Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.27

BERT-Based Blended Approach for Fake News Detection

Shafqaat Ahmad
Penn State University
Sua845@psu.edu

Satish Mahadevan Srinivasan
Penn State University
sus64@psu.edu

Abstract

This paper presents a new approach for detecting fake news on social media. Previous works in this domain have demonstrated that context is an important factor when attempting to distinguish subtle differences within text. Fake news itself presents different level of difficulty due the vast similarity that exists between genuine and fake news contents. Therefore, we propose a collaborative approach which uses probabilistic fusion strategy to combine the knowledge gained from modelling two language models, Bidirectional Encoder Representations Transformers (BERT)-long short-term memory network (LSTM) and BERT-convolutional neural network (CNN). To achieve the fusion, we exploit the Bayesian method. Our experiments are conducted on two fake news detection datasets. The detection accuracy attained in these experiments attest to the efficiency of the proposed method, as our approach is very competitive compared to the state-of-the-art methods.

Keywords *natural language processing, language modelling, deep neural network, machine learning, BERT.*

1. Introduction

The importance of social media nowadays as a medium for human interaction cannot be overemphasized. From mere sharing of memes, posting of pictures and videos, to making live broadcast, most people tend to spend a huge amount of their time daily engaging with contents on social media. This has also become one of the reasons why most people use social media as their main source of news as opposed to the traditional news outlets. Moreover, the option of sharing, commenting, and liking news contents, coupled with the flexibility and speed at which these actions can be performed is another motivating factor changing people's interests from traditional news outlets to social media-based news. Despite the great advantage social media offers, the quality of news on social media is incomparable to traditional news. While the cost of quickly posting news on social media is extremely negligible, a considerable proportion of these news are fake, intentionally prepared to propagate false information and negative agenda (Pérez-Rosas et al. 2017). Some news is propagated for political and financial gains, and some are mainly to

divide opinions and create distrust among the public. This inherently changes the belief system of people toward news that are actually genuine (Reis et al. 2019).

Fake news is challenging to detect because the fine line between genuine and fake content is only so obscure, that even for humans, detection of some fake news cannot be 100% guaranteed as context places a huge factor in fake news detection (Sharma et al. 2019). Moreover, attempting to manually detect fake news will lead to a highly laborious process that will at the same time, required huge amount of human resources.

Hence, several research studies have been focusing on development of algorithms and techniques that can be leveraged to automatically detect fake news, particularly on social media contents. This automated approach will save huge amount of time and effort, at time same being more effective. Approaches in machine learning toward fake news detection have typically been based on natural language processing (NLP) techniques, starting from the simple bag-of-words (BoW) representations to more sophisticated, advanced techniques like Word2vec (Zhang et al. 2018; Jang et al. 2019). Exploiting context within text to delineate genuine from fake is difficult, nonetheless, attempts are being made to push the boundaries of NLP toward context understanding. Recently, Bidirectional Encoder Representations Transformers (BERT) have become a revolutionary language model that has become the standard for solving several NLP problems such as machine translation, text summarization, entity extraction and so on (Sanh et al. 2019; Mozafari et al. 2020).

In this work we exploit the representation advantage of BERT, by using the model as an embedding for convolutional neural network (CNN) and long short-term memory network (LSTM). We then further propose a collaborative approach for fusing the knowledge learnt from modelling the two language models as a convoluted hybrid framework. Using BERT-LSTM and BERT-CNN as the baseline models, we propose a fusion method that relies on Bayesian theorem for score level fusion to boost fake news detection performance. Our idea is established on the fact that subtle differences that cannot be captured by a single model can be complemented for, using a hybrid model.

2. Related Works

Solving the problem of fake news detection using datasets retrieved from social media platform has been approached from different perspectives. Some research works have focused on extracting features based on solely on the content of the news articles, while others explore social context to perform feature extraction (Shu et al. 2019a).

