Hindawi
Complexity
Volume 2021, Article ID 5557784, 8 pages
https://doi.org/10.1155/2021/5557784

Research Article

Fake Detect: A Deep Learning Ensemble Model for Fake News Detection

Nida Aslam, Irfan Ullah Khan, Farah Salem Alotaibi, Lama Abdulaziz Aldaej, and Asma Khaled Aldubaikil

Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

Correspondence should be addressed to Nida Aslam; naslam@iau.edu.sa

Received 19 February 2021; Revised 7 March 2021; Accepted 26 March 2021; Published 15 April 2021

Academic Editor: M. Irfan Uddin

Pervasive usage and the development of social media networks have provided the platform for the fake news to spread fast among people. Fake news often misleads people and creates wrong society perceptions. The spread of low-quality news in social media has negatively affected individuals and society. In this study, we proposed an ensemble-based deep learning model to classify news as fake or real using LIAR dataset. Due to the nature of the dataset attributes, two deep learning models were used. For the textual attribute "statement," Bi-LSTM-GRU-dense deep learning model was used, while for the remaining attributes, dense deep learning model was used. Experimental results showed that the proposed study achieved an accuracy of 0.898, recall of 0.916, precision of 0.913, and *F*-score of 0.914, respectively, using only statement attribute. Moreover, the outcome of the proposed models is remarkable when compared with that of the previous studies for fake news detection using LIAR dataset.

1. Introduction

Progression and advancement of the hand-held devices and high-speed Internet have exponentially increased the number of digital media users. According to digital global report 2020, the number of users for digital media reached 4.75 billion, and the social media users reached 301 million in 2020 [1]. This digitalization converts the world into the global village. Due to this advancement, individuals are just one click away from the information worldwide. Despite several advantages, this transformation has raised some challenges. Fake news is one of the challenges faced by the digital community nowadays.

Fake news is pervasive propaganda that spreads misinformation online, using social media like Facebook, twitter, and Snapchat to manipulate public perceptions. Social media can have two sides for news consumption, i.e., can be utilized to update the community about the latest news and, on the other hand, can be a source of spreading false news. However, social media is a low cost, quick access, and fast distribution of news and information and to know what is happening worldwide. Moreover, due to its simplicity and lack of control on the Internet, it allows "fake news" to be widespread.

Fake news has become a focal point of discussion in the media over the past three years due to its impact on the 2016 US Presidential election [2]. Reports showed that human's capability for detecting deception without special assistance is only 54% [3]. Therefore, there is a need for an automated way to classify fake and real news accurately. Some studies have been conducted but still there is a need for further attention and exploration. The proposed study attempts to eliminate the spread of rumors and fake news and helps people to identify the news source as trustworthy or not by automatically classifying the news.

The organization of this paper is as follows. Section 2 includes a review of previous studies. Section 3 explains the proposed methodology, which contains the "LIAR" dataset description, preprocessing, and classification models used. Section 5 includes experimental setup results and discussion. Finally, Section 6 contains the conclusion of this paper.

2. Related Studies

One of the earlier studies on fake news detection and automatic fact-checking with more than a thousand samples was done by [4] using LIAR dataset. The dataset contains 12.8 K human-labeled short statements from POLITIFACT.COM. The statements were labeled in six different categories, such as pants fire, false, barely true, half true, mostly true, and true. The study used several classifiers such as logistic regression (LR), support vector machine (SVM), a bidirectional long short-term memory (Bi-LSTM) networks model), and a convolutional neural network (CNN) model. For LR and SVM, the study used the LIBSHORTTEXT toolkit and showed significant performance on short text classification problems. The study compared several techniques using text features only and achieved an accuracy of 0.204 and 0.208 on the validation and test sets. Due to overfitting, the Bi-LSTMs did not show good performance. However, the CNN outperformed all models, resulting in an accuracy of 0.270 on the holdout data splitting.

