Neural Computing and Applications (2022) 34:771–782
https://doi.org/10.1007/s00521-021-06450-4

ORIGINAL ARTICLE

AENeT: an attention-enabled neural architecture for fake news detection using contextual features

Vidit Jain[2] · Rohit Kumar Kaliyar[1] · Anurag Goswami[1] · Pratik Narang[2] · Yashvardhan Sharma[2]

Received: 5 March 2021 / Accepted: 17 August 2021 / Published online: 29 August 2021

Abstract
In the current era of social media, the popularity of smartphones and social media platforms has increased exponentially. Through these electronic media, fake news has been rising rapidly with the advent of new sources of information, which are highly unreliable. Checking off a particular news article is genuine or fake is not easy for any end user. Search engines like Google are also not capable of telling about the fakeness of any news article due to its restriction with limited query keywords. In this paper, our end goal is to design an efficient deep learning model to detect the degree of fakeness in a news statement. We propose a simple network architecture that combines the use of contextual embedding as word embedding and uses attention mechanisms with relevant metadata available. The efficacy and efficiency of our models are demonstrated on several real-world datasets. Our model achieved 46.36% accuracy on the LIAR dataset, which outperforms the current state of the art by 1.49%.

Keywords Fake News · Social Media · Contextualized Features · Deep learning · Neural Network

1 Introduction

Fake news has seen steep growth since the onset of the digital age. Every year, the ratio of Internet users [1] to the population of the world is increasing, currently standing at as high as 57%. With this increasing ratio, the reliability of any piece of news decreases, demeaning the credibility of the press and media. In the course of recent years, the popularity of smartphones and social networking websites has also increased at an exponential rate. Any end user can deceive hundreds of people in a short period and cause harm to individuals or society using the social media platforms available online. In recent times, fake news has also been responsible for increasing political polarization among different cohorts. A few examples include the controversy created during the 2017-Trump visit with pope Francis and the 2016 presidential election campaign in the USA [1–3].

Sharing and publishing counterfeit content over any social media platform are always questionable in terms of security perspective. Few well-known examples of fake news that had been trending during the 2016 US Presidential General Election [3] and COVID-19 (corona-virus) have been shown with the help of Fig. 1. Because of our inability in distinguishing real from false news, these news items lead to a negative impact [3, 5] in society. Maligned facts are mostly promoted to support a cause, as it happened in the UK, where the public was misinformed about the UK immigration policies to follow after Brexit to influence the referendum. They might also overshadow the more critical issues that must be known to the public [3].

The problem of fake news has been addressed quite extensively by researchers, which is not only challenging but also requires checking the facts [1]. Many datasets

✉ Pratik Narang
pratik.narang@pilani.bits-pilani.ac.in

Vidit Jain
f2016064@pilani.bits-pilani.ac.in

Rohit Kumar Kaliyar
rk5370@bennett.edu.in

Anurag Goswami
anurag.goswami@bennett.edu.in

Yashvardhan Sharma
yash@pilani.bits-pilani.ac.in

[1] Department of Computer Science Engineering, Bennett University, Greater Noida, India