2.1 Approaches Based on News Content Features

A lot of research works are exploring raw metainformation such as the title, body, headline, news source, and possibly digital information like videos and images to extract features (Shu et al. 2019b). Since fake news are usually intention coined to mislead the general public, it is often possible that the common words or phrases in these news are written to attract more clicks, views, and likes from social media (Shu et al. 2017). In other words, making the news go viral. Meanwhile such words or phrases can also turn out to be an advantage when extracting features, as lexical level features which are mainly words and characters can be represented using text vectorization algorithms and syntactic level features which mainly consist of sentences, statements, and phrases can be extracted using algorithms based on BoW and word2vec (Guthrie et al. 2006; Wallach 2006; Goldberg and Levy, 2014).

Furthermore, with the recent advances in deep neural networks, language models based on recurrent encoder-decoder networks have also been similarly explored, with techniques such as LSTM (Hochreiter and Schmidhuber, 1997), BERT (Devlin et al. 2018), Generalized Autoregressive Pretraining for Language Understanding (XLNET; Yang et al. 2019) making significant breakthrough in this domain. Pretrained large cased BERT model has been directly applied to detect the possibility of an author on social media to spread fake news (Baruah et al. 2020). Similarly, Rodríguez and Iglesias (2019) applied different architecture of pretrained BERT on fake news detection.

We have seen a few cases where BERT has been integrated with other linear and deep neural network models. For instance, Wu and Chien (2020) combined pretrained BERT with linear classifier by basically using the BERT model to capture the text representation and performing classification of fake news spreaders using a linear classifier. Kula et al. (2019) proposed integrating BERT model with recurrent neural network (RNN) for fake news detection. Kaliyar et al. (2021) propose fake BERT model which combines different parallel components of CNN having different kernel sizes and filters with BERT. The goal of using such an approach is to minimize the effect of ambiguity on text understanding.

BERT has somewhat attracted a considerable amount of attention in fake news detection, either using the model directly or exploring it combination with other machine learning models. However, to the best of our knowledge,

we have not seen a situation where a hybrid combination of two BERT-based deep neural network models have been explored for fake news detection. Moreover, there are no reports of using fusion strategies to combine BERT model with other model, particularly probabilistic score fusion methods.

Hence, in this paper, we the follow the approach of learning representation from news contents and our key contributions are summarized as follow:

- We use BERT as an embedding for training 1D CNN and LSTM.
- We introduce using Bayesian model for performing score level fusion of BERT-CNN and BERT-LSTM.
- We assess the performance of the proposed methods under different parameter settings.
- Finally, we evaluate the performance of the proposed hybrid model, and demonstrate that it performs better than other models that use only BERT, word2vec, or BoW techniques.

3. Proposed Method

This section describes the propose learning methods for achieving fake news detection. The architecture is illustrated in Figure 1, where we follow a couple of processes including text preprocessing, text representation using languages models and postclassification score fusion using Bayesian method.

In the text preprocessing stage, we used regular expressions to remove unnecessary or unwanted characters such as hashtags, punctuations, numbers, html tags, and so on.

After obtaining the cleansed text, we pass the text to the base language models which are explained in the subsequent sections. The models perform both text representations with pretrained embedding and classification of texts into fake or genuine classes. The resulting classification scores are finally passed to Bayesian model for score fusion to obtain the final classification results.

4. BERT

BERT (Devlin et al. 2018), initially developed by Google, has its origins from pretraining contextual representations learning in NLP. It can handle NLP tasks such as supervised text classification, question-answering, text summarization, without the need for human intervention. This technique is widely popular in academics and industry because of its versatility in dealing with any corpus while producing excellent results. BERT is also an encoder-decoder type of model but adds an attention mechanism which performs mapping of a query and a set of key-value pairs to an output helping the model to maintain the relative importance of input (Liu et al. 2019). BERT has two steps, which are: pretraining and fine-tuning. In the pretraining stage, the model is trained on unlabeled data over different pretraining tasks. During fine-tuning, the BERT model is finetuned by first initializing it with the pretrained parameters using Masked LM and Next Sentence Prediction (NSP) and then fine-tuning all of the parameters using labeled data from the downstream tasks, fine-tuning is straightforward since the self-attention mechanism in the transformer allows BERT to model many downstream tasks.