Similarly, another study compared three datasets such as LIAR datasets, fake or real news dataset [5], and the dataset generated by collecting fake news and real news from Internet [6]. The study made a comparison among various conventional machine learning models such as SVM, LR, decision tree (DT), AdaBoost (AB), Naive Bayes (NB), and K nearest neighbor (KNN), respectively, using lexical, sentiment, unigram, and bigram techniques with term frequency and inverse document frequency (TF-IDF). Furthermore, several CNN models such as NN, CNN, LSTM, Bi-LSTM, hierarchical attention network (HAN), convolutional HAN, and character level C-LSTM were also used with Glove embedding and character embedding to train the model. They found that the performance of the LSTM model highly depends upon the size of the dataset. The result showed that NB, with n-gram (bigram TF-IDF), features produced the best outcome of approximately 0.94 accuracy with the combined corpus dataset.

Conversely, the study by [4] indicated that the CNN model outperformed the LIAR dataset. However, the study by [6] showed that the CNN model is the second-best for all the datasets. The NB model showed the best performance for the LIAR dataset with 0.60 accuracy and 0.59 F1-score. For the fake or real news, dataset Char-level C-LSTM showed the best performance with 0.95 accuracy and 0.95 F1-score. LSTM-based models showed the best outcome on the combined corpus dataset, where both Bi-LSTM and C-LSTM produced an accuracy of 0.95 and F1-score of 0.95.

Furthermore, another study was performed by Girgis et al. [3] regarding the spread of fake news and used recurrent neural network (RNN) models (Vanilla RNN, Gated Recurrent Unit (GRU)) and long short-term memories (LSTMs) on the LIAR dataset to predict fake news. They compared and analyzed their results with Wangs [4] findings. Although similar results were achieved, GRU (0.217) outperformed the other models. Nevertheless, in comparison with the findings of Wang, they found that CNN is better in terms of speed and outcomes. Similarly, the authors in [7] used the LSTM model on LIAR dataset. They found that adding the speaker profile enhances the performance of the algorithm. The model achieved an accuracy of 0.415.

Moreover, the study by [8] proposed a novel approach to overcome the problem of fake news detection using two metaheuristic algorithms, salp swarm optimization (SSO) and grey wolf optimization (GWO). The study performed experiments using three different datasets, which are BuzzFeed Political News, Random Political News, and LIAR Benchmark. The results showed that the GWO algorithm outperformed as compared with the SSO and other algorithms. GWO obtained the best accuracy in all datasets and produced highest precision and F1-score in two out of three datasets. Moreover, the precision of the SSO within two out of three datasets performed better than all the algorithms. The results obtained from the two algorithms were very promising because of the representation structure, and flexible fitness function handled many different objectives simultaneously and efficiently. The study recommended that using different similarity metrics in model construction and testing improves the performance of their model. In the converted document vector, binary versions of metaheuristic optimization techniques can also be used. Similarly, to improve the results of the study, adaptive and hybrid versions of the SSO and GWO algorithms were proposed.

Another study [9] used self multihead attention-based CNN (SMHACNN). The study implemented CNN and self multihead attention (SMHA) techniques and evaluated the truthfulness of news based on its content. The experiments were conducted on a public dataset that was collected from fakenews.mit.edu. The study conducted two experiments using 5-fold cross-validation, and their results showed that the model produced effective outcomes in detecting the fake news with the precision of 0.95 and the recall of 0.95. Besides, they have compared their results with previous work and have shown that their proposed technique using the self multihead attention with the CNN made a remarkable performance.

Additionally, the authors in [10] developed an exploratory analysis model using Facebook news during the 2016 US Presidential election based on the elaboration likelihood model as well as numerous cognitive and visual indicators of information, which most of them have already been shown to impact the quality of online information. The study investigated how news posts' cognitive, visual, affective, and behavioural clues, together with the addressed user communal, can be used by machine learning models to automatically detect the fake news. The study used a BuzzFeed dataset of Facebook posts. They trained many machine learning models appropriate for binary classification. The classifiers were LR, SVM, DT, random forest (RF), and extreme gradient boosting (XGB) and were trained with the same features set. The study achieved the highest accuracy of 0.80 and an approximately 0.90 recall.