[2] Department of CSIS, BITS, Pilani, Rajasthan, India

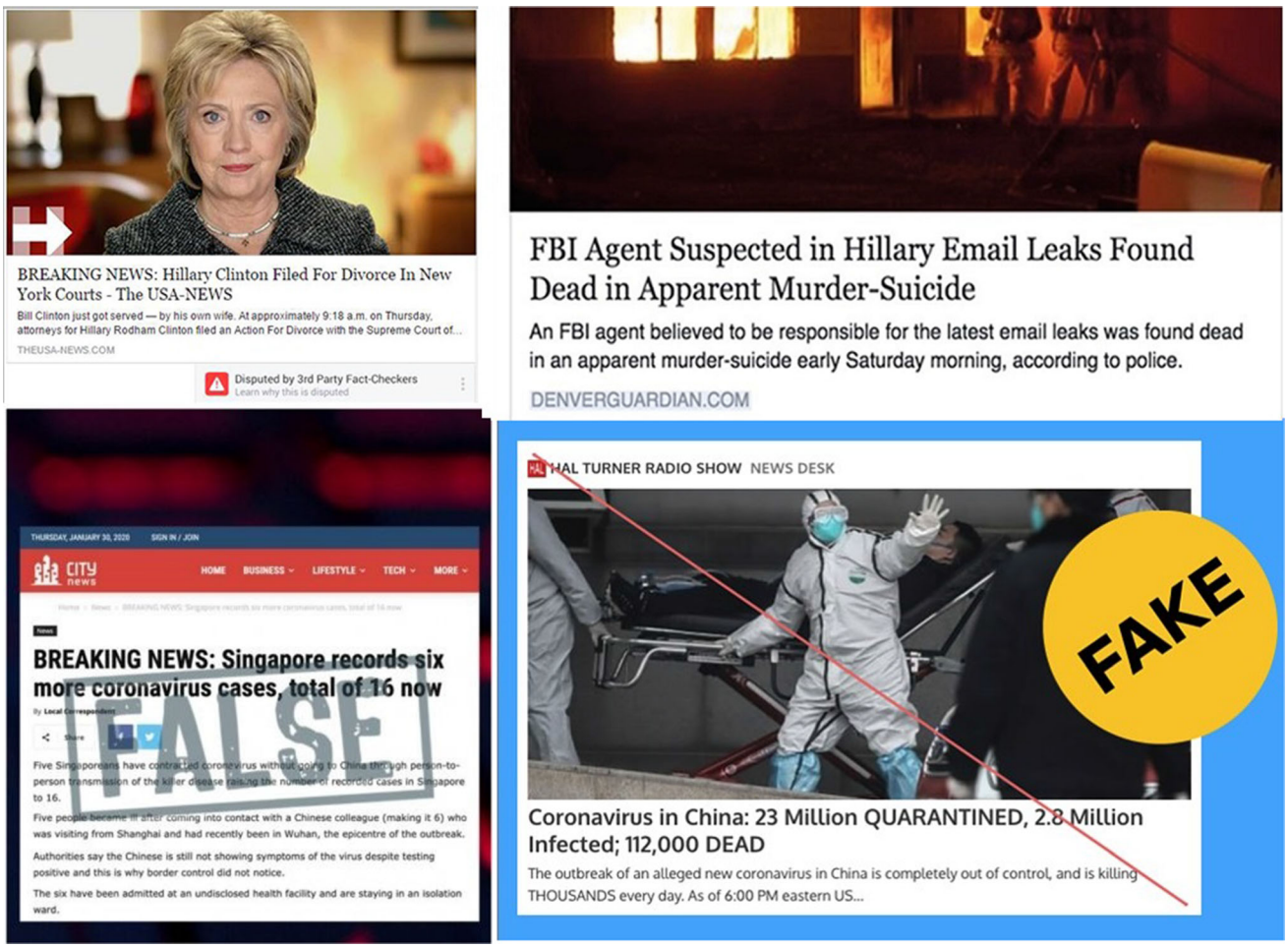

Fig. 1 Examples of some Fake News spread over social media (Source: Facebook and Twitter)

(e.g., LIAR [4], FakeNewsCorpus [2], Fake or Real news [6]) have been made available to the public for accurate fake news detection. In this paper, experiments have been conducted using the LIAR dataset, which contains news statements collected from debates, Facebook posts, etc., along with the information about the speaker having six class annotations that decide the degree of the fakeness of a statement where a lengthy analysis report grounds each judgment.

With recent advances [1, 3, 5, 26], fake news has been investigated by utilizing supervised and unsupervised methods [1, 7, 8, 27, 28]. Another notable model which is based on word vectors for pre-training the classification models is Global Vectors (GloVe) [11, 12]. These methods are widely used in the natural language processing (NLP) domain for classifying fake news [16]. In the NLP domain, for fake news detection, it has become imperative to develop deep neural architectures that are capable to learn hierarchical representations [9, 13, 14] of complete sentences.

1.1 Motivation and research goal

Fake news detection is one of the most active research areas with a focus of attention from various researchers across the world. A huge scope of improvement exists in the field of fake news detection due to insufficient context-specific news data for training. Employing deep learning techniques for detecting fake news gives a particular advantage over classical approaches because of its ability to engineer high-level features from the data. The above-mentioned issues motivate us to build an effective deep learning model for fake news detection.

Research Goal: Utilizing our proposed Neural Architecture using contextual features to improve fake news detection.

1.2 Our contribution

Deep neural architecture [2, 10, 33–35] has gathered huge attention from researchers due to its automatic feature extraction capability and demonstrated successful results

[8, 15, 31, 32]. The progress of NLP [2, 3, 13, 17] has motivated us to design of our proposed model for improving fake news detection. We propose a model architecture relying on the attention mechanisms to draw dependencies between metadata attributes and news statements and contextualized embedding to capture the semantic meaning of the sentence. The model is evaluated on the dataset provided by Wang [4]. The results of our experiment show that the use of contextual embedding and incorporating attention improves performance on the dataset. Our accuracy reaches 46.36% on the LIAR dataset, which is 1.49% higher than the current state-of-the-art methods. Our proposed model has shown an improvement in the classification results on real-world fake news datasets, which makes it effective in detecting fake news accurately.

The research paper is organized in the following manner. Section 2 describes the many existing approaches used in the field of fake news detection. Section 3 introduces the model architecture used. Section 4 specifies the experiment design and training details. Section 5 shows the comparison results with our proposed model. Section 6 evaluates the proposed model and presents experimental results with analysis. Section 7 shows the conclusion and future directions.

2 Related work

The goal of detecting fake news and classifying news statements into one of the six classes denoting the degree of fakeness has mostly been done using a word embedding such as Google word2Vec [8] or GloVe embedding [11] to capture the semantic meanings of words followed by adding metadata information to the model by either concatenating their fixed representations to the sentence or using attention mechanisms. However, in this model, instead of just capturing the meaning of the news statements at the word level, contextualized word representation [18] was used to model complex characteristics of word use across various linguistic contexts.

In one of the investigation, the authors [6] proposed a hybrid LSTM model in their research, where the LSTM layer is used for obtaining the representation of news articles on which attention is applied. Two attention factors are constructed; one used only the speaker information, whereas the other used topic information of the news articles. These speaker profiles are passed through a second LSTM to obtain vector representation of speakers which are concatenated and fed through a dense layer to obtain a prediction.

In one of the research [19], the authors proposed a deep ensemble model that used bi-directional long short-term memory (Bi-LSTM) to capture sequential information and convolutional neural networks (CNN) to capture the hidden features efficiently. These combined representations are used to capture the relations among the various attributes, which are all merged by a six-neuron dense layer at the end.

In one of the investigation [37], the authors proposed a novel attention mechanism for effective fake news detection. In their model, the authors have utilized an active learning framework to enhance learning performance in the case of labeled data. The authors have used two real-world datasets to validate the results. In another similar research [38], the authors proposed a novel knowledge-driven graph-CNN for fake news detection. The authors have utilized the concept of jointly modeling the different features (visual, textual, knowledge-based features) into a unified framework. The authors have used two real-world social media datasets, PHEME and WEIBO, for their investigations. They have achieved improved classification results. In one of the exploration [39], the authors proposed a deep face clustering method using residual graph convolutional network. In their network, the authors considered more hidden layers. For each node in the network, the authors used k-nearest neighbor algorithm to construct its sub-graphs. The authors have achieved more efficient and better clustering results in the experiments. Authors utilized different real-world datasets to validate the results.

However, in this model, instead of just capturing the meaning of the news statements at the word level, contextualized word representation [11] was also used to model complex characteristics. Attention mechanisms are used to emphasize important words in the sentence by computing a representation of the sequence using some additional information such as metadata or the sentence itself, in which case, it is referred to as self-attention. These have been very useful in tasks like question answering system, text summarization, etc. In the following sections, we will describe contextualized embedding, attention function, and network architecture.

