effective fake news detection. The presented FNDNet model automatically learns the discriminative feature vectors for classifying the fake news. Several performance metrics like accuracy, false positive, f-score, precision, true negative, recall, and Wilcoxon were utilized for evaluating the effectiveness of the presented FNDNet model on Kaggle fake news dataset. H. Yuan, *et al*, [19] presented a new model named Domain Adversarial and Graph Attention Neural Network (DAGANN) for detecting the fake news. The presented DAGANN model extracts domain invariant feature vector to identify the fake news across domains/events. The simulation results showed that the presented DAGANN model obtained significant performance in fake news detection across different domains/events on two multi-media datasets of Weibo and Twitter.

J.A. Nasir, et al, [20] has combined Recurrent Neural Network (RNN) and Convolutional Neural Network for an effective fake news detection. The developed CNN-RNN model obtained superior performance in fake news detection related to the non-hybrid baseline models on the FA-KES, and ISO datasets. S.R. Sahoo, and B.B. Gupta, [21] used multiple feature extraction techniques for an automatic fake news detection on the social networks. This literature incorporates the behavior of multiple features that were associated with the Facebook accounts, and analyze the behavior of the Facebook accounts using deep learning based analyzer. A. Choudhary and A. Arora, [22] presented a new linguistic model that extracts readability, sentimental, grammatical and syntactic features of news for an effective fake news detection. The extensive experiments and the comparative analysis on the benchmark datasets showed the superiority of the linguistic model than the existing models. S. Sheikhi, [23] combined content based feature extraction techniques, Extreme Gradient Boosting Tree (xgbTree), and Whale Optimization Algorithm (WOA) for detecting fake news articles. The presented model majorly includes two phases: (i) used content based feature extraction techniques for extracting the informative feature vectors, and (ii) applied xgbTree-WOA model for classifying the news articles using the informative feature vectors. In the resulting section, the presented model effectiveness was validated on 40,000 news articles by means of f-measure and accuracy.

Y.F. Huang, and P.H. Chen, [24] introduced a novel ensemble learning model for fake news detection, where the ensemble-learning model combines four classifiers like N-gram CNN, LSTM, depth-LSTM and Linguistic Inquiry and Word Count (LIWC) CNN. The optimized weights of the ensemble-learning model were determined by utilizing self-

adaptive harmony search algorithm for achieving better classification accuracy in fake news detection. F.A. Ozbay and B. Alatas, [25] has presented a two-step technique to identify fake news on the social media. In the initial step, many data pre-processing methods such as document term matrix and term frequency weighting methods were used for converting the unstructured datasets into the structured datasets. In the second step, around 23 artificial intelligence approaches were applied in the transformed datasets for fake news detection. P. Bahad, et al [26] has combined RNN and Bi-directional LSTM network for fake news detection. In this literature, two unstructured online datasets were used for assessing the efficiency of the developed model by means of accuracy. A. Kumar, et al, [27] presented a hybrid model for rumor detection that includes CNN and a filter wrapper optimized Naïve Bayes classifier. Initially, the CNN model was employed for extracting the texture feature vectors and then information gain-ant colony optimization technique was used for discriminative feature selection. Lastly, the Naïve Bayes classifier used the discriminative feature vectors for rumor classification. A. Zubiaga, et al, [28] utilized conditional random field as a sequential classifier for classifying the non-rumors and rumors sequences on the PHEME dataset. Correspondingly, E. Kochkina, et al, [29] implemented LSTM based sequential model for rumor stance classification on the SemEval-2017 task 8 dataset.

Yuhang Wang, et al. [30] introduced the Elementary Discoursed Unit whose granularity is between word and sentence to detect fake news as soon as it is published. Zhou, Y., Ying et al. [31] proposed FND-CLIP frame work to make use of images as well as text present in the news. The authors concatenated the text and image deep learning features using a ResNet-based encoder and a BERT-based encoder, respectively. Ali M. Meligy, et al., [32] presented a fake profile identification technique called Fake Profile Recognizer (FPR), which makes use of Regular Expression (RE) and Deterministic Finite Automaton (DFA) to detect fake profiles. Fake profile detection mechanism would help improve the identification of fake news. Shubham Bauskar, et al., [33] presented a neural language model for fake news detection which makes use of both content and profile to detect fake news. Momina Shaheen, et al., explored various techniques such as Support Vector Machines (SVM), SGD, Gradient Boosting Classifier, LSTM, and CNN for sentiment analysis. Some of the techniques have also been implemented and tested in this study for bench mark performance evaluation purposes. Momina Shaheen. et. al., [34] used different machine learning algorithms to perform sentiment analysis on social media data.

Existing studies have not considered the difference between the standard form of communication in English and the informal form of communication on social media platforms. The informal form of communication has an integral relationship with standard forms of communication. Most existing work requires a large labelled dataset to train deep learning models. It has been proven that [35] AWD-LSTM based universal language models outperformed state-of-the-art models even with a dataset of size 100. Hence, the proposed framework would like to address these issues by making use of pre-trained language models with AWD-LSTM. These pre-trained language models will be further trained on domain-specific target task datasets by achieving the objective of meeting the gap between standard forms of communication and informal communication normally existing on social media platforms.

3. Problem Statement

Due to various reasons, the language used in social media deviates from the standards of a professional language; the blogs and posts in social media contains most of the times the shortcuts, abbreviations, local language implications, idioms etc. Most of the existing language models do not consider this inherent property of social media language. In order to reduce the gap between official language usage standards and the social media language, a neural language transfer-learning model is proposed in this article. The deep learning based language models are explored as they are found to be successful in all most all of the natural language processing tasks. In this manuscript, an AWD-LSTM architecture is proposed to build a deep learning-based language model to encode the tweets for detecting the rumors via stance classification. Once the model is developed, the early detection of fake news is accomplished automatically without requiring to wait for the stances of the response tweets. This is the first work that adopts the neural language model for fake news detection through multi-stage transfer learning approach.