A study used a hybrid approach by combining deep learning, natural language processing (NLP), and semantics using LIAR and PolitiFact datasets [11]. The study compared the performance of some classical machine learning models like multinomial Naïve Bayes (MNB), stochastic gradient boosting (SGD), LR, DT, and SVM. The study compared the performance of some classical machine learning models like multinomial Naïve Bayes (MNB), stochastic gradient boosting (SGD), LR, DT, SVM, and DL models like CNN, Basic LSTM, Bi-LSTM GRU, and CapsNet, respectively. The study found that CapsNet outperformed the other model with an accuracy of 0.649 using LIAR dataset. The integration of semantic features such as named entity recognition (NER) sentiments in LIAR dataset enhanced the performance of the classification model. Similarly, another study also compared the performance of machine learning and DL models and found similar performance of SVM and Bi-LSTM with an accuracy of 0.61 using LIAR dataset [12]. However, the training time of Bi-LSTM was very huge. Recently, the study used ensemble-based machine learning approach for the classification of fake news using two datasets LIAR and ISOT dataset [13]. The ensemble model used DT, RF, and extra tree classifiers. The study achieved testing accuracy of 44.15%.

Despite of several studies already made in the Fake news detection, there is still a room for further improvement and investigation. The studies mentioned above highlight the significance of the CNN and deep learning models for classification of fake news. It is also found that LIAR dataset is one of the widely used benchmark dataset for the detection of fake news. In our study, we attempt to develop an ensemble-based deep learning model for fake news classification that produced better outcome when compared with the previous studies using LIAR dataset.

3. Material and Methods

This section presents an overview of dataset, preprocessing techniques, and description of the deep learning model used for classification. Figure 1 represents the proposed study methodology. The dataset contains two types of feature such as short textual feature, i.e., statement and other features like speaker job title, subject, and venue. Therefore, the features were initially divided according to the category. For the statement attribute, several NLP techniques like tokenization, lemmatization, and stop word removal were used. However, for the other category of features, different data preprocessing techniques were applied that will be discussed further in the preprocessing section.

4. Dataset Description

The study used “LIAR” dataset [4] that contains 12.8 K human-labeled short statements from POLITIFACT.COM, and each statement is checked for its truthfulness by a POLITIFACT.COM editor. It has six categories for the label to rate accuracy, which are pants fire, false, mostly true, half true, mostly true, and true. The dates for the statements are primarily from 2007 to 2016. The speakers include a combination of democrats and republicans, and for each speaker, there is a rich collection of metadata that includes historical counts of false statements for each speaker. Such statements are sampled from different contexts/venues, and also the speakers are discussing a diverse set of subjects. Table 1 shows the description of the dataset. The statistical analysis of the historical counts of inaccurate statements for each speaker is also presented in the table. For the numeric variable mean (μ), standard deviation (σ) and range are used. However, categorical variable number of categories has been used. The table also contains the number of missing values per attribute. In the dataset, only three attributes have missing values, namely, speaker’s job title, state info, and the context. The study used the records with the class label true and false with the total number of records 4557. The number of news records with true class label is 2053 and with false class label is 2504, respectively.

4.1. Preprocessing. Several preprocessing techniques were applied on the dataset. Initially, the dataset consists of 14 attributes. Three attributes have missing values, namely, state_info, speaker job title, and venue. State info was removed from the study due to low relevance of the attribute. However, the other two attributes with missing values, namely, speaker’s_job title and venue were included for further analysis. In the speaker’s job title and venue, attribute missing values were replaced with the unique category unknown. Party affiliation feature consists of 24 categories and is converted into four categories, namely, republican, democrat, unknown, and other, respectively. The category none is replaced with unknown while all other 19 categories are replaced with other except republican and democrat. Normalization was performed on four columns, namely, barely true counts, false counts, half true counts, mostly true counts, and pants on fire counts, respectively. The data were normalized in the range (0-1).

$$X\prime = \frac{\left(x - \min_a\right)}{\max_a - \min_a}\left(\text{new}_{\max_a} - \text{new}_{\min_a}\right) + \text{new\_min}_a. \tag{1}$$

Figure 2 represents number of records per category for party affiliation attribute. Venue attribute was reduced to 8 categories, namely, interview (1686 records), other (766 records), ad (685 records), news (427 records), social media (316 records), website (72 records), unknown (39 records), and show (20 records), respectively. The distribution of categories for subject attribute is shown in Figure 3. Similarly, speaker job title attribute was converted into 9 categories such as unknown (1089 records), other (968 records), state representative (730 records), president (469 records), US representative records, media (159 records), government (157 records), company (58 records), and office director (31 records), respectively. Figure 4 represents number of records per category for subject attribute after reduction.

