Fake News Detection Using BERT

Sai Deep Nigidala

Department of Computing Science

Texas A&M University-Corpus Christi Corpus Christi, TX 78412, USA

snigidala@islander.tamucc.edu

Bharath Nekkanti

Department of Computing Science

Texas A&M University-Corpus Christi

Corpus Christi, TX 78412, USA bnekkanti@islander.tamucc.edu

Wenlu Wang

Department of Computing Science

Texas A&M University-Corpus Christi

Corpus Christi, TX 78412, USA wenlu.wang@tamucc.edu

***Abstract— The news ecosystem has evolved from conventional print media to social media networks in the digital age of computing. Since social media sites allow us to receive news much more quickly and with less restrictive editing, fake news spreads at an unprecedented rate and size. Fake news, intentional misinformation, hoaxes, parodies, and satire are all examples of how to manipulate people in order to harm an organization, entity, or person, and/or to profit financially or politically. Because of how it can affect a country's political circumstance, fake news has recently been in the spotlight of mainstream media and the general public (Fake News). In this paper, we propose a FakeBERT deep learning approach based on BERT (Bidirectional Encoder Representations from Transformers) by combining different parallel blocks of a single-layer deep Convolutional Neural Network (CNN) with different kernel sizes and filters with the BERT.***

Keywords—Fake news, Deep learning, BERT, political, Neural network.

# Introduction

With artificial intelligence's rapid development, a significant number of experiments are being performed in order to solve problems that were never considered in the context of computer science. One such issue is the identification of fake news. Since the accessibility to news media has become so simple, as soon as a notable occurrence happens anywhere on the planet, various news outlets strive to make their stories stand out in order to reach as many people as possible through the internet. As a result, news is quickly disseminated to millions and billions of people through a variety of news outlets, including news networks, blogs, websites, and social networking sites.

"Fake News" is a term that refers to false news or propaganda that spreads disinformation across conventional and non-traditional news organizations such as print and television, as well as non-traditional news organizations such as social media. The general motivation for disseminating such information is to deceive readers, damage the credibility of any person, or profit from sensationalism. It is widely regarded as one of the most serious challenges to democracy, freedom of expression, and the Western order [1]. Fake news is becoming more popular on social media sites like Twitter and Facebook. These channels provide a platform for the general public to express themselves in an unfiltered and uncensored manner. As opposed to direct views from news organizations websites, certain news stories hosted or shared on social media platforms receive more views. According to research on the spread of fake news, false information spreads six times faster on Twitter than real information [3].

Fake news differs significantly from traditional suspicious content, such as spam, in many ways: (1) Social impact: Spams are typically found in personal emails or specific review websites, and they only have a local impact on a limited number of people, while fake news in online social networks can have a huge global impact due to the massive user numbers. This is enhanced even more by the widespread knowledge sharing and dissemination among these users. (2) Audience initiative: rather than passively receiving spam emails, users of online social networks will actively search out, receive, and exchange news information with no assurance of its accuracy. (3) Identification difficulty: Spams are normally easier to distinguish when compared to a large number of daily messages (in emails or on review websites) however, detecting false news with incorrect details is extremely difficult, as it necessitates both tedious evidence gathering and diligent fact-checking due to the lack of other equivalent news articles available [4].

The project's primary goal is to identify and classify fake news. Social networking is the predominant means of disseminating such news, and it occasionally finds its way into the mainstream media as well. Because of the grave influence fake news can have on the political outcome of an election, it is becoming increasingly important to identify and recognize it as such. The primary difficulty in resolving the issue of fake news is the ambiguous meaning of the word. However, advancing computationally intensive, broad coverage models in this direction is still hampered by the lack of manually labelled fake news datasets. Vlachos and Riedel (2014) were the first to make a public fake news identification and fact-checking dataset accessible, but it only contains 221 sentences, making machine learning evaluations impossible [8]. The LIAR dataset, which includes 12,836 short statements labeled for truthfulness, topic, context/venue, speaker, state, group, and prior history, is used to address these issues. LIAR is an order of magnitude greater than comparable resources presently available, with such a large volume and a ten-year time frame.