Existing state-of-the-art work is given in Table 1. Most published works treat fake news detection as a binary

Table 1 Existing results using LIAR dataset

Authors	Models	Accuracy(%)
William Yang Wang [4]	Hybrid CNN	27.40
Long et al. [6]	Hybrid LSTM	41.50
Bi-LSTM Model	Bi-LSTM	42.65
CNN Model	CNN Model	42.89
Roy et al. [19]	A Deep Ensemble Model	44.87

classification problem, but it is vital to know how true a statement is. William Y Wang proposed a convolutional neural network approach for the six-way classification and achieved an accuracy of 0.27. Long et al. [6] proposed a hybrid attention-based LSTM model, which bettered the results by 14.5%. Roy et al. [19] established a new state of the art accuracy of 44.87% by using deep learning ensemble architecture based on CNN and Bi-LSTM to capture hidden features and information in both directions, respectively.

3 Model architecture

In order to proceed, we must formally define the notations required for understanding the architecture.

Given n news items $X = \{x_1, x_2, ..., x_n\}$, our model has to predict the degree of fakeness of each news item, denoted by $Y = \{y_1, y_2, ..., y_n\}$, $y_i \in Y$ where $Y = \{$pants fire, false, barely true, half true, mostly true, true$\}$. The true labels $l = \{l_1, l_2, ..., l_n\}$, $l_i \in Y$ are given at the time of training.

3.1 Pre-trained word embedding

Pre-trained word embedding is an essential step [12, 13, 20, 21] obtained by training a model in an unsupervised fashion on a context-specific dataset. Figure 2 shows the current state-of-the-art word embedding models for pre-training. These models are capable enough for pre-training with large-sized context-related datasets. Logically, it represents geometrical encoding [19] of words based on their frequency in the text corpus. Training and classification time are reduced to a great extent using these pre-trained models. They can also be categorized as both context-free and contextual-based. We can further divide the contextual-based models as unidirectional, as well as bidirectional [22–24], for pre-training.

3.2 Contextualized embedding

Contextualized embedding models typically consist of multiple stacked layers of representations (e.g., recurrent layers or transformer outputs). More context is incorporated with increasing layers added to the network. We can divide the contextual embedding (refer Fig. 4) into two forms (unidirectional and bi-directional).

3.2.1 ELMo

Contextualized embedding like ELMo (Embedding from Language Models) is used to learn context-dependent word embedding [18] from character level embedding (refer Fig. 3 for more details). ELMo has proven itself in pre-training in natural language processing by outperforming existing word embedding techniques. ELMo is trained on a massive dataset in the language of our dataset and then used as a component in other models to accomplish tasks in that particular language. Given a sequence consisting of N tokens, in our case, the news statement $s_i = (s_1, s_2, ..., s_N)$, a context-independent token representation x_{ij}, is computed using convolution over characters.

The sequence is passed through L layers of forward and backward LSTMs which output a context-dependent representation at each level. For the forward LSTMs, the model is trained so as to maximize the probability of token s_i given tokens $h_1, h_2, ..., h_{i-1}$ at the topmost layer. Similarly, the probability of token s_i is maximized with tokens $h_{i+1}, h_{i+2}, ..., h_N$ at the topmost layer in case of backward LSTMs. Since i^{th} layer tokens are dependent on $(i-1)^{th}$ layer tokens, all L layers end up getting trained.