After performing all preprocessing steps with the data, the dataset contains 10 features and a target variable. One of the features in the dataset, namely, “statement” contains textual data.

For the statement attribute initially, wordnet tokenizer was applied. Similarly, for the lemmatization, we used WordNetLemmatizer. After the lemmatization stop words were removed, we used English stop words. The word cloud before and after preprocessing is shown in Figures 5 and 6 .

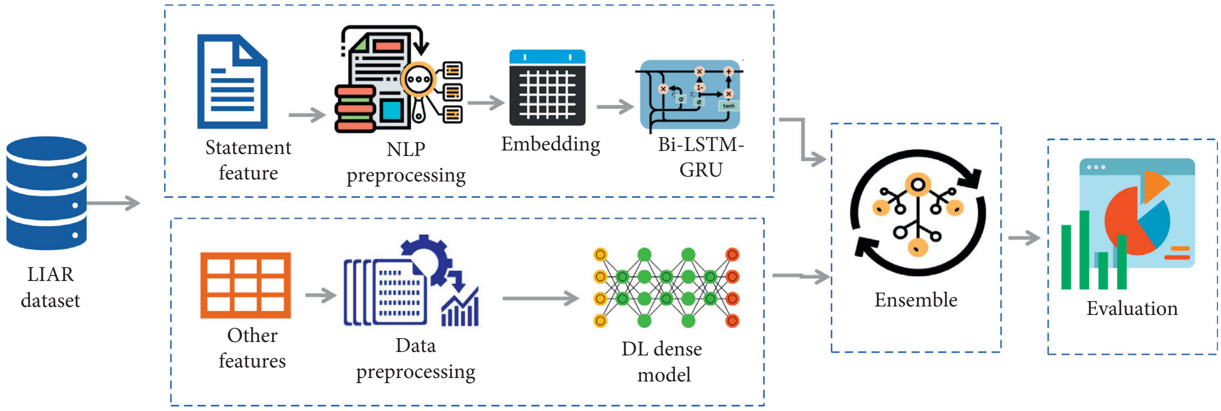


Figure 1: Block diagram of the proposed study methodology.

Table 1: Description of the LIAR dataset.

No.	Feature name	Datatype	Missing values	Mean (μ) ± Std (σ)	Range	No. of categories
1	ID of the statement	Object	—	—		
2	Label	Object	—	—		2
3	Statement	Object	—	—		4007
4	Subject(s)	Object	—	—		1823
5	Speaker	Object	—	—		9
6	Speaker’s job title	Object	1184	—		656
7	State info	Object	926	—		
8	Party affiliation	Object	—	—		4
9	Barely true counts	Int (64)	—	11.59 ± 18.98	0–70	
10	False counts	Int (64)	—	13.36 ± 24.14	0–114	
11	Half true counts	Int (64)	—	17.19 ± 35.85	0–160	
12	Mostly true counts	Int (64)	—	16.50 ± 36.17	0–163	
13	Pants on fire counts	Int (64)	—	6.25 ± 16.18	0–70	
14	The context (venue/location of the speech or statement)	Object	52			

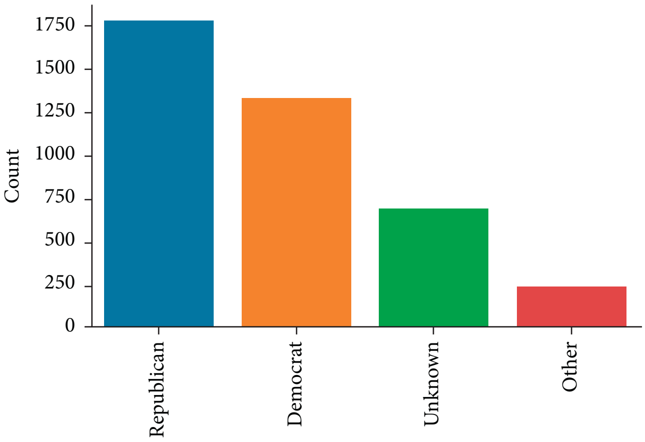


Figure 2: Number of records per category for party affiliation attribute.