Let us consider a dataset containing D tweets, where the response tweets are generated in response to base tweets. The authors propose multi-stage model for predicting the veracity of tweets to detect rumors automatically without requiring the stances associated with the rumor. For each tweet $tw_i \in D$ in the training set, multiple response tweets are collected as $\{tw_{i1}, tw_{i2}, .. tw_{im}\}$. Each of these response tweet tw_{ij} express a specific stance either posted directly or indirectly towards original tweet i. The stance classification is a multi-class classification problem as there are multiple class labels like support, deny, query, and comment. Supervisory learning occurs during stance classification to further fine tune the task specific language model, this model is used as a task specific encoder for feature extraction from news tweets to determine the rumor veracity without requiring corresponding stances to gain time for faster identification fake news.

4. Methodology

The proposed multi-stage transfer-learning framework is developed which involves both unsupervised and supervised transfer learning for building and refining the language models for fake news detection. The multi-stage

transfer learning approach is a variant of inductive transfer learning, which is proposed for stance prediction of response, comment tweets followed by fake news detection. The figure 1 depicts an overview of multi-stage transfer learning approach, which is composed of four stages: language model pre-training, language model fine-tuning, stance classification, and fake news identification.

Once the general language model is fine-tuned based on the domain specific text corpus, the annotated tweets are required for supervised training. In the third stage, the task specific encoder is used in the process of stance prediction of each reply tweet into one of the four classes such as support, deny, query, and comment. In the fourth stage, the task specific encoder is used for feature extraction to build a classifier for predicting the veracity of a tweet into one of the three classes labeled as true, false, and un-verified. Optimal utilization of labelled data is achieved in the multi-stage transfer-learning framework to build the fake news detector capable in the early detection of fake news even before sufficient number of comments, and their stances are available in the social media.

The figure 2 illustrates the mechanism for multi-stage transfer learning for fine tuning and predicting the stance of response tweets and fake news detection. The first component is to build a language model from large text data that learns lexical, syntactic, semantic, pragmatic features of the language (English) in general context. It involves Averaged SGD Weight Dropped LSTM as a basic unit of learning shared weights while processing the sequence of words in the text. The second component is to fine-tune the general language model for domain specific language (twitter corpus /IMDB corpus) to generate the domain specific language model. Inductive transfer learning is applied on the encoder of the domain specific language model in the third stage, to learn a stance classifier along with a task specific encoder on top of the domain specific language model. It utilizes stance labelled response tweet data to learn the model and classify the response and comment tweets for accurately predicting the stances. The stance specific encoder is employed in the final stage to extract task specific language features of a tweet to predict the news tweet veracity.

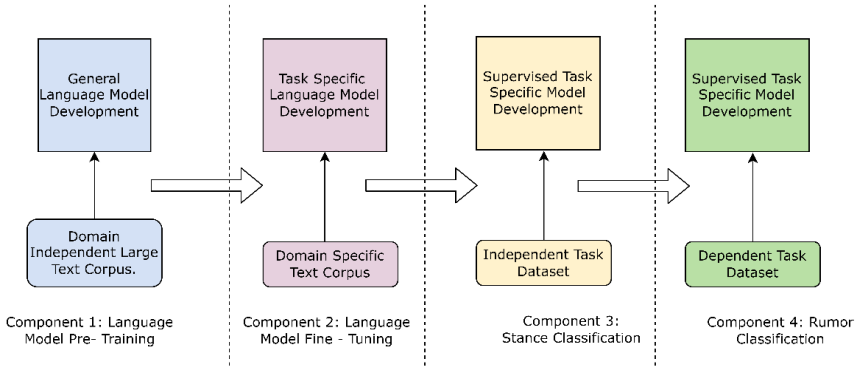

Fig.1. Multi-stage transfer learning mechanism for Fake News Detection.

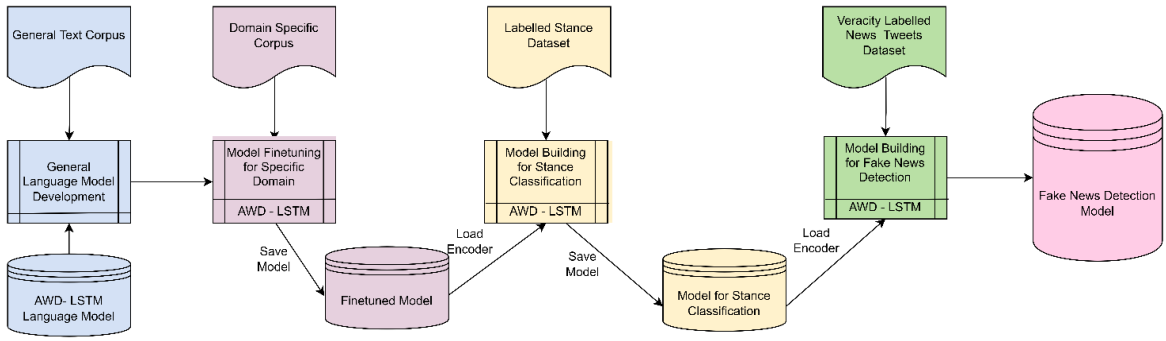

Fig.2. Multi-stage transfer learning approach for predicting stance and veracity (This entire image/Architecture has been redesigned)