In this work, we use the pretrained uncased Small BERT provided by huggingface with $L = 4$ hidden layers (i.e., Transformer blocks), a hidden size of $H = 256$, and $A = 4$ attention heads. Additional information about the pretrained uncased Small Bert can be obtained from this link https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-4_H-256_A-4/2.

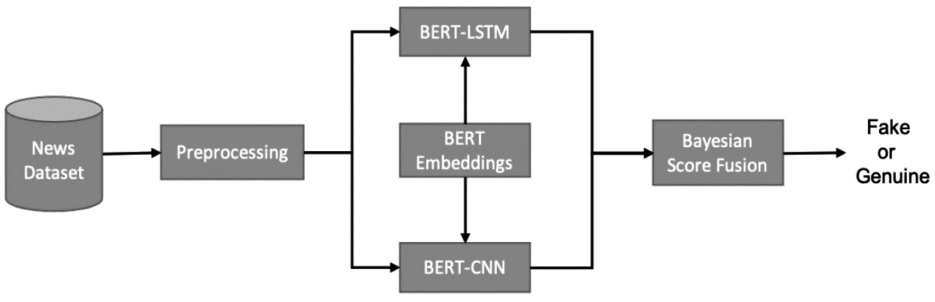

Figure 1: Illustration of the proposed system.

4.1 BERT-CNN

CNN is a one of the most successful deep learning algorithms which combines feature learning with trainable classifiers. It usually has multilayers of convolution, nonlinear transformation, pooling operation, and a fully connected network (FCN).

In CNN, convolutional layers perform operations of convolution on the input data by applying a set of filter banks with varying properties to generate some feature maps. This operation is performed in an iterative manner and the resulting feature maps from a preceding layer are transferred to the subsequent layers to learn more features via convolution. These feature maps are then approximated using an activation function, with common examples are sigmoid and rectified linear units (ReLU). Between two convolutional layers, we normally perform pooling operation to obtain features with strong affinity, which also indirectly eliminates weaker features in the feature map, and at the same time reduces the size of the feature map by replacing the values in a particular region with the statistical summarization of its neighbors.

Two of the most popular techniques are max pooling which replaces values of the feature map with the max value, and average pooling that simply computes the average of the feature map. Another operation generally applied in the learning process is regularization, and the dropout method has proven to be the most successful technique. The final layers of the network consist of FCN and loss layer. The FCN connects every single neuron in one layer to that of another layer, while the loss layer is used for making predictions.

In order to train CNN with BERT, we use the pretrained uncased BERT model embedding vectors as the embedding layer in our CNN, and the CNN model further has three convolutional layers, one global max pooling layer, one dropout layer, and one fully connected layer. We generally used ReLU activation function and sigmoid in the fully connected layer. The loss function is binary crossentropy and ADAM optimizer.

4.2 BERT-LSTM

LSTM has been introduced as a variant of RNN which incorporates memory units into the network. This effectively allows the network to determine the instances to forget previous hidden states or when to update hidden states when new data is fed into the network.

Standard RNN in its most conventional form aims to construct a model with temporal dynamics flow by mapping sequential input data to a hidden state. The hidden states are then mapped to outputs which can be expressed with the following Equation (1), given an input sequence data X:

$$\begin{aligned} h_s &= f(W_{xh}X_s + W_{hh}h_{s-1} + b_s) \\ z_s &= f(W_{hz}h_s + b_z), \end{aligned} \tag{1}$$

where f is a nonlinear activation function computed elementwise, h_s is the hidden state, W is the weight, b is the bias, and z_s is the output at time s. One of the major challenges of RNN is the inability to remember interaction in long-term sequence due to the problem of exploding gradients. As a result, LSTM.