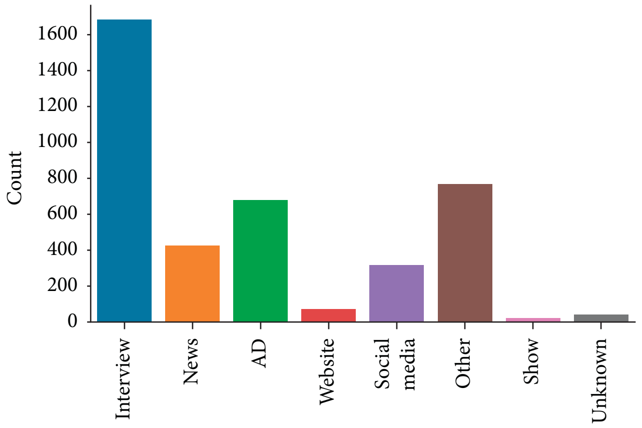


Figure 3: Number of records per category for venue attribute after reduction.

After the basic NLP, word embedding technique was applied. Word embedding is a technique that enhances the performance of the deep learning models for NLP tasks [14]. The words are converted into the real value numbers (vectors) that can be easily executed by the neural network models. The words containing similar meaning have same representation in a vector space. The details of the word embedding are further discussed in the Bi-LSTM-GRU-dense deep learning model.

5. Deep Learning Model

Based on the nature of the features in the dataset, two deep neural network models were designed as discussed below.

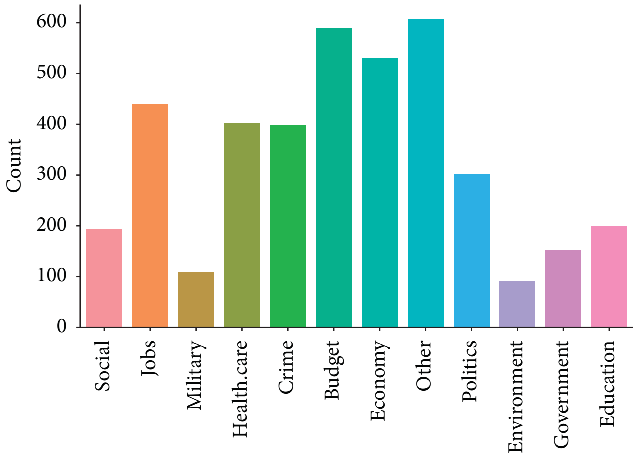


Figure 4: Number of records per category for subject attribute after reduction.

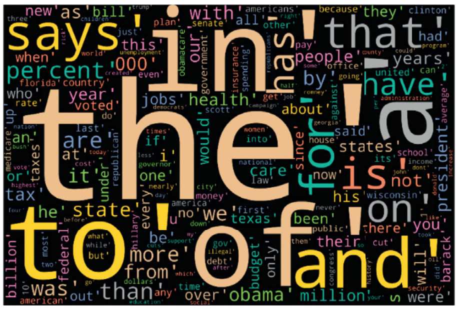


Figure 5: Word cloud before preprocessing.

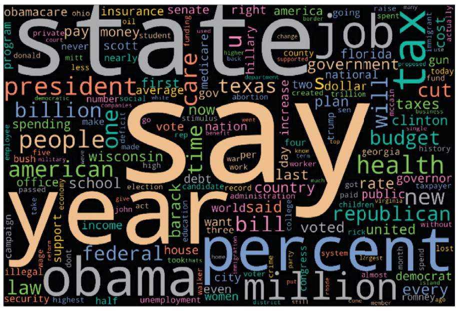


Figure 6: Word cloud after preprocessing.

The first dense model was used for other features. However, the Bi-LSTM-GRU model was used for statement feature.

6. Deep Learning Dense Model

The first model was designed with 10 fully connected dense layers with 9 feature variables as input. The structure of the layers was 512, 256, 256, 256 (dropout_layer), 64, 64, 64 (dropout layer), and 1 (output layer) neurons, respectively. The addition of dropout layers with value 0.5 was used to force the model to learn more robust features. The rectified linear unit "ReLU" activation function was used for input and all hidden layers, while "sigmoid" was used an activation function for output layer. The "Adam" optimizer was used as an optimization algorithm [15], while "binary_crossentropy" was used for loss of the model. The "accuracy" was used to evaluate the model accuracy. The 150 epochs were used which has batch size = 128 with callbacks setting to monitor the "validation accuracy" and save only the best model.