We propose a BERT-based deep learning approach in this paper by combining different parallel blocks of single-layer CNNs with Transformers' Bidirectional Encoder Representations (BERT). We use BERT as a sentence encoder, which can reliably extract a sentence's context representation. This work differs from previous studies in which researchers looked at a text sequence in a particular direction. With sequential neural networks to encode the relevant information, several current and useful methods had been presented. A deep neural network with a bidirectional training strategy, on the other hand, could be the most optimal and accurate solution for detecting fake news. With the powerful ability to capture semantic and long-distance dependencies in sentences, our proposed approach enhances the efficiency of fake news detection.

## Motivation:

In today's modern era, where there are thousands of information exchange channels through which false news or disinformation can circulate, the widespread issue of fake news is extremely difficult to fight. It has become a bigger problem as AI advances, bringing with it artificial bots that can be used to build and distribute false news. The problem is critical because many people believe everything they read on the internet, and those who are inexperienced or unfamiliar to modern media are sensitive to being duped. Fraud is another problem that may arise as a result of spam or malicious emails and tweets.

# RELATED WORK

In the context of computational linguistics, deception detection is a text classification issue in which our method must identify an unseen document as either truthful or deceptive. A device like this is first trained on proven cases of deception. Vlachos and Riedel [8] published one of the first papers on the automated identification of false news. The authors characterized fact-checking as a classification activity, collected a dataset from two common fact-checking websites, and used k-Nearest Neighbors classifiers to handle it. Wang [5] published the LIAR dataset, which contains 12.8K manually labeled short PolitiFact statements. Several classifiers were compared on a six-level truthfulness classification challenge, and those that included additional textual and contextual features including topic, speaker, and history got better results.

In this paper [3] they present the solution to the task of fake news detection by using Deep Learning architectures. Automated identification of false news is difficult to achieve because it requires the model's understanding of natural language nuances. Furthermore, the majority of current fake news detection models treat the issue as a binary classification function, limiting the model's ability to understand how closely the broadcast news is connected to the real news. To fill in the gaps, they present a neural network architecture that can reliably predict the stance of a given pair of headlines and article bodies. When the stances between the headline and the news article are ‘unrelated,' ‘agree,' and ‘discuss,' their model does fairly well, but the prediction accuracy for the “disagree” stance is poor. Using a finely tuned Tf-IDF – Dense neural network (DNN) model they are able to outperform existing model architectures. They also played around with various hyperparameters and found that using regularization strategies like Dropout, L2 regularization, cross-validation, and early stopping, they could achieve a very smooth and consistent learning process. Finally, they would like to build on this work by conducting similar research on a completely different dataset, such as Twitter or Facebook and hope to get one step closer to creating an automated fake news identification mechanism by classifying fake news from social media platforms. This research establishes a foundation for future research.

The main goal in this paper [9] is to create a classifier that can predict whether a piece of news is fake or not based solely on its content, using RNN technique models (vanilla, GRU) and LSTMs to approach the problem from a purely deep learning perspective. They'll show the differences and evaluate the results by adding them to the LAIR dataset and using word embedding to generate vectors of terms, which then feed into deep learning technique. They discovered that the results of the experiments are similar, but GRU (Gated Recurrent Unit) is the best because it solves the problems of Vanilla, which is well-known for its gradient vanishing problem, and LSTMs (long short-term memories), which GRU is simple to change and does not need memory units, so they gain faster training than LSTM, which has an effect on the output. However, when compared to other papers, it is discovered that CNN's (Convolutional Neural Networks) is the better model from the rest because of its speed and best results and efficiency for Windows. In the future, they will improve this consistency by combining GRU and CNN results to achieve the best result.

In this paper [10], both BERT and ELMo have been shown to be capable of recognizing and classifying hyperpartisan news. Despite the fact that the models performed poorly at first, the analysis improved the performance of the two models after further investigation and tuning. Since this analysis only used BERT CASE, a future study might use BERT LARGE or even expand to train the BERT in a semi-supervised and supervised manner to see how it performs on the dataset. The ELMo model currently only works with the by-articles dataset. Incorporating the by-publisher dataset into the ELMo model to classify hyperpartisan news would be an important project. The research can also be used as a benchmark for evaluating other hyperpartisan-related works. The results of the optimized BERT and ELMo models has achieved 68.4% and 60.8%, respectively.