For us to train LSTM with BERT, we use the pretrained uncased BERT model embedding vectors also as the embedding layer in the LSTM model and the LSTM has one spatial dropout layer, one recurrent layer, and two dense layers. The hyperparameters of the LSTM model were fine-tuned to obtain the optimal values. Similarly, the ReLU and sigmoid are used in the dense layers, and the loss is binary cross entropy and ADAM optimizer.

5. Fusion Strategy

In order to fuse the classification scores from the language models we used these techniques:

$$\text{Sum Rule}: \ \text{FS} = \sum_{i=1}^{n} s_i. \tag{2}$$

$$\text{Weighted Sum Rule}: \ \text{Weighted} = \sum_{i=1}^{n} w_i s_i, \tag{3}$$

where, s_i is predicted scores of a classifier and w_i is a weight value assigned to the classifier based on recognition performance.

Bayesian Fusion: this is a probabilistic approach for fusing classification scores originating from the language models. Similar approach has recently been used for fusing multiple sensors in Chen et al. (2021). In this case, we

are able to compensate for the inadequacy of a model with another model. Assume we have fake news labels y, associated to each language model BERT-CNN s_1 and BERT-LSTM s_2 and that the scores from the models are conditionally independent of the news labels y. Hence, we can write the Bayesian expression as:

$$P(s_{1,}, s_2|y) = P(s_1|y)P(s_2|y), \tag{4}$$

which is also equivalent to:

$$P(s_1|y) = P(s_1|s_2, y). \tag{5}$$

We can notice that conditional independence is valid in the equation because given a news label y, predicting the BERTCNN score s_1 will not have any influence on the knowledge gained from BERT-LSTM s_2.

Hence, applying Bayes rule to the above expression will result in:

$$P(y|s_{1,}, s_2) = P(s_{1,}, s_2|y)P(y)\,P(s_{1,}, s_2). \tag{6}$$

6. Experimental Results

Experiments were conducted on two benchmark datasets, retrieved from Kaggle, which are described below. Experimental environment was setup on Google Colaboratory (also known as Colab). The python scripts used for executing the experiments discussed in this manuscript are available in a GitHub repository. The link to the GitHub repository is https://github.com/shafqaatahmad/BERT_based_Blended_approach_for_Detecting_FakeNews.

6.1 Dataset

UTK Machine Learning Club: This dataset was downloaded from Kaggle instigated by the UTK Machine Learning Club 2. It mainly consists of 20,800 news articles which are fairly distributed equally between the two classes namely genuine and fake news. However, out of the 20,800 news articles we randomly selected only 4,000 samples for this experiment. Due to the limitation in the availability of computational resources we decided not to use the entire dataset for this experiment. The distribution of the samples selected for this experiment are shown in Figure 2. In order to build the model, we split the data into 70% training set and 30% test set.

WELFake Dataset: This is a publicly available dataset for fake news detection recently published by Verma et al. (2021). The dataset contains approximately 72,000 news articles, which is a combination of four different datasets: Kaggle, McIntire, Reuters, and BuzzFeed. The dataset is compiled with the aim removing bias or class imbalance from the two news categories. The resulting dataset is 48.55% genuine news and 51.45% fake news.

Due to the large volume of the dataset and limited access to larger computing resources, we randomly select 3,575 samples from the dataset for our experiments. The data distribution is depicted in Figure 3.

6.2 Results

1) *UTK Machine Learning Club*: conducting the experiments on this dataset using the individual algorithms: BERT-LSTM and BERT-CNN, we achieved a prediction rate of 91.09% and 79.2%, respectively, as shown in

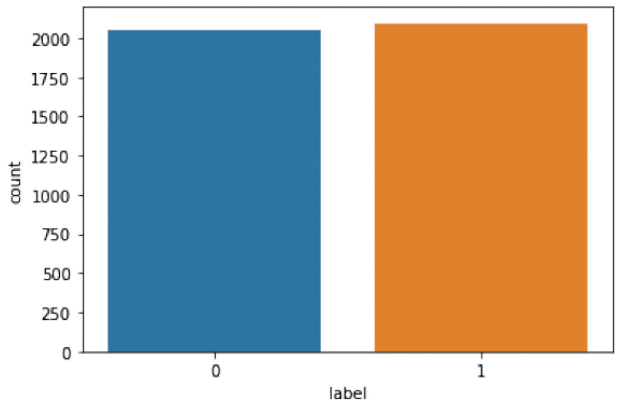

Figure 2: Distribution of samples selected from UTK dataset. 1 = Genuine, 0 = Fake.