6.1. Deep Learning Dense Model. The main architecture of the second bidirectional LSTM gated recurrent unit (GRU) model was dense neural network with 9 layers with 200 as input as per the size of the vector for each word. The embedding layer in the deep learning model is added along with vocabulary size of 5000, size of the real-valued vector space, i.e., EMBEDDING_DIM is 300, and the maximum length of input documents is 200. The two deep neural models bi-LSTM with 50 units and return_sequence is TRUE and (b) bidirectional Gated Recurrent Unit (bi-GRU) with 50 units, return_sequence is TRUE, and return_state is also TRUE. The addition of global maximum and average pooling layers at the output of the LSTM and GRU is used to make the resulting feature map to more robust to the positional changes of features. The outcome of the both models (i.e., bi-LSTM and bi-GRU) after the global max and global average pool is concatenated to get a single value. The output layer is set to a single dense layer with 1 output. The model was trained with 10 epochs, batch size to 64, and with class weight of {0: 1.1304960541149944, 1: 0.8965131873044255}.

6.2. Experimental Setup and Results. The model was implemented in python 3.9.0, using several libraries such as sklearn, Keras, and matplot. Based on the nature of the dataset features, the experimental setup was prepared accordingly. The dataset was divided into three sets as presented in Table 2. The dataset 1, dataset 2, and dataset 3 represent the feature combination used in the previously mentioned experimental setup.

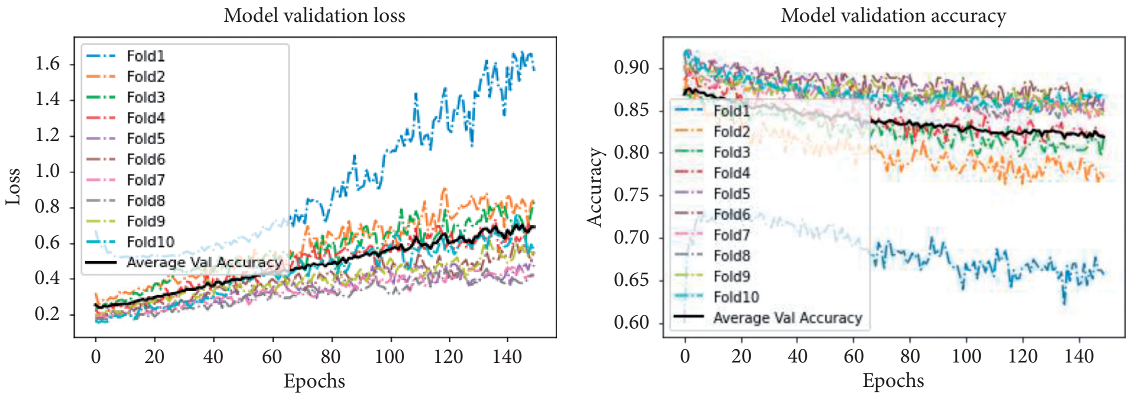
To prepare the dataset for experiment 1 using deep neural network, the word embedding techniques were applied during the preprocessing. The embeddings consider

Table 2: Experimental setup, features, and description.

Experiment-setup	Features	No. of features	Experimental description
Experiment 1	Contain only single attribute “statement”	1	Dataset 1 contains only single feature “statement” and provides the embedding vector of dataset 1 as input to Bi-LSTM-GRU-dense deep learning model, and results were recorded
Experiment 2	Numeric and categorical features excluding statement	9	The dataset 2 included only categorial or numeric data. For this dataset, the first model, i.e., deep learning dense model was used, and results were collected
Experiment 3	Contain all features including “statement”	10	For the dataset 3, the ensemble technique of the proposed model was applied, i.e., for the “statement” feature, the Bi-LSTM-GRU-dense deep learning model was used, while for the rest of 9 features, the deep learning dense model was used. The result of each model is then ensembled by using ensemble voting techniques and is recorded

Table 3: Performance comparison of the proposed fake news detection model for different experiments.