4.1. Architecture of AWD LSTM

The AWD-LSTM architecture is effective in handling overfitting, which is problematic, when the language model is used in the low resource tasks. Therefore, it is used as a basic learning mechanism in all the four stages of proposed transfer learning system. The AWD-LSTM architecture utilizes Averaged Stochastic Gradient Descent (ASGD), and weight dropout methods for an effective fake news classification. The ASGD is very similar to the stochastic gradient descent algorithm, where it utilizes average gradients in all the previous steps for estimating gradients. In the weight dropout technique, randomly selected subset of the weights is forgotten by making them equal to zero, thus discarding partial information of the weight matrix propagated to the next layer. The AWD-LSTM architecture comprises of input gate i_t, forget gate f_t, cell c_t, and output gate o_t, that are expressed in the equations (1), (2), (3), and (4).

$$i_t = \sigma(W_{ih} h_{t-1} + W_{ia} a_t + b_i) \quad (1)$$

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fa}a_t + b_f) \quad (2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\,(W_{ch}h_{t-1} + W_{ca}a_t + b_c) \quad (3)$$

$$o_t = \sigma(W_{oh}h_{t-1} + W_{oa}a_t + b_0) \quad (4)$$

At the time step t, $a_t = A[t,.] \in \mathbb{R}^F$, which is denoted as the quasi-periodic feature. Further, the W and b are represented as work coefficients, $\sigma(.)$ is specified as sigmoid activation function, $tanh\,(.)$ denotes hyperbolic tangent activation function, and h_{t-1} is denoted as output of the prior LSTM unit, which is defined in equation (5).

$$h_t = o_t \odot tanh(c_t) \quad (5)$$

As represented in the figure 3, $\boldsymbol{h_t}$ has the information of the previous time steps of a cell $\boldsymbol{c_t}$ and an output gate $\boldsymbol{o_t}$. The cell states $\{\boldsymbol{c_t}|\boldsymbol{t} = \boldsymbol{1,2,..T}\}$ learns the memory information of the temporal quasi-periodic feature vectors for both short and long time period based on the dependency relation during the training process. Finally, the extracted feature vectors are represented by the last LSTM unit $\boldsymbol{h_T}$ output [36,37].

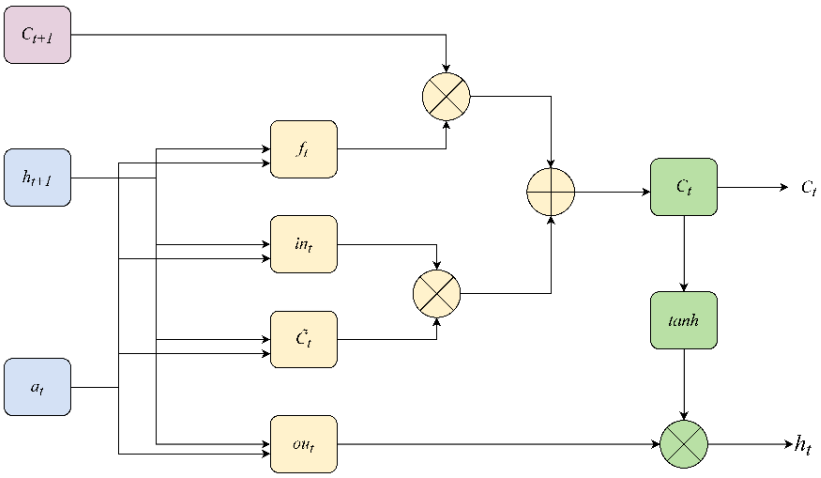

Fig.3.Architecture of the LSTM unit

4.2. Language Model Fine-tuning for the Domain Specific Data

General-purpose language like Wikipedia text will certainly have a different word distribution compared with the domain specific word distribution like finance, healthcare, etc. Similarly, the language used in the social media has larger number of non-standard tokens, emoji's, expressions in the short forms, which inherently contains the opinion of the social media user. Hence, the pre-trained language model is fine-tuned on domain specific data i.e. SemEval 2017 task 8 data in an unsupervised manner. The discriminative fine-tuning and slanted triangular learning rates has been used in order to avoid the catastrophic forgetting normally exhibited by language models to fine-tune the pre-trained model for domain specific data.

4.3. Classifier Developed for the Independent Target Task

The encoder of the model obtained by fine-tuning pre-trained language model on domain specific data that is used on top of the classifier, which is built with two fully connected layers with Softmax activation for output layer. The weights of the classifier are learnt by gradual unfreezing of weights starting from the top most layer to the lower layers in successive iterations to fine-tune the weights of the classifier for learning the independent target specific task i.e. stance detection to capture the public response to the tweet.

4.4. Classifier Developed for the Dependent Target Task

The public opinion towards a trending message is captured in the previous step; the encoder of the independent task classifier is used to build a rumor/veracity classifier with two fully connected layers with softmax activation at the end. The method for learning the weights of the classifier is similar to the stance detection module, which explained in the previous section. Gradual unfreezing of weights and triangular learning rate variation are the key for optimal training schedule to achieve better accuracy in detecting the tweets, which propagate fake news. In this manuscript, the rumor stance classification and rumor detection has been independently evaluated by the neural language model as described in figure 4, and then evaluated as a joint task using the proposed architecture for multi-stage transfer learning as described in figure 2.