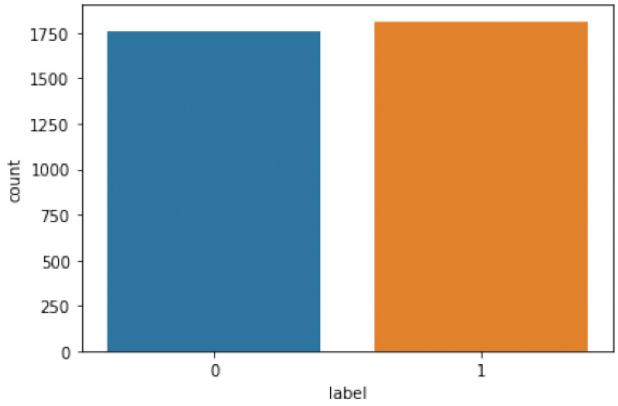

Figure 3: Distribution of samples selected from WELFake dataset. 1 = Genuine, 0 = Fake.

Table 1. In this experiment a 90% weightage was assigned to the classification output of the BERT-CNN model and the remaining 10% weightage was assigned to the classification output of the BERT-LSTM model.

Looking at the results, we did not notice a trend where Genuine class is solely predicted with higher accuracy than the Fake class and vice versa for both language models. As mentioned in the Introduction, the subtle differences between genuine and fake news could be very marginal and that it is very likely that the words used to coin fake news contents are similar to those in genuine contents. However, with the use of language models which are able to properly represent words and understand the context around those words, we can notice that both classes of news are well predicted with considerably high accuracy.

We also tried different architectures of BERT-CNN and BERT-LSTM by basically changing the parameters of the CNN and LSTM models. However, we observed similar trend in classification accuracies of the two models. Nonetheless, we attempted using our proposed score fusion method Bayesian formulation. From the experiments, we observed an improvement in performance by at least 1.5%, yielding a prediction rate of 96.7% as shown in Table 1.

Furthermore, this approach helps in bridging the gap between the two individual language model, moving the accuracy of Fake news to 97.2% and Genuine to 94.1% as opposed to the low classification rates attained using BERT-LSTM. This shows that it is quite beneficial to apply score fusion to compensate inadequacy of a model with another model. Furthermore, we attempted using the standard SUM rule for score fusion, which also produced impressive performance but not as effective as Bayesian method.

2) *WELFake*: Using this dataset, we follow the same experimental setting used on UTK dataset in terms of training/ test split, and CNN/LSTM architecture. With BERT-CNN, we attained a prediction rate of 94.5% as shown in Table 2. In this experiment a 50% weightage was assigned to the classification outputs of both the BERT-CNN and BERT-LSTM models.

In the case of BERT-LSTM, we attained a prediction rate of 80.7% as shown in Table 2. We also noticed that the BERT-CNN generally outperformed BERT-LSTM in this experiment. Also, similar classification trend can be observed where neither of the two classes significantly attained higher accuracy than the other class in all cases.

Meanwhile, using the proposed Bayesian score fusion method, the performance of the model improved to 96.2%, with the fake class increasing to 95.3%. Using SUM rule, we attained an accuracy of 92.8%. This also corroborates the point that the inadequacy of a single model can be compensated by using a hybrid model.

Table 1: Results of experiments on UTK dataset.