Experiment	Accuracy	Precision	Recall	*F*-score
Experiment 1	0.898	0.913	0.916	0.914
Experiment 2	0.819	0.828	0.852	0.840
Experiment 3	0.859	0.870	0.884	0.877



(a)

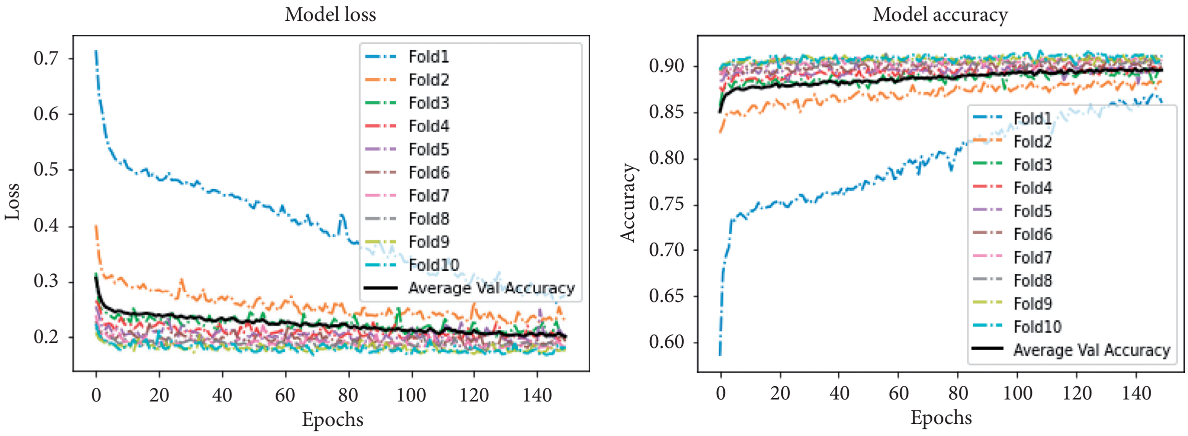
(b)

Figure 7: (a) Model validation loss and the number of epochs for Experiment 1. (b) Model validation accuracy and the number of epochs for Experiment 1.

the context and semantics meaning of the words by producing n-dimensional vector. During the embedding process, the data are encoded to represent each word with a unique integer. Prior to embedding, the Tokenizer API from tensor flow Keras is used to perform the tokenization. Padding was added to make vectors for all words of same length (i.e., max. length set to 200). The embedding matrix was created by using the FastText “cc.en.300.vec” [16] pretrain vector.

The performance of the proposed model was compared in terms of accuracy, precision, recall, and *F*-score. K-fold cross-validation technique was applied for data partitioning with $K = 10$. The results of the proposed model are presented in Table 3. The experiments proved the significance of the proposed deep learning model for fake news detection. The model produced best results with using statement as a feature. The performance of adding other features with the statement degraded the prediction performance. However, the outcome of the proposed model using all the other features excluding statement feature produced the least results. The highest accuracy of 0.898, precision of 0.913, recall of 0.916, and *F*-score of 0.914, respectively, were achieved using only one attribute, i.e., statement. Moreover, Figures 7(a)–8(b) present the model validation and testing loss and accuracy for the first feature set.

Additionally, Table 4 contains the comparison of the proposed model with the benchmark studies in the literature. The criteria for the selection of benchmark were the studies in the literature that used LIAR dataset. The study outperformed the previous studies with an accuracy of 0.898. Like the previous studies [3, 6], the proposed study also achieved the highest results using statement feature only. However, the author in [4] used 12 feature including statement. Long et al. [7] found the speaker profile as one of



(a)

(b)

Figure 8: (a) Model testing loss and the number of epochs for Experiment 1. (b): Model testing accuracy and the number of epochs for Experiment 1.

Table 4: Comparison of the proposed model with the baseline studies using LIAR dataset.