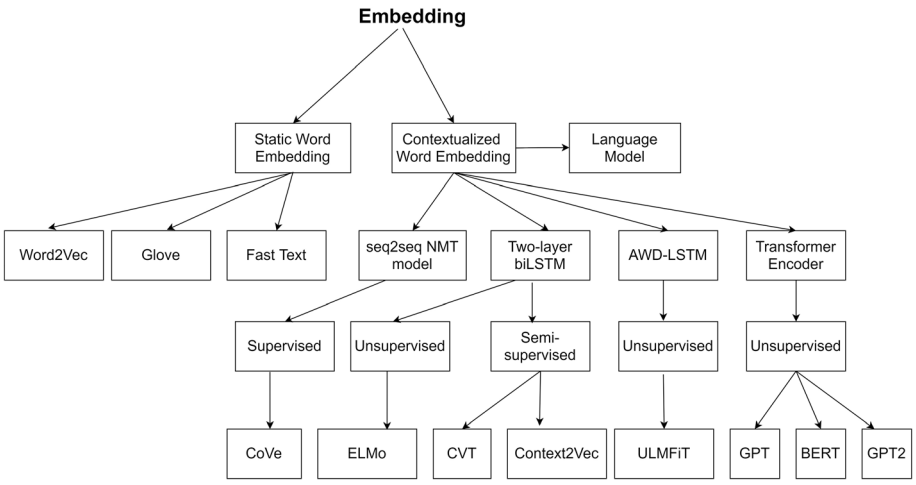

Fig. 2 Word Embedding State-of-the-Art Models

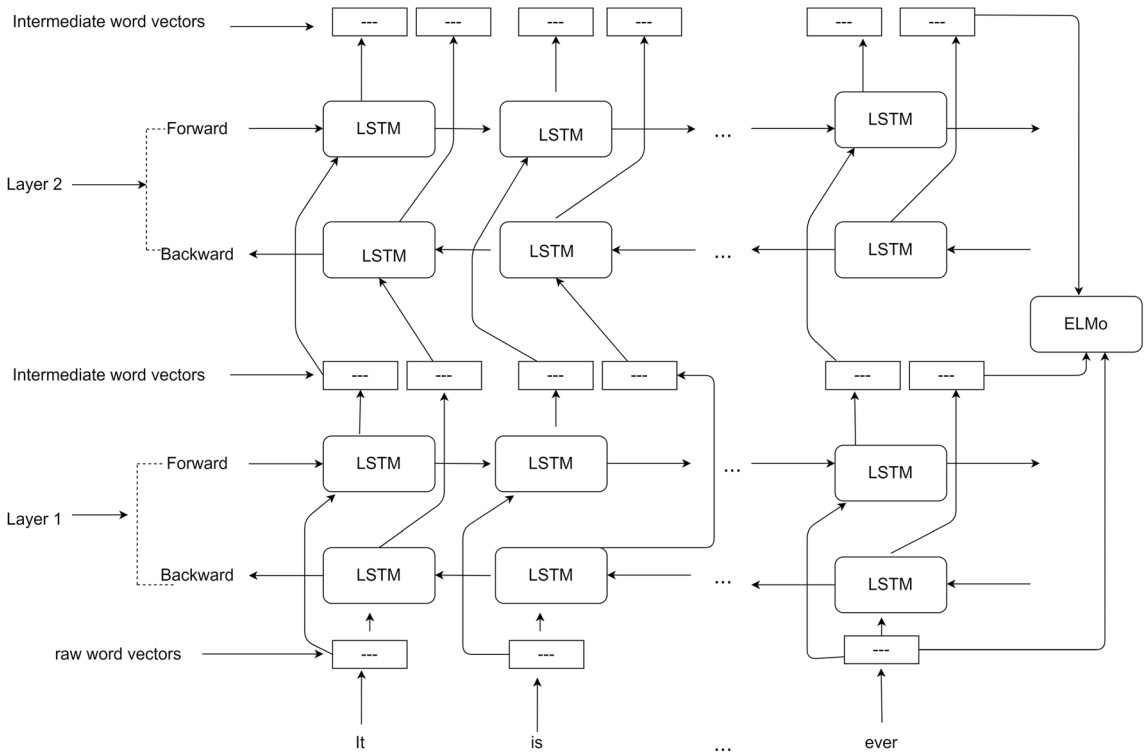

Fig. 3 Contextualized Embedding Model Architecture

ELMO(*s*) collapses the output representations at each layer along with the initial context-independent representations into a single vector by either their concatenation or softmax-normalized averaging of vectors.

3.2.2 GloVe

GloVe is a static word embedding technique that combines two of the most successful techniques in capturing word information [36]. First is the Continuous Bag of Words model, which is trained by predicting the center word in a particular window size. This approach is scalable with corpus size because of its window size limitation. The other model is the Count-based Co-occurrence matrix method, which tries to capture the global statistics of the words in the corpus but is high dimensional because the word embedding is directly proportional to the vocabulary size. GloVe combines the loss functions of both approaches in such a way that it outperforms both approaches. These word vectors encode similarity between two words, the similarity being their distance in n-dimensional vector space.

3.3 Attention mechanism

The attention mechanism used here is "Scaled Dot-Product Attention" [25]. The inputs consist of queries Q and key-value pairs K and V. Q is calculated by adding a fully connected layer (in our case) or a convolutional layer on top of the fixed vector with whom attention is to be taken. If the number of tokens is variable, mean pooling is done over sequence length to get a fixed vector. Then, attention is calculated by computing the dot product of the query with the key followed by the softmax function to obtain weights on the values.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

In our case, queries are the different metadata information compressed into a fixed vector, whereas the keys and values are the same news statement representations. Dot-Product Attention is space-efficient and much faster than other attention mechanisms since they all mostly work in conjunction with recurrent networks.