Methods	Genuine	Fake	Total
BERT-CNN	90.27	91.2	91.09
BERT-LSTM	80.53	78.9	79.2
SUM rule	77.8	96.2	92.9
Bayes fusion	**94.1**	**97.2**	**96.7**

Table 2: Results of experiments on WELFake dataset.

Methods	Genuine	Fake	Total
BERT-CNN	95.3	93.8	94.5
BERT-LSTM	91.9	69.6	80.7
SUM rule	89.6	95.9	92.8
Bayes fusion	**96.4**	**95.3**	**96.2**

Table 3: Results of experiments on UTK dataset.

Papers	Methods	Results (%)
Trivedi et al. (2021)	BERT	83
Barbosa et al. (2020)	MLP	96.4
Agarwal et al. (2020)	CNN-LSTM	94.7
Our proposed	**Bayes fusion**	**96.7**

6.3 Statistical Analysis

In this study we have considered two statistical techniques namely the Kruskal–Wallis test and the pairwise Wilcox test. Kruskal–Wallis is a nonparametric method for testing whether samples are from the same distribution. The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same. Kruskal–Wallis test is roughly equivalent to one-way ANOVA on ranks. The nonparametric Kruskal–Wallis test does not assume a normal distribution of the underlying data. Thus, Kruskal–Wallis test is more suitable for analysis of dataset where the sample size is small (<30). For the dataset that is not normally distributed and contain some strong outliers, it is more appropriate to use ranks rather than actual values to avoid the testing being affected by the presence of outliers or by the nonnormal distribution of data. This test also assumes that the observations are independent of each other.

The pairwise Wilcox test is performed as a *post hoc* test to determine which groups are different from other groups. Upon randomly varying the sample size of both the datasets we created 17 different samples from both the datasets. On the 34 sample datasets we applied the BERT-CNN, BERT-LSTM and our proposed Bayesian score fusion algorithm. The detection accuracy attained across all the samples of the dataset were analyzed using the Kruskal–Wallis nonparametric test. This test was performed to determine whether there is a significant difference in the performance of these algorithms. On the 17 samples of instances obtained from the UTK Machine Learning Club dataset, we observed a significant difference in the mean detection accuracy of the three algorithms at $a = 0.05$. The Kruskal–Wallis nonparametric test resulted in a p-value of 1.159e-08. In addition to that we also performed a pairwise Wilcox test to determine which group of algorithms differed from each other in terms of the detection accuracy of fake news. At $a = 0.05$, we observed a significant different in the detection accuracy of our proposed Bayesian score fusion algorithm with p-value of 0.008 and 2.6e-09 against the BERTCNN and BERT-LSTM, respectively. Also, there was a significant difference in the detection accuracy of the BERT-CNN against the BERT-LSTM (p-value = 1.1e-06, $a = 0.05$). On the 17 samples of instances obtained from the WELFake dataset, we observed a significant difference in the mean detection accuracy of the three algorithms at $a = 0.05$. The Kruskal–Wallis nonparametric test resulted in a p-value of 1.798e-08. In addition to that we also performed a pairwise Wilcox test to determine which group of algorithms differed from each other in terms of the detection accuracy of fake news. At $a = 0.05$, we observed a significant different in the detection accuracy of our proposed Bayesian score fusion algorithm with p-value of 0.024 and 1.1e-06 against the BERT-CNN and BERT-LSTM, respectively. Also, there was a significant difference in the detection accuracy of the BERT-CNN against the BERT-LSTM (p-value = 2.6e-09, $a = 0.05$).

6.4 Performance Comparison

Finally, in order to ascertain the effectiveness of the proposed method, we compared its performance with existing techniques in the literature which have been implemented on UTK and WELFake datasets as shown in Tables 3 and 4.

From the comparisons it can be noticed that our proposed Bayesian based score fusion method is highly competitive with existing methods. In most cases, we outperformed existing techniques by at least 3% increase in classification accuracy.

Table 4: Performance comparison on WELFake dataset.