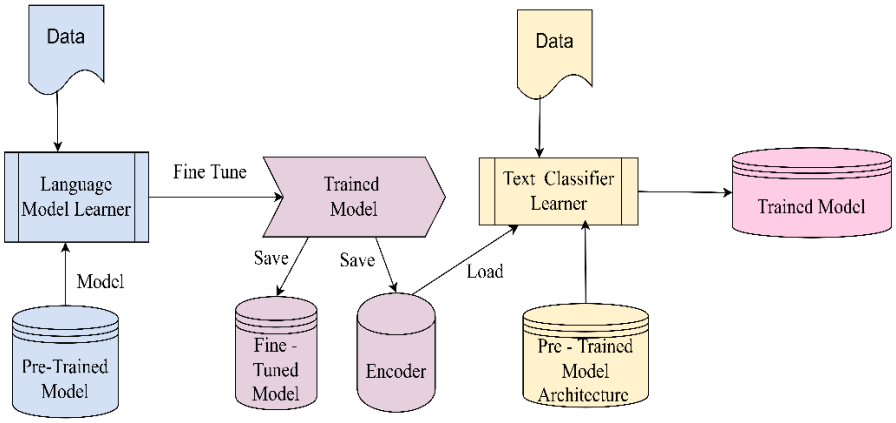

Fig.4. Inductive transfer learning for predicting rumor stance and veracity

5. Experimental Results

The AWD-LSTM architecture's effectiveness is validated using anaconda navigator 3.5.2.0 (64-bit), Python 3.7 software environment on a system with 16 GB random access memory, Intel core i7 processor, 4TB hard disk and windows 10 (64 bit) operating system. The efficiency of the AWD-LSTM architecture is evaluated by using the performance metrics like f-measure, accuracy, recall, MCC, and precision on two benchmark datasets such as semEval-2017 task 8 dataset, and PHEME dataset. In the classification procedure, the MCC performance measure estimates the percentage of the news items, which are predicted in the correct classes. On the other hand, the classification accuracy measures the closeness between the forecasted records, and the total number of fake and legitimate news articles. The precision performance measure estimates the percentage of the news items, which are precisely classified as fake news by the AWD-LSTM architecture. Similarly, the recall performance metric estimates the performance of the fake news articles, which are precisely classified as fake news. Lastly, the f-measure is the measurement of harmonic mean of precision, and recall. The mathematical representation of the f-measure, accuracy, recall, MCC, and precision is defined in the equations (6-10).

$$F - measure = \frac{2TP}{FP+2TP+FN} \times 100 \tag{6}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \times 100 \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \times 100 \tag{10}$$

Where, False Negative (FN) indicates the percentage of the fake news items, which are incorrectly determined as legitimate news. False Positive (FP) represents the percentage of the legitimate news items, which are incorrectly detected as fake news. True Negative (TN) denotes the percentage of the legitimate news items, which are precisely classified by the AWD-LSTM architecture. Further, True Positive (TP) represents the percentage of the fake news items, which are precisely classified by the proposed architecture.

5.1. Dataset Description

A. SemEval-2017 task 8 dataset

This dataset contains conversational tweets, which are discussed about 10 events with 325 threads, each discussion starts by a rumors tweet. The discussions consist of tweets reacting to those rumors' tweets. The reacted tweets are clarified for Support, Deny, Query or Comment. The dataset has been split into train, development and test sets. Conversational threads of eight events are present in train, development-sets, and conversational threads of the two events are present in the test set [38].

B. PHEME dataset

This dataset consists of 2,402 tweet conversations that belongs to nine events. Out of nine events, eight events tweet conversations are utilized for training, the left out event tweet conversations are used for testing. The outputs of

all nine experiments are integrated for computing the final output. The PHEME dataset cannot be used for stance classification as only a subset of the conversations has labels for stance [38,39]. Hence, this dataset has been used for veracity verification only, and the table 1 elaborates the statistics of both datasets.

Table 1. Statistics of the datasets

Datasets	#Discussion	#Tweets	Stance Labels				Veracity Labels		
			#Support	#Deny	#Query	#Comment	#True	#False	#Unverified
SemEval-2017	325	5568	1004	415	464	3685	145	74	106
PHEME	2402	105,354	-	-	-	-	1067	638	697

5.2. Experimental Set-up

In this manuscript, the proposed AWD-LSTM language model is used as a pre-trained language model. The AWD-LSTM architecture includes three layers, 1150 hidden activations per layer, and an embedding size of 400. The AWD-LSTM encoder with a two layer fully connected classifier has been used in this study. The hidden layer of the classifier is 50, and the batch size of 32 is used for training the model. In order to perform rumor stance classification, a 3-stage inductive transfer learning is proposed. To perform rumor veracity classification in a single task setting, an inductive transfer learning is used. The 3-stage approach is used for a multi-task setting, where stances have been utilized to predict rumor veracity.

Table 2. Experimental results of the AWD-LSTM architecture on semEval-2017 task 8 dataset

Models	**Cross-folds**	**F-measure (%)**	**Accuracy (%)**	**Recall (%)**	**MCC (%)**	**Precision (%)**
Random forest	5-folds	30.98	37.68	40.56	41.29	40.90
Naïve Bayes		40.34	42.39	44.63	43.44	43.84
Autoencoder		41.02	48.80	47.87	38.92	40.85
LSTM		42.37	42.19	41.02	30.34	42.94
AWD-LSTM		44.20	43.28	43.26	52.36	44.60
Random forest	10-folds	56.80	59.60	55.55	56.20	56.98
Naïve Bayes		54.64	59.30	58.60	58.45	59.89
Autoencoder		52.42	50.85	50.84	50.92	52.80
LSTM		53.38	53.88	53.12	53.35	54.90
AWD-LSTM		**65.95**	**64.30**	**65.56**	**66.38**	**65.22**