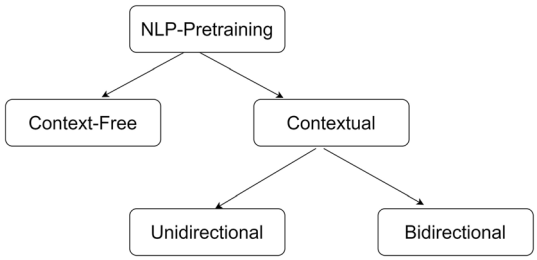

Fig. 4 Natural Language Models for Pre-training

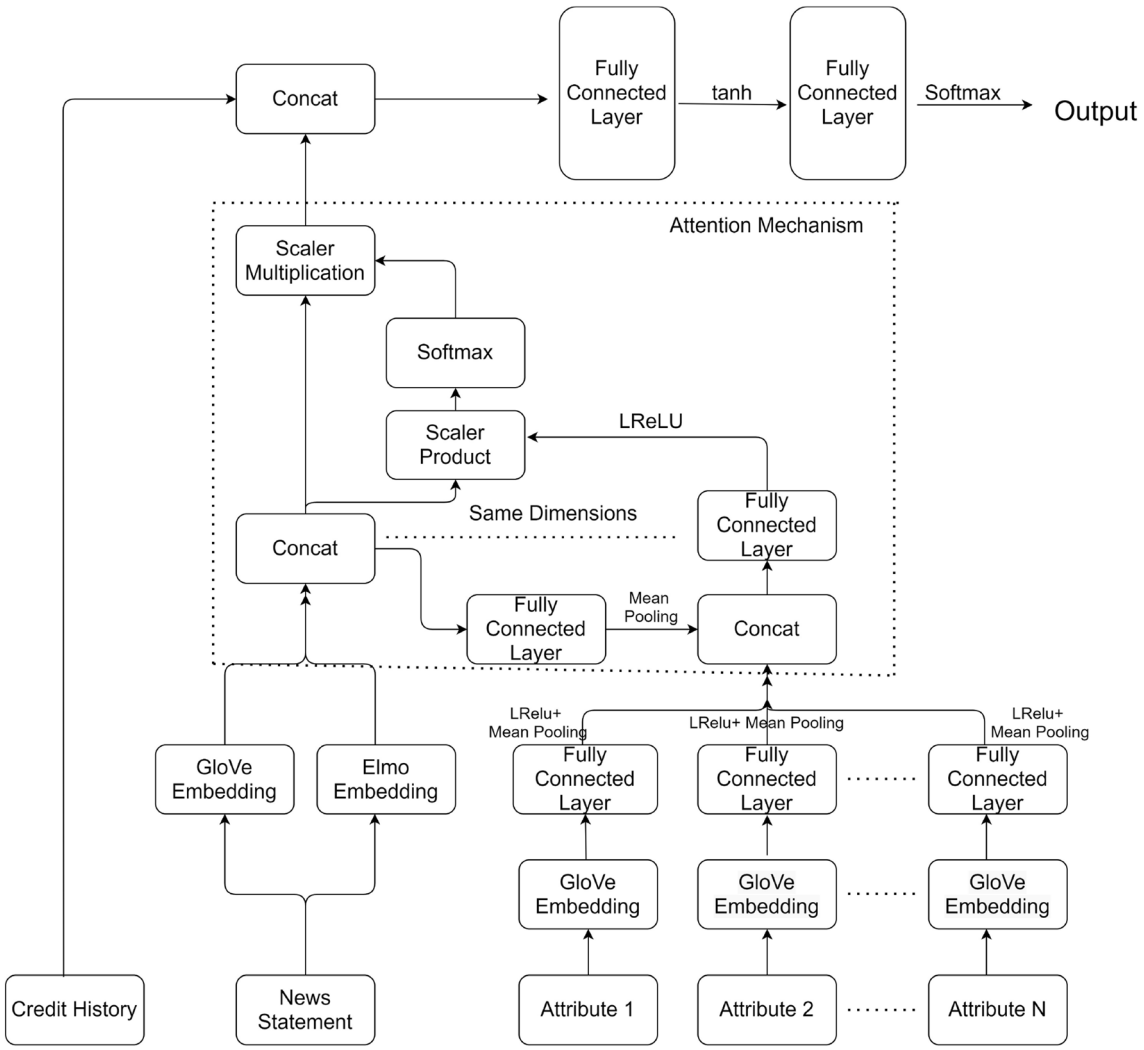

Fig. 5 Proposed model

3.4 Proposed model

We propose a model architecture for classifying a news statement into one of the six classes (false, half-true, true, barely true, mostly true, pants-fire) of fake news. In this research, we have used a combination of word embedding and sentence embedding because of the sequential nature of the data. The architecture of our proposed model is shown with the help of Fig. 5. This architecture allows us to utilize the semantic meaning of the words obtained from word embedding and context-dependent meaning obtained from contextualized word embedding.

3.4.1 Pre-processing

Before transforming the input into vectors, certain standardized NLP pre-processing techniques had to be used on the news statement, to reduce the noise of data and make the model robust. This involved lowercasing the sentences, tokenizing them and performing lemmatization using spacy for efficient GloVe embedding lookup. Stop words were also removed from the sentences to give more focus to words that can contribute toward better news classification.

3.4.2 Embeddings

Deep contextualized word representations have been proven effective in various natural language processing tasks to provide embedding as a function of the whole sentence's character embedding, whereas GloVe representations assign a unique vector for each word that captures the semantic meaning of that word. The Elmo embedding is obtained from a three-layer trainable bi-directional LSTM network with dropout in between layers. This network is run on pre-trained character embedding with the output of each layer being concatenated to get the final Elmo

embedding of the sentence. In the case of GloVe, the pre-trained embedding (trained on 6 billion words) was directly used. So, the news statement is embedded using both Elmo embedding and 300-dimensional GloVe embedding, and the results are concatenated to form (Elmo-size+300)-dimensional embedding.

The vectors of all attributes are embedded using 300-dimensional GloVe representations only and not Elmo embedding since these attributes are very short in nature and not actual sentences. Hence, using contextualized embedding is not propitious for attributes. These vectors are then passed through a fully connected layer to get the same dimensionality as the sentence vector for attention, followed by mean pooling over the sequence length to get a fixed-length vector. Mean pooling was preferred over LSTMs to remove the variability of the attributes because they are computationally less expensive and more efficient in this case. Also, the length of the metadata is not very long, making Mean Pooling a better method since there are no long dependencies to capture. In the case of multiple attributes, instead of passing through a fully connected layer of dimensionality (Elmo-size+300) neurons, all attributes are passed through a dense layer of 100 neurons. Then, these attribute vectors are made of constant dimensionality by the Mean Pooling layer of sequence length wherever necessary. These constant length vectors are concatenated and finally passed through a fully connected layer of (Elmo-size+300) neurons to get the same dimensionality as the sentence vector obtained. Leaky ReLU is used as activation functions after fully connected layers because of its faster convergence and not suffering from vanishing gradients in the future.

3.4.3 Attention

The method of "Scaled Dot-Product Attention" [25] is mostly preferred over other attention mechanisms because of its time and space efficiency. However, in the case of higher dimensionality, additive attention performs better where a one-layer feed-forward network is used to calculate the attention alignment. In our model, we decided to leverage the benefits of both the attention mechanisms, using the feed-forward layer on attributes and doing a scalar product instead of the usual dot product with the sentence vector $(s_k = w_{k1}w_{k2}...w_{kN})$ to get attention scores at the feature level instead of word level. In this mechanism, the complexity of the algorithm remains the same as the two models, and under-fitting is avoided despite the higher dimensionality of GloVe vectors being used. A softmax layer is applied to the calculated attention alignment, and a dot product is taken of this output with the sentence vector to get the final vector with attention. Attention weights are calculated for every word in the news sentence in the following manner.

$$\alpha_i = \frac{exp((W_x X + b_x).s_i)}{\sum_{j=1}^{N} exp((W_x X + b_x).s_j)} \tag{2}$$

$$s_i = \alpha_i.s_i \tag{3}$$

where X is the attribute vector with whom attention is to be applied, W_x is a $300 \times (elmo - size + 300)$ weight matrix and b_x is a $(elmo - size + 300)$-dimensional bias. This is the most important step in the model is all the metadata information decides the importance of each and every word in the statement to the meaning of the news.

3.4.4 Final output

After applying attention to attributes to the sentence, which helped in assigning higher weights to all the relevant information, the credit history of the speaker is concatenated to the highly important sentence vector, as it contributes the most toward the accuracy of the model and is the most significant indicator of the credibility of the speaker and, hence, the news item. This vector is passed through two fully connected layers, first having 300 neurons and tanh activation function applied to it, followed by a six neuron layer and a softmax function to calculate the probabilities of each category for the given news statement.