Papers	Methods	Results (%)
Verma et al. (2021)	CNN	92.48
Verma et al. (2021)	BERT	93.79
Verma et al. (2021)	TFIDF+SVM	96.7
Our proposed	**Bayes fusion**	**96.2**

6.5 Internal and External Threats to Validity

Here, there are no threats to internal validity of this study since both the datasets analyzed in this study are large-scale datasets containing both real and fake news obtained from both the trustworthy and untrustworthy sources, respectively. More precisely, the trustworthiness of the source has been used as a proxy for the real labels. It is quite possible that both these data sets may suffer from false positives (since untrustworthy sources can spread a mix of real and fake news), and false negatives (false information spread by trustworthy sources, e.g., by accident). However, collecting all the news from a specified set of sources over a period of time mitigates the problems of biases in the dataset. There might be some threat to the external validity of this research because here we reported the performance measures of all the algorithms by comparing the predicted labels from the algorithms with the actual labels i.e., real, or fake which are indeed low-quality labels as the dataset is not manually curated.

7. Conclusion

In this paper, we have presented a new approach for detecting fake news from news posted on social media. We proposed using a probabilistic fusion strategy to combine the knowledge gained from two language models BERT-CNN and BERT-LSTM, at a classification score level. The experiments on these methods were conducted on two benchmarked datasets. Under varying parameter settings, the detection accuracy attained supersede the existing fake news detection methods by at least 3%.

The core objective of the fake news detection system is to effectively monitor and counter the dissemination of misleading content and misinformation with the potential to manipulate public opinion, thoughts, and behaviors on a societal level. Some real-world applications where fake news detection system can be implemented includes, Social Media Platforms, News Organizations, Government Agencies, Public Awareness Campaigns, and Fact-Checking Services.

The detection of fake news is an evolving field with very vast future potential. A particularly intriguing avenue for future research is the integration of multiple data modalities, including images and videos alongside textual data, to create advanced multimodal fake news detection system. Such an approach holds promise for detecting manipulated media, deepfakes, and complementing textual misinformation identification. Additionally, the incorporation of user feedback emerges as another fascinating enhancement for improving the accuracy and reliability of fake news detection system proposed in this paper.

References

Agarwal, A., M. Mittal, A. Pathak, and L. M. Goyal. 2020. "Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning." *SN Computer Science* **1**, no. 3: 143. doi: 10.1007/s42979-020-00165-4

Barbosa, V., Carina de Oliveira, and B. B. Reinaldo. 2020. "AuFa-Automatic Detection and Classification of Fake News Using Neural Networks." *8th International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE 2020)*, Universidad Autonoma De Yucatan, Cancun, January 27–29.

Baruah, A., K. A. Das, F. A. Barbhuiya, and K. Dey. 2020. "Automatic Detection of Fake News Spreaders Using BERT." CLEF (Working paper). https://ceur-ws.org/Vol-2696/paper_237.pdf

Chen, Y. T., J. Shi, C. Mertz, S. Kong, and D. Ramanan. 2021. "Multimodal Object Detection via Bayesian Fusion." Preprint, submitted July 2022. arXiv Preprint arXiv:2104.02904. https://arxiv.org/pdf/2104.02904.pdf

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pretraining of Deep Bidirectional Transformers for Language Understanding." Preprint, submitted May 2019. arXiv:1810.04805. https://arxiv.org/pdf/1810.04805.pdf

Goldberg, Y., and O. Levy. 2014. "word2vec Explained: Deriving Mikolov *et al.*'s Negative-Sampling Word-Embedding Method." Preprint, submitted Feb 2014. arXiv Preprint arXiv:1402.3722. https://arxiv.org/pdf/1402.3722.pdf

Guthrie, D., B. Allison, W. Liu, L. Guthrie, and Y. Wilks. 2006. "A Closer Look at Skip-Gram Modelling." *LREC* **6**: 1222–1225.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* **9**, no. 8: 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jang, B., I. Kim, and J. W. Kim. 2019. "Word2vec Convolutional Neural Networks for Classification of News Articles and Tweets." *PloS One* **14**, no. 8: e0220976. doi: 10.1371/journal.pone.0220976