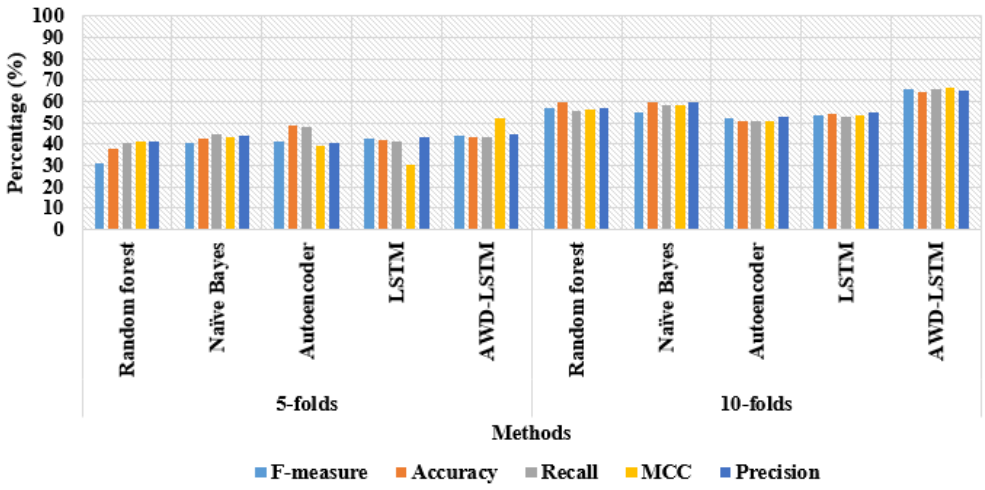

Fig.5. Comparison results of the AWD-LSTM architecture on semEval-2017 task 8 dataset

5.3. Quantitative study on SemEval-2017 Task 8 Dataset

In the fake news detection, the AWD-LSTM architecture's efficiency is validated on semEval-2017 task 8 dataset. In this scenario, the performance analysis is carried-out with different machine learning techniques and deep learning techniques with two different cross-fold validations such as 5-folds and 10-folds. Further, the performance evaluation is accomplished with 80% data training and 20% data testing. As represented in table 2, the proposed AWD-LSTM architecture attained a maximum performance in fake news detection related to the comparative classifiers like random forest, Naïve Bayes, autoencoder, and LSTM. By investigating table 2, the proposed AWD-LSTM architecture attained 65.95% of f-measure, 64.30% of accuracy, 65.56% of recall, 66.38% of MCC, and 65.22% of precision in the 10-fold

cross validation, which are higher related to the comparative classifiers on semEval-2017 task 8 dataset. Graphical presentation of the AWD-LSTM architecture on semEval-2017 task 8 dataset is depicted in figure 5.

Table 3. Experimental results of the AWD-LSTM architecture on PHEME dataset

Models	Cross-folds	F-measure (%)	Accuracy (%)	Recall (%)	MCC (%)	Precision (%)
Random forest	5-folds	49.93	47.62	40.55	40.99	50
Naïve Bayes		43.38	46.38	44.68	40.45	42.82
Autoencoder		41.42	43.85	46.88	45.90	49.84
LSTM		47.39	49.18	40.08	48.89	40.94
AWD-LSTM		40.28	50.24	41.27	50.24	42.20
Random forest	10-folds	53.84	59.67	55.50	55.57	53.18
Naïve Bayes		52.66	50.35	56.68	54.40	55.80
Autoencoder		58.49	50.65	52.80	57.28	50.80
LSTM		52.30	52.80	53.02	50.45	52.94
AWD-LSTM		**63.90**	**65.30**	**63.58**	**63.34**	**64.38**

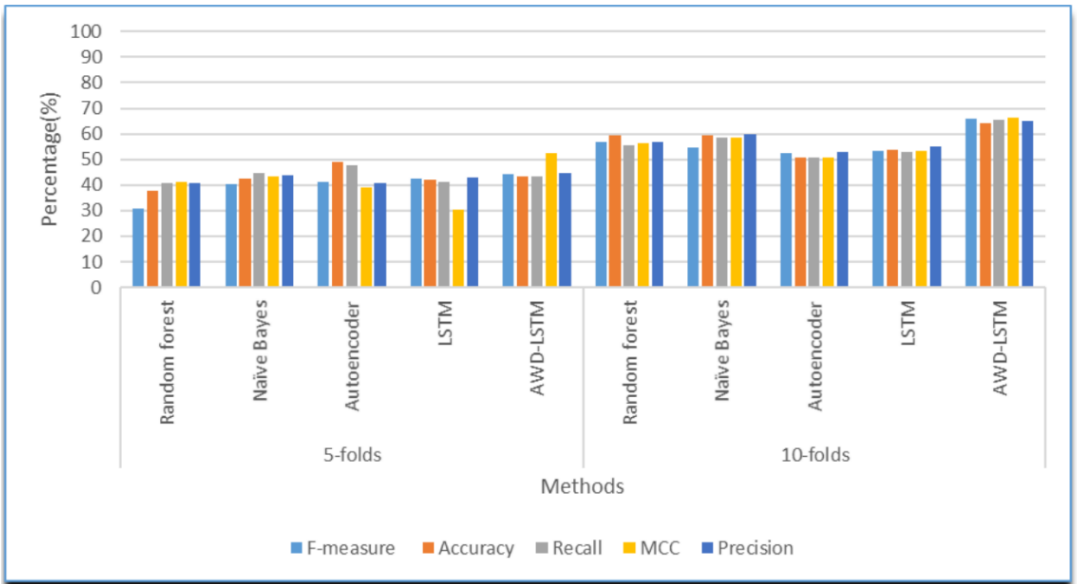

Fig.6. Comparison results of the AWD-LSTM architecture on PHEME dataset

5.4. Quantitative Study on PHEME Dataset

In this scenario, the AWD-LSTM architecture's effectiveness is validated on PHEME dataset by means of f-measure, accuracy, recall, MCC, and precision. By inspecting table 3, the developed AWD-LSTM architecture attained 63.9% of f-measure, 65.30% of accuracy, 64.38% of precision, 63.58% of recall, and 63.34% of MCC in the fake news detection, particularly in the 10-fold cross validation. The obtained experimental results of the AWD-LSTM architecture are better compared to the existing models such as random forest, naïve Bayes, Autoencoder and LSTM on the PHEME dataset. Graphical presentation of the AWD-LSTM architecture on PHEME dataset is indicated in figure 6.