Reference	Year	Techniques	Accuracy
[4]	2017	Hybrid CNN, NLP	0.274
[7]	2017	LSTM	0.415
[3]	2018	CNN, NLP	0.27
[6]	2019	NB, bigram, TF-IDF	0.60
[11]	2020	CapsNet	0.649
[12]	2020	SVM, Bi-LSTM	0.61
[13]	2021	RF	0.442
Current study		Deep learning (ensemble)	0.898

the most significant feature. However, in the current study, the proposed model performance was not enhanced after the inclusion of speaker profile attribute. Conversely, the authors in [11] also found the highest outcome using the combination of the textual and other features. However, the highest accuracy achieved in [11] using the statement feature was 0.565. Thus, the performance of their model was greatly enhanced with the integration of other features like speaker job title and speaker info. In [4, 12], the authors focused on binary classification like the proposed study. They converted the news into two categories as fake and real.

7. Conclusion

The primary goal of this paper is to reduce the drawback of social media, which is the fast spread of fake news that often misleads people, creates wrong perceptions, and has a negative influence on society. Therefore, an ensemble-based deep learning model is constructed to classify the news into fake or real. Several preprocessing techniques were applied initially on the dataset. Furthermore, NLP techniques were applied on statement attribute. Two deep learning models were used, deep learning dense model for the other 9 attributes excluding statement and Bi-LSTM-GRU-dense model for statement attribute. The results achieved by the proposed study is significant with an accuracy of 0.898 using statement feature. This model performance surpassed the other studies on the same dataset, and it is very effective in detecting fake news. Finally, fake news detection using machine learning is still a new topic and challenging. Despite of the significant results achieved by the proposed study, there is still a room for the improvement. The model needs to be investigated using other fake news datasets.

Data Availability

The study used open-source dataset and is accessed from the weblink https://www.kaggle.com/mrisdal/fake-news.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] 2021 https://datareportal.com/social-media-users Global Social Media Overview.

[2] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] M. G. Sherry Girgis and E. amer, “Deep learning algorithms for detecting fake news in online text,” in *Proceedings of the ICCES*, pp. 93–97, Cairo, Egypt, July 2018.

[4] W. Y. Wang, ““Liar, liar pants on fire”: a new benchmark dataset for fake news detection,” in *Proceedings of the Annu. Meet. Assoc. Comput. Linguist*, pp. 422–426, Vancouver, Canada, July 2017.

[5] 2020 https://www.kaggle.com/mrisdal/fake-news Fake or Real News.

[6] J. Y. Khan, M. T. I. Khondaker, A. Iqbal, and S. Afroz, “A benchmark study on machine learning methods for fake news detection,” pp. 1–14, 2019, https://arxiv.org/abs/1905.04749.

[7] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, “Fake news detection through multi-perspective speaker profiles,” in *Proceedings of the Eighth Int. Jt. Conf. Nat. Lang*, pp. 252–256, Taipei, Taiwan, November 2017.

[8] F. A. Ozbay and B. Alatas, “A novel approach for detection of fake news on social media using metaheuristic optimization algorithms,” *Elektron. Ir Elektrotechnika*, vol. 25, no. 4, pp. 62–67, 2019.

[9] Y. Fang, J. Gao, C. Huang, H. Peng, and R. Wu, “Self multi-head attention-based convolutional neural networks for fake news detection,” *PLoS One*, vol. 14, no. 9, pp. 1–14, 2019.

[10] C. Janze and M. Risius, “Automatic detection of fake news on social media platforms,” in *Proceedings of the 21st Pacific-Asia Conference on Information Systems*, pp. 261–276, Langkawi Island, Malaysia, July 2017.

[11] A. M. P. Braşoveanu and R. Andonie, “Integrating machine learning techniques in semantic fake news detection,” *Neural Processing Letters*, vol. 52, no. 2, 2020.

[12] T. C. Truong, Q. B. Diep, I. Zelinka, and R. Senkerik, “Supervised classification methods for fake news identification,” in *Proceedings of the ICAISC*, pp. 445–454, Zakopane, Poland, July 2020.

[13] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, “An ensemble machine learning approach through effective feature extraction to classify fake news,” *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.

[14] N. C. Sahar Ghannay, B. Favre, and Y. Estève, “Word embeddings evaluation and combination,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 300–305, Portorož, Slovenia, May 2016.

[15] J. L. B. Diederik and P. Kingma, “ADAM: a method for stochastic optimization,” in *3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[16] 2021 https://fasttext.cc/docs/en/english-vectors.html “English word vectors.