Kaliyar, R. K., A. Goswami, and P. Narang. 2021. "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach." *Multimedia Tools and Applications* **80**, no. 8: 11765–11788. doi: 10.1007/s11042-020-10183-2

Kula, S., M. Choraś, and R. Kozik. 2019. "Application of the BERT-Based Architecture in Fake News Detection." In *Computational Intelligence in Security for Information Systems Conference*, Herrero, A., Cambra, C., Urda, D., Sedano, J., Quintian, H., Corchado, E., editors, Cham, Switzerland: Springer, 239–249.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." Preprint, submitted July 26. arXiv Preprint arXiv:1907.11692

Mozafari, M., R. Farahbakhsh, and N. Crespi. 2020. "Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model." *PloS One* **15**, no. 8: e0237861. doi: 10.1371/journal.pone.0237861

Pérez-Rosas, V., B. Kleinberg, A. Lefevre, and R. Mihalcea. 2017. "Automatic Detection of Fake News." Preprint, submitted August 23. arXiv Preprint arXiv:1708.07104

Reis, J. C., A. Correia, F. Murai, A. Veloso, and F. Benevenuto. 2019. "Supervised Learning for Fake News Detection." *IEEE Intelligent Systems* **34**, no. 2: 76–81. doi: 10.1109/MIS.2019.2899143

Rodríguez, A. I., and L. L. Iglesias. 2019. "Fake News Detection Using Deep Learning." Preprint, submitted September 29. arXiv Preprint arXiv:1910.03496

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. "DistilBERT, a Distilled Version of BERT: smaller, Faster, Cheaper and Lighter." Preprint, submitted October 2. arXiv Preprint arXiv:1910.01108

Sharma, K., F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. 2019. "Combating Fake News: A Survey on Identification and Mitigation Techniques." *ACM Transactions on Intelligent Systems and Technology* **10**, no. 3: 1–42. doi: 10.1145/3305260

Shu, K., D. Mahudeswaran, and H. Liu. 2019a. "FakeNewsTracker: A Tool for Fake News Collection, Detection, and Visualization." *Computational and Mathematical Organization Theory* **25**, no. 1: 60–71. doi: 10.1007/s10588-018-09280-3

Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter* **19**, no. 1: 22–36. doi: 10.1145/3137597.3137600

Shu, K., S. Wang, and H. Liu. 2019b. "Beyond News Contents: The Role of Social Context for Fake News Detection." *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Melbourne, VIC, Australia, February 11–19.

Trivedi, A., S. Alyssa, M. Prathamesh, M. Subhiksha, P. D. Meghana, M. Malvika, B. Meredith, S. Arathi, J. Ashish, and D. Rahul. 2021. "Defending Democracy: Using Deep Learning to Identify and Prevent Misinformation." Preprint, submitted June 3. arXiv Preprint arXiv:2106.02607

Verma, P. K., A. Prateek, A. Ivone, and P. Radu. 2021. "WELFake: Word Embedding over Linguistic Features for Fake News Detection." *IEEE Transactions on Computational Social Systems* **8**, no. 4: 881–893. doi: 10.1109/TCSS.2021.3068519

Wallach, H. M. 2006. "Topic Modeling: Beyond Bag-of-Words." *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 25–29. doi: 10.1145/1143844.1143967

Wu, S. H., and S. L. Chien. 2020. "A BERT Based Two-Stage Fake News Spreader Profiling System." CLEF (Working paper). https://ceur-ws.org/Vol-2696/paper_177.pdf

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. "Xlnet: Generalized Autoregressive Pretraining for Language Understanding." *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver BC, Canada, December 8–14.

Zhang, S., Y. Wang, and C. Tan. 2018. "Research on Text Classification for Identifying Fake News." *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Jinan, China, December 14–17. doi: 10.1109/SPAC46244.2018.8965536