5.5. Comparative Study

The comparative study between the existing models and the proposed AWD-LSTM architecture is represented in the tables 4 and 5. A. Kumar, et al, [27] used CNN to extract the textual features from the collected PHEME dataset. Additionally, the information gain and ant colony optimization algorithm were used to select the discriminative feature vectors. Finally, the Naïve Bayes classifier was used to classify the non-rumors and rumors sequences on the PHEME dataset. A. Zubiaga, et al, [28] used seven feature extraction techniques such as word count, word vectors, part-of-speech tags, use of period, capital ratio, use of question mark, and use of exclamation mark for extracting the feature values from the content of the tweets. Then, the conditional random field was employed for classifying the non-rumors and rumors sequences on the PHEME dataset. Correspondingly, E. Kochkina et al, [29] presented LSTM based sequential classifier for rumor stance classification on the SemEval-2017 task 8 dataset. The presented model achieved accuracy of 58.40% on the testing set, which is better related to the comparative models in the rumor stance classification. As denoted in the tables 4 and 5, the proposed AWD-LSTM architecture obtained effective performance in the fake news detection related to the comparative models on both PHEME and SemEval-2017 task 8 datasets.

The comparative results have been presented in the table 4, table 5. The results from the table 4 describes the

comparative results of the existing works [27, 28] and proposed AWD-LSTM. Table 5 describes the comparative results of the existing works [29] and proposed AWD-LSTM. PHEME dataset contains the data event-wise such as Germanwings crash, Ferguson unrest Charlie Hebdo, Sydney siege, Ottawa shooting. These event wise dataset size is comparatively very small compared with the entire dataset size. Hence the existing deep learning state of the art models failed to achieve results on these event-wise data. The proposed multi-stage transfer learning technique outperformed the existing state of the art deep leering models.

Table 4. Comparative results between the existing models and the AWD-LSTM architecture on PHEME dataset

Events	CNN + Naïve Bayes [27]			Conditional random field [28]			AWD-LSTM		
	Precision (%)	Recall (%)	F-measure (%)	Precision (%)	Recall (%)	F-measure (%)	**Precision (%)**	**Recall (%)**	**F-measure (%)**
Germanwings crash	56.70	56.10	56.30	54.30	56.80	50.40	**64.90**	**62.93**	**63.45**
Ferguson unrest	57.40	54.10	55.70	56.60	39.40	46.50	**65.48**	**63.30**	**62.30**
Charlie Hebdo	55.60	54.10	54.80	54.50	56.20	63.60	**62.83**	**64.38**	**62.10**
Sydney siege	54	59.90	51.90	56.40	38.50	51.20	**62.03**	**63.46**	**64.39**
Ottawa shooting	54.90	60.10	57.30	54.10	58.50	59	**63.78**	**62.94**	**63.50**

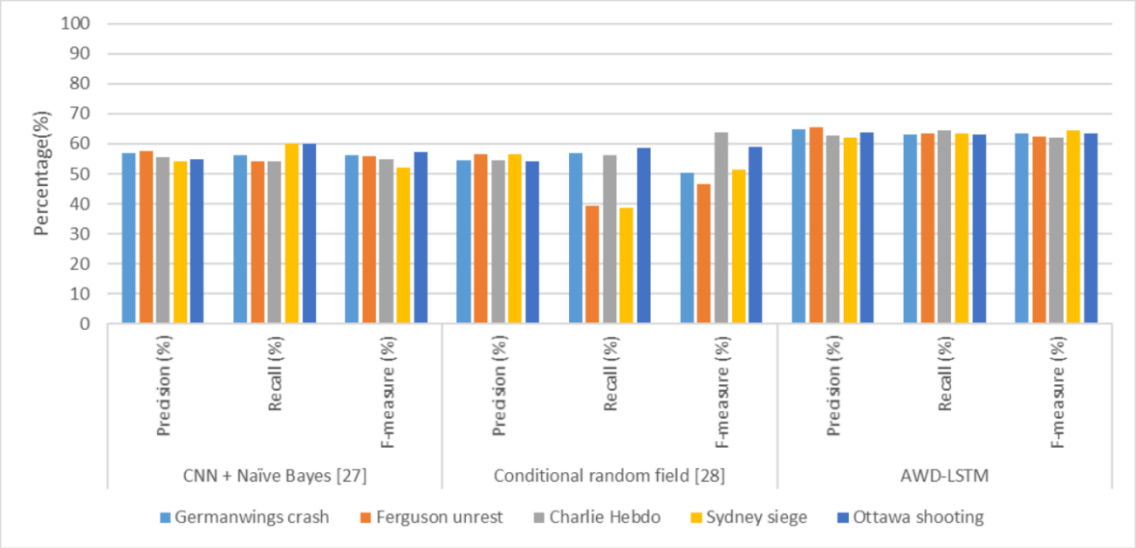

Fig.7. Comparison results of the AWD-LSTM architecture with Existing models

6. Conclusion

In this manuscript, a multi-stage transfer-learning framework is proposed for fake news detection based on the availability of relevant social media data. The existing fake news detection methods completely rely on stances of the comments related to the news tweets, where the existing methods cannot detect the fake news until sufficient number of stances are available. The proposed multi-stage transfer-learning framework is capable in detecting the fake news earliest, as it does not have to wait for the stances related to a tweet in order to check its veracity. In this study, the efficiency of the proposed multi-stage transfer-learning framework is tested on two benchmark datasets such as semEval-2017 task 8 and PHEME by means of f-measure, accuracy, recall, MCC, and precision. The proposed multi-stage transfer-learning framework attained classification accuracy of 64.30% and 65.30% on the semEval-2017 task 8 and PHEME datasets, which are superior related to other comparative models. The experimental outcome showed that the dependent tasks such as veracity is verified effectively by utilizing the knowledge gained through independent tasks such as crowd opinion through comments. This research has also shown that the classification of individual events in the dataset is as accurate as the classification accuracy of the entire dataset, even though the individual event's dataset size is small. The results achieved in this study have shown that even if the dataset size is very small, with the multi-stage transferred learning approach, effective results can be achieved. As a future extension, the multi-stage transfer-learning framework is extended by including the meta-data such as user profile, network structure, and web search engine results for early prediction of veracity without waiting for the comments.