4 Experiment design and training

We trained on the standard LIAR dataset consisting of 10240 rows of news statement along with their metadata and labels embedded using deep contextualized embedding and GloVe embedding consisting of 6 billion words mapped to their corresponding 300-dimensional vector. The model was trained on NVIDIA P100 GPU. The model was trained for a total of 3000 steps or half an hour with a batch size of 16.

4.1 Dataset description

In this research, we have conducted several experiments using context-related fake news dataset[1]. It consists of three files (i) train.tsv (ii) test.tsv: It contains 1267 news statements along with their metadata, and (iii) validation.csv. This dataset includes 12.8K human-labeled short statements from PolitiFact.com's. In this dataset, (LIAR) six class labels exist for the truthfulness ratings: mostly true, barely true, pants-fire, false,half-true, and true. From

[1] The dataset can be downloaded from https://www.cs.ucsb.edu/william/data/liar_dataset.zip.

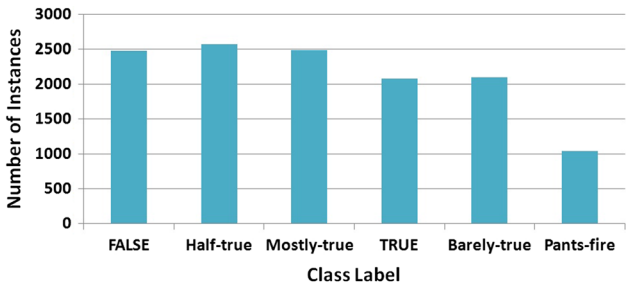

Fig. 6 Distribution of classes of LIAR dataset (Complete dataset-12791 Instances)

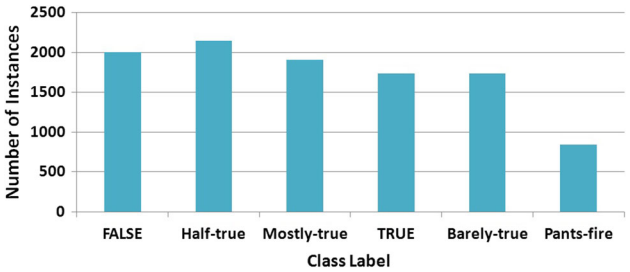

Fig. 7 Distribution of classes of LIAR dataset (Training samples-10240 Instances)

the perspective of class distribution, the LIAR dataset is well balanced: except for 839 pants-fire cases. Distribution of all separate classes, as well as a single dataset, is shown with the help of Figs. 6 and 7. An increase in the number of features can be seen after pre-processing (refer to Table 4) since we have added GloVe vectors and the length of sentences (statements, subject, venue, and job) in the input itself in order to avoid re-computation of GloVe embedding sentences for each iteration.

4.2 Optimizer

We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate varied according to the following formula:

$$learningrate = max(0.9^{epochnumber} \times 0.001, 5 \times 10^{-5}) \tag{4}$$

This corresponds to the exponential decay of 0.9 every epoch with an initial learning rate of 0.001 and a saturation learning rate 5×10^{-5}.

4.3 Regularization

Two types of regularization were used during training: 1) Residual Dropout of $P_{drop} = 0.1$ is applied in between the forward and backward LSTM layers of ELMo for efficient training. 2) L2 Regularization is applied with coefficient $\alpha = 3 \times 10^{-7}$ to avoid over-fitting.

5 Comparative analysis with existing results

Since last few years, fake news detection has become an active research topic among researchers (Table 1 for detailed description). Detailed analysis of our proposed model is given in Table 3 with the existing state-of-the-art methods using the LIAR dataset. Authors [19] have used a combination of the LSTM and CNN model and reported an accuracy of 44.87% using the LIAR dataset. Before this hybrid model, the best classification results were tabulated with an accuracy of 42.89 %. Several experiments have been conducted to evaluate the performance of our proposed model. Using our Elmo-enabled deep learning model, we achieved an accuracy of 46.36%. We motivate the researchers to use our model for future research in the area of fake news detection.

6 Analysis and results

In this section, detailed description of experiments results is explained. The selection of optimal hyperparameters is also explained in this section.

6.1 Hyperparameter tuning

In our research, we set the values of hyperparameters in such a way that we should get optimal results. In deep learning, the main functionality of selecting hyperparameters is related to use optimal memory and less cost of execution. In our approach, the selection of optimal hyperparameters is shown with the help of Table 2. The selection of optimal parameters is different for each classification task, as well as a context-dependent dataset. For

Table 2 Hyperparameters for our Proposed Model

Hyperparameter	Description or Value
Batch Size	16
Emb_dim	300
max_sub_len	1000
max_sub_len	1000
max_job_len	1000
max_job_len	1000
Learning Rate	0.001
Number of epochs	60
Optimizer	Adam
Attention-size	100
Elmo-size	1024
LSTM_output-size	64

Table 3 Existing classification results with our proposed model using LIAR dataset

Authors	Models	Accuracy(%)
William Yang Wang [4]	Hybrid CNN	27.40
Long et al. [6]	Hybrid LSTM	41.50
Bi-LSTM Model	Bi-LSTM	42.65
CNN Model	CNN Model	42.89
Roy et al. [19]	A Deep Ensemble Model	44.87
Our Proposed model	ELMo-enabled Attention-based Model	46.36%

Table 4 Shape of Input Data before and after pre-processing

Data	Shape
Training Data	10240x14
Testing Data	1267x14
Training Data after pre-processing	10240x22
Testing Data after pre-processing	1267x22

selecting optimal hyperparameters, there exist two basic selection techniques: manual and automatic selection. In the case of automatic selection, a high computational cost is required, while a deep understanding of the classification model is required for manual selection.