Table 5. Comparative results between the existing model and the AWD-LSTM architecture on SemEval-2017 task 8 datasets (Table has been modified to include the Precision, Recall, F-Measure)

Models	Accuracy (%)	Precision(%)	Recall(%)	F-Measure(%)
LSTM based sequential classifier [29]	58.40	44.21	44.28	43.47
AWD-LSTM	**64.30**	**65.22**	**65.56**	**65.95**

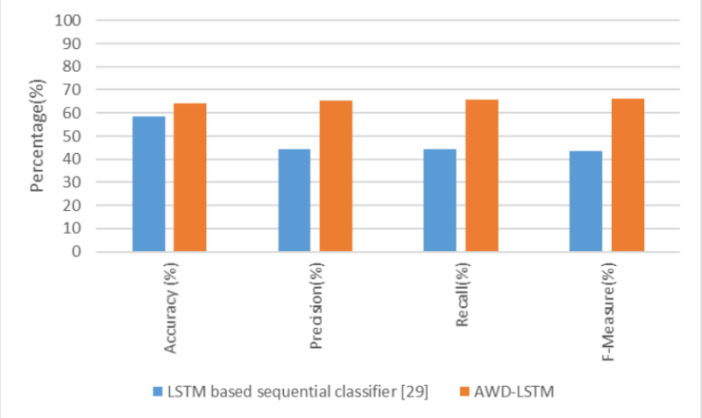

Fig.8. Comparison results of the AWD-LSTM architecture with LSTM based sequential classifier (This figure has been modified to reflect the changes)

References

[1] A. Aker, A. Sliwa, F. Dalvi, and K. Bontcheva, "Rumour verification through recurring information and an inner-attention mechanism", *Online Social Networks and Media*, vol.13, pp.100045, 2019.

[2] J.P. Singh, A. Kumar, N.P. Rana, and Y.K. Dwivedi, "Attention-based LSTM network for rumor veracity estimation of tweet", *Information Systems Frontiers*, pp.1-16, 2020.

[3] A.R. Pathak, A. Mahajan, K. Singh, A. Patil, and A. Nair, "Analysis of techniques for rumor detection in social media", *Procedia Computer Science*, vol.167, pp.2286-2296, 2020.

[4] M. Guo, Z. Xu, L. Liu, M. Guo, and Y. Zhang, "An Adaptive deep transfer learning model for rumor detection without sufficient identified rumors", *Mathematical Problems in Engineering*, 2020.

[5] Q. Lv, Y. Wang, B. Zhang, and Q. Jin, "RV-ML: An effective rumor verification scheme based on multi-task learning model", IEEE Communications Letters, vol.24, no.11, pp.2527-2531, 2020.

[6] N. Bai, Z. Wang, and F. Meng, "A Stochastic Attention CNN Model for Rumor Stance Classification", *IEEE Access*, vol.8, pp.80771-80778, 2020.

[7] A. Kumar, "Rumour Stance Classification using A Hybrid of Capsule Network and Multi-Layer Perceptron", Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol.12, no.13, pp.4110-4120, 2021.

[8] R. kumari Mukiri, and B.V. Babu, "Prediction of rumour source identification through spam detection on social Networks-A survey", *Materials Today: Proceedings*, 2021.

[9] A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, and I. Augenstein, "Discourse-aware rumour stance classification in social media using sequential classifiers", Information Processing & Management, vol.54, no.2, pp.273-290, 2018.

[10] A. Bondielli, and F. Marcelloni, "A survey on fake news and rumour detection techniques", Information Sciences, vol.497, pp.38-55, 2019.

[11] J.C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection", IEEE Intelligent Systems, vol.34, no.2, pp.76-81, 2019.

[12] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection", Expert Systems with Applications, vol.128, pp.201-213, 2019.

[13] R.W. Filice, "Deep-learning language-modeling approach for automated, personalized, and iterative radiology-pathology correlation", Journal of the American College of Radiology, vol.16, no.9, pp.1286-1291, 2019.

[14] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, and C. Zhang, "SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model", IEEE Access, vol.8, pp.10896-10906, 2020.

[15] H. Jwa, D. Oh, K. Park, J.M. Kang, and H. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (Bert)", Applied Sciences, vol.9, no.19, pp.4062, 2019.

[16] Stephen Merity, Nitish Shirish Keskar, Richard Socher, 2017. "Regularizing and Optimizing LSTM Language Models", Cornell University, https://doi.org/10.48550/arXiv.1708.02182.

[17] H. Allcott, and M. Gentzkow, "Social media and fake news in the 2016 election", Journal of economic perspectives, vol.31, no.2, pp.211-36, 2017.

[18] R.K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet-a deep convolutional neural network for fake news detection", Cognitive Systems Research, vol.61, pp.32-44, 2020.

[19] H. Yuan, J. Zheng, Q. Ye, Y. Qian, and Y. Zhang, "Improving fake news detection with domain-adversarial and graph-attention neural network", Decision Support Systems, vol.151, pp.113633, 2021.

[20] J.A. Nasir, O.S. Khan, and I. Varlamis, “Fake news detection: A hybrid CNN-RNN based deep learning approach”, International Journal of Information Management Data Insights, vol.1, no.1, pp.100007, 2021.
[21] S.R. Sahoo, and B.B. Gupta, “Multiple features based approach for automatic fake news detection on social networks using deep learning”, Applied Soft Computing, vol.100, pp.106983, 2021.
[22] A. Choudhary, and A. Arora, “Linguistic Feature Based Learning Model for Fake News Detection and Classification”, Expert Systems with Applications, 2020.
[23] S. Sheikhi, “An effective fake news detection method using WOA-xgbTree algorithm and content-based features”, Applied Soft Computing, pp.107559, 2021.
[24] Y.F. Huang, and P.H. Chen, “Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms”, Expert Systems with Applications, vol.159, p.113584, 2020.
[25] F.A. Ozbay, and B. Alatas, “Fake news detection within online social media using supervised artificial intelligence algorithms”, Physica A: Statistical Mechanics and its Applications, vol.540, pp.123174, 2020.
[26] P. Bahad, P. Saxena, and R. Kamal, “Fake news detection using bi-directional LSTM-recurrent neural network”, Procedia Computer Science, vol.165, pp.74-82, 2019.
[27] A. Kumar, M.P.S. Bhatia, and S.R. Sangwan, “Rumour detection using deep learning and filter-wrapper feature selection in benchmark twitter dataset”, Multimedia Tools and Applications, pp.1-18, 2021.
[28] A. Zubiaga, M. Liakata, and R. Procter, “Exploiting context for rumour detection in social media”, In International Conference on Social Informatics, Springer, Cham, pp.109-123, 2017.
[29] E. Kochkina, M. Liakata, and I. Augenstein, “Turing at semeval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM”, arXiv preprint arXiv:1704.07221, 2017.
[30] Wang, Y., Wang, L., Yang, Y. and Zhang, Y., 2022. Detecting fake news by enhanced text representation with multi-EDU-structure awareness. arXiv preprint arXiv:2205.15139.
[31] Zhou, Y., Ying, Q., Qian, Z., Li, S. and Zhang, X., 2022. Multimodal Fake News Detection via CLIP-Guided Learning. arXiv preprint arXiv:2205.14304.
[32] Ali M. Meligy, Hani M. Ibrahim, Mohamed F. Torky, "Identity Verification Mechanism for Detecting Fake Profiles in Online Social Networks", International Journal of Computer Network and Information Security(IJCNIS), Vol.9, No.1, pp.31-39, 2017. DOI:10.5815/ijcnis.2017.01.04
[33] Shubham Bauskar, Vijay Badole, Prajal Jain, Meenu Chawla, "Natural Language Processing based Hybrid Model for Detecting Fake News Using Content-Based Features and Social Features", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.4, pp. 1-10, 2019. DOI:10.5815/ijieeb.2019.04.01
[34] Momina Shaheen, Shahid M. Awan, Nisar Hussain, Zaheer A. Gondal, "Sentiment Analysis on Mobile Phone Reviews Using Supervised Learning Techniques", International Journal of Modern Education and Computer Science(IJMECS), Vol.11, No.7, pp. 32-43, 2019. DOI:10.5815/ijmecs.2019.07.04
[35] Howard, J., & Ruder, S. (2018), “Universal language model fine-tuning for text classification”, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), https://doi.org/10.18653/v1/p18-1031.
[36] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network”, Physica D: Nonlinear Phenomena, vol.404, pp.132306, 2020.
[37] M. Sundermeyer, R. Schlüter, and H. Ney, “LSTM neural networks for language modeling”, In Thirteenth annual conference of the international speech communication association, 2012.
[38] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, “Analysing how people orient to and spread rumours in social media by looking at conversational threads”, PloS one, vol.11, no.3, pp.e0150989, 2016.
[39] M. Askarizadeh, B.T. Ladani, and M.H. Manshaei, “An evolutionary game model for analysis of rumor propagation and control in social networks”, Physica A: statistical mechanics and its applications, vol.523, pp.21-39, 2019.

Authors’ Profiles

S. Kanthi Kiran received M.Tech. degree in Computer Science and Engineering from Jawaharlal Nehru Technological University Hyderabad in 2008. He is pursuing Ph.D in JNTUK, Kakinada. Presently he is working as Associate Professor in department of Information Technology at Gayatri Vidya Parishad College of Engineering(A), Visakhapatnam, Andhra Pradesh, India. His research interests include big analytics, Machine Learning, Data Science. He published research papers in International Journals and conferences.

M. Shashi is a Professor and Chairperson of Board of Studies of the Department of Computer Science & Systems Engineering, A.U. College of Engineering(A), Andhra University, Visakhapatnam, Andhra Pradesh. She received the AICTE Career Award in 1996, Best Ph.D thesis prize from Andhra University in the year 1994 and AP State Best teacher award in 2016. 13 Ph.D.’s was awarded under her guidance. She co-authored more than 60 technical research papers in International Journals and 50 International Conferences and delivered many invited talks in such academic events. She is a member of IEEE Computational Intelligence group, Fellow of Institute of Engineers (India) and life member of Computer Society of India. Her current research interests include Data warehousing and Mining, Data Analytics, Artificial Intelligence, Soft Computing and Machine Learning.

K.B. Madhuri received M.Tech. degree in Computer Science and Technology from Andhra University in 1999. She obtained Ph.D from JNTU, Hyderabad in 2009. Presently she is working as Professor & Dean, School of CSE, IT & Computer Applications at Gayatri Vidya Parishad College of Engineering(A),Visakhapatnam, Andhra Pradesh, India. Her research interests include Data Mining, Pattern Recognition, Data warehousing and RDBMS. She is currently guiding two Ph.D scholars. She published research papers in National and International Journals. She is a member of IEEE and associate member of Institute of Engineers (India).

How to cite this paper: Sirra Kanthi Kiran, M. Shashi, K. B. Madhuri, "Multi-stage Transfer Learning for Fake News Detection Using AWD-LSTM Network", International Journal of Information Technology and Computer Science(IJITCS), Vol.14, No.5, pp. 58-69, 2022. DOI:10.5815/ijitcs.2022.05.05