6.2 Performance parameters

Evaluations are performed on the test set of the LIAR dataset. The labels take discrete values from 0 to 5 corresponding to pants-fire, false, barely true, half-true, mostly true, and true. Classification results are outlined in Table 7. The results show that the speaker profile information improves fake news detection significantly. Apart from the speaker's credit history, which gives an improvement of 11% as should have been the case, the five major attributes are the speaker's job, subject of news statement, venue of the speech, the party of the speaker and the state in which speech took place. The experiments were conducted with all attributes present in the dataset, but we did not see any improvement of accuracy from the model without attention for attributes other than the five listed in Table 7, and concluded that those attributes had no correlation with the degree of fakeness of the news. As evident by the results, applying self-attention or attention to the job provides the best improvement in the case of individual attributes. However, the best accuracy of 46.46% is seen when attention is applied to the concatenation of job, subject, venue, and the statement itself. Even without attention, our model outperforms Roy et al. [19] by nearly 1%.

To validate the performance of our model, different performance parameters (Precision, Recall, and F_1-Score) have been taken into consideration. From Table 6, we can observe the values of these performance parameters. Our proposed model has shown an improvement in the fake

Table 5 Representation of Confusion Matrix using our proposed model

	Predicted-1	Predicted-2	Predicted-3	Predicted-4	Predicted-5	Predicted-6
Actual-1	53	14	8	1	2	5
Actual-2	21	143	43	40	33	37
Actual-3	7	21	70	12	6	15
Actual-4	10	33	60	153	58	1
Actual-5	1	35	29	58	135	66
Actual-6	0	2	0	1	0	32

Table 6 Classification results using our proposed model

Class	n(truth)	n(classified)	Precision(%)	Recall(%)	F_1-Score(%)
1	92	83	63.85	57.61	60.57
2	248	317	45.11	57.66	50.62
3	210	131	53.44	33.33	41.05
4	265	374	48.57	57.74	52.76
5	241	324	41.67	57.69	48.39
6	208	35	91.43	20.51	33.50

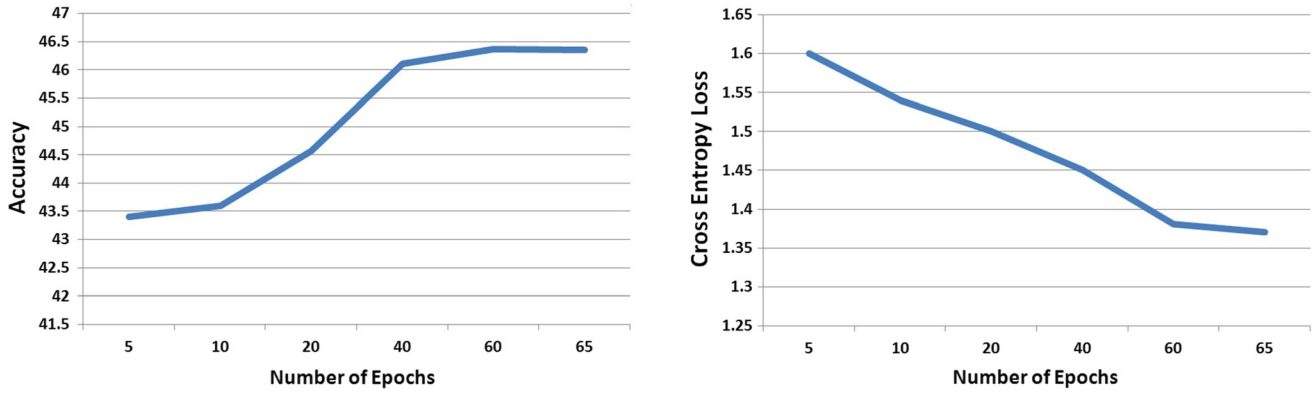

Fig. 8 Accuracy of our proposed model with Testing samples

Table 7 Evaluation Results

Models	Accuracy
William Yang Wang [4]- Hybrid CNN	0.2740
Long et al. [6]- Hybrid LSTM	0.4150
Roy et al. [19]- Deep ensemble model	0.4487
Elmo+No attention	0.4514
Elmo+GloVe+No attention	0.4565
GloVe+No attention	0.4525
Elmo+GloVe+Subject-attention	0.4573
Elmo+GloVe+Venue-attention	0.4517
Elmo+GloVe+State-attention	0.4509
Elmo+GloVe+Party-attention	0.4509
Elmo+GloVe+Self-attention	0.4597
Elmo+GloVe+Job-attention	0.4612
Elmo+GloVe+(Job+Subject+Venue+Self)-attention(our proposed model)	**0.4636**

news classification. We have achieved an accuracy of 46.36% which is 1.49% higher than the state-of-the-art models. In our investigation and analysis (refer Fig. 8 for more details), the cross-entropy loss decays rapidly with our attention-enabled deep learning-based model in comparison with the standard embedding-layer-based model with testing samples. Cross-entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1. So predicting a probability of .015 when the actual observation label is 1 would be wrong and result in a high loss value in the prediction.

6.3 Results

It has been shown that the best accuracy is achieved via our proposed model. From Table 7, we can observe the values of different performance parameters (Precision, Recall, and F_1-Score). These results validate the performance of our proposed model with other classification models. From the above analysis (refer Fig. 8), it can be seen that the cross-entropy loss decays rapidly with the attention-enabled model in comparison with the standard embedding-layer-based model. Cross-entropy loss measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probabilities diverge from the actual label. So predicting a probability of .015 when the actual observation label is 1 would be wrong and result in a high loss value in the prediction. The training loss for pre-trained embedding-based models decays relatively fast and without any fluctuations. Cross-entropy loss reduces significantly using our model, and it achieved the highest accuracy in comparison with traditional learning-based models, as well as other deep learning-based models, with minimal losses.

7 Conclusion

In this research, we propose an attention-based model for fake news detection. To compute the attention weights, the main attributes of the speaker were used. These attributes are the job of the speaker, the subject of the speech, etc. An

intermediate layer is added with the credit history. These techniques contribute to the improvement of the model. Our model was able to reach an accuracy of 46.36%, which outperforms the state-of-the-art model by 1.49%. In the future, we plan to use multi-model based approaches for fake news detection with BERT-Score also. Our further plan would be to utilize a hybrid approach (using content, context, and graph-based information of news) for classification.

In future, we can explore multiple parallel channel-based deep neural networks with different kernel sizes. These neural networks would be a milestone to learn from different feature vectors of word length for more accurate classification. Beyond the text information, visual information-based analysis can be more helpful to build a real-time detection system for video and image investigation. Future directions to explore the knowledge of domain experts and fact-check techniques are promising.

Declarations

Conflicts of interest All the authors of this manuscript certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

1. Srijan K, Neil S (2018) False information on web and social media: a survey
2. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newslett 19(1):22–36
3. Persily N (2017) The 2016 US election: Can democracy survive the internet?. J Democracy 28(2):63–76
4. Wang WY (2017) Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 2: Short Papers), pp. 422-426
5. Conroy NJ, Rubin VL, and Chen Y (2015) Automatic deception detection: methods for finding fake news. In: Proceedings of the 78th ASIS&T annual meeting: information science with impact: research in and for the community, pp. 1-4
6. Long, Yunfei, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. "Fake News Detection Through Multi-Perspective Speaker Profiles." In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 252-256. 2017
7. Dougherty J, Ron K, and Mehran S (1995) Supervised and unsupervised discretization of continuous features. In: Machine learning proceedings 1995, pp. 194-202. Morgan Kaufmann
8. Mikolov T, Kai C, Greg C and Jeffrey D (2013) Efficient Estimation of Word Representations in Vector Space
9. Ahmed H, Issa T, and Sherif S (2017) Detection of online fake news using N-gram analysis and machine learning techniques. In: International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, pp. 127-138. Springer, Cham,
10. Vasudevan V, Barret Z, Jonathon S, and Le QV (2019) Neural architecture search for convolutional neural networks. U.S. Patent 10,521,729, issued December 31
11. Pennington J, Socher R, and Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543
12. Zhang Y, Wallace BC (2017) A sensitivity analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In: Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers), pp. 253-263
13. Caliskan Aylin, Bryson Joanna J, Narayanan Arvind (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186
14. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 1: Long Papers), pp. 1555-1565
15. Zhong Botao, Xing Xuejiao, Peter Love Xu, Wang, and Hanbin Luo. (2019) Convolutional neural network: Deep learning-based classification of building quality problems. Adv Eng Inform 40:46–57
16. Zhang T, Wang D, Chen H, Zeng Z, Guo W, Miao C, Cui L (2020) BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection. In: 2020 international joint conference on neural networks (IJCNN), pp. 1-8. IEEE
17. Shu K, Wang S, Liu H (2019) Beyond news contents: the role of social context for fake news detection. In: Proceedings of the Twelfth ACM international conference on web search and data mining, pp. 312-320. ACM
18. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, and Zettlemoyer L (2018) Dep contextualized word representations. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, Volume 1 (Long Papers), pp. 2227-2237
19. Roy A, Basak K, Ekbal A, Bhattacharyya P (2018) A deep ensemble framework for fake news detection and classification
20. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, pp. 849-857
21. Camacho-Collados J, Pilehvar MT, and Navigli R (2016) Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artif Intell 240:36-64
22. Iyyer M, Manjunatha V, Boyd-Graber J, and Hal Daumé III (2015) Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pp. 1681-1691
23. Kamkarhaghighi M, Makrehchi M (2017) Content tree word embedding for document representation. Expert Syst Appl 90:241–249

24. Cerisara C, Kral P, Lenc L (2018) On the effects of using word2vec representations in neural networks for dialogue act recognition. Comput Speech Language 47:175–193
25. Vaswani A, Noam S, Niki P, Jakob U, Llion J , Gomez AN, Lukasz K, and Illia P (2017) Attention is All you Need. InL NIPS
26. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y (2019) Combating fake news: a survey on identification and mitigation techniques. ACM Trans Intell Syst Technol (TIST) 10(3):1–42
27. Ruchansky N, Sungyong S, and Yan L (2017) Csi: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp. 797-806. ACM
28. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188
29. Feng S, Ritwik B, and Yejin C (2012) Syntactic stylometry for deception detection. In: Proceedings of the 50th annual meeting of the association for computational linguistics: Short Papers-Volume 2, pp. 171-175. Association for Computational Linguistics
30. PÉrez-Rosas V, Bennett K, Alexandra L, and Rada M (2018) Automatic Detection of Fake News. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3391-3401
31. Shu K, Limeng C, Suhang W, Dongwon L, and Huan L (2019) Defend: explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 395-405
32. Zhang J, Bowen D, and Yu Philip S (2020) Fakedetector: effective fake news detection with deep diffusive neural network. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1826-1829. IEEE
33. Rohit KK , Anurag Gi, Pratik N, and Soumendu S (2020) FNDNet–a deep convolutional neural network for fake news detection. Cognit Syst Res 61: 32–44
34. Kaliyar RK, Anurag G, and Pratik N (2021) EchoFakeD: improving fake news detection in social media with an efficient deep neural network. Neural Computing and Applications 1-17
35. Kaliyar RK, Anurag G, and Pratik N (2021) MCNNet: generalizing Fake News Detection with a Multichannel Convolutional Neural Network using a Novel COVID-19 Dataset. In: 8th ACM IKDD CODS and 26th COMAD, pp. 437-437
36. Pennington J, Richard S, and Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543
37. Ren Y, Bo W, Jiawei Z, and Yi C (2020) Adversarial active learning based heterogeneous graph neural network for fake news detection. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 452-461. IEEE,
38. Wang Y, Shengsheng Q, Jun H, Quan F, and Changsheng X (2020) Fake News Detection via Knowledge-driven Multimodal Graph Convolutional Networks. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 540-547
39. Qi C, Zhang J, Jia H, Mao Q, Wang L, Song H (2021) Deep face clustering using residual graph convolutional network. Knowl Based Syst 211:106561