# DATASET

The LIAR dataset (Wang, 2017) is made up of 12,836 short statements from POLITIFACT that have been labeled by humans for truthfulness, topic, context/venue, speaker, state, group, and prior background. The LIAR dataset has six labels for truthfulness: pants-fire, false, mostly-false, half-true, mostly-true, and true. The sizes of these six label sets are fairly evenly distributed. The declarations were gathered from a number of broad-casting mediums, including TV interviews, speeches, tweets, and debates, and they address a wide range of subjects, including the economy, health care, taxation, and election.

We expand the LIAR dataset to the LIAR-PLUS dataset by automatically extracting the argument given by humans in the fact-checking article associated with each assertion in the LIAR-PLUS dataset. The majority of the papers conclude with a review headlined "our ruling" or "summing up." Several reasoning sentences relevant to the argument are normally included in this overview. We take all sentences from these overview sections, or the last five sentences from the fact-checking article if there isn't one. The sentence with the verdict and related terms is filtered out. These extracted sentences may either support or contradict the argument, which should improve the classification accuracy.

# PROPOSED WORK

BERT:

BERT is a transformer encoded architecture-based advanced pre-trained word embedding model. BERT is used as a sentence encoder, which can reliably extract a sentence's meaning representation. Using a mask language model, BERT eliminates the unidirectional constraint (MLM). It masks some tokens from the input at random and predicts the original vocabulary id of the masked word using only the input. As compared to previous embedding approaches, MLM has improved BERT's ability to outperform. It is a highly bidirectional mechanism capable of handling unlabeled text in all layers by conditioning on both left and right context. We extracted embeddings for a sentence or a group of words in this study, as well as pooling the hidden-state sequence for the entire input sequence. A deep left-to-right and right-to-left model is more effective than a shallow left-to-right or right-to-left model. For context-specific tasks, two types of BERT models have been investigated in previous research:

* BERT Base: Smaller in size, less expensive to compute, and unsuitable for complex text mining operations.
* BERT Large: It's bigger, more computationally intensive, and crunches a lot of text data to get the best results.

|  |  |  |
| --- | --- | --- |
| Parameter Name | Value of parameter (BERT-Base) | Value of parameter (BERT-Large) |
| Number of Layers | 12 | 24 |
| Hidden Size | 768 | 1024 |
| Attention Heads | 12 | 16 |
| Number of parameters | 110M | 340M |

Table 1: Information about parameter settings.

Diagram

Description automatically generated

Fig 1: BERT Pre-training Architecture.

Pre-training and fine-tuning are the two stages in our system. The model is trained on unlabeled data through various pre-training tasks during pre-training. The BERT model is fine-tuned using labeled data from downstream tasks after it is initialized with the pre-trained parameters. Even though they are all initiated with the same pre-trained parameters, each downstream task has its own fine-tuned model.

Diagram

Description automatically generated

Fig 2: BERT Fine-training Architecture.

Enhanced claim/statement representation captures the additional information which is useful such as hedging. From our experiments we found BERT can be tuned to work on classification to some extent. We have tested with different training strategies with BERT as a base architecture which is then tuned for the text classification. We prefer to choose BERT as the base architecture because of state of art performance in language translation and language modeling tasks.

# EXPERIMENT AND RESULTS

We use three training strategies and architecture to get the desired results.

**Finetuning BERT:**

Fine-tuning is simple since the Transformer's self-attention function helps BERT to model a wide range of downstream functions, whether they include single texts or pairs of texts, by switching out the necessary inputs and outputs. A typical pattern for applications involving text pairs is to encode text pairs separately before implementing bidirectional cross focus. Instead, BERT combines these two steps using the self-attention process, since encoding a concatenated text pair with self-attention essentially requires bidirectional cross attention between two sentences.

We use the Finetuned BERT architecture for classification by passing the tensor from BERT into a linear layer which gives binary output logics. We only use the new statements to train the network. No metadata can be used. With this, we were able to achieve 60 % accuracy on binary classification task.

Diagram

Description automatically generated

Fig 3: Finetuning BERT Block Diagram.

**Siamese network with BERT as the base network:**

Constructed a Siamese network with two branches and with each branch having BERT as the base model. The tokens corresponding to new statements for which we need to predict the labels which will be the first branch input. The tokens corresponding to justification of the specific news statement passed to the first branch will be second branch input. Each BERT layer will produce a 1D shape tensor as its output (768). We will get two 1-D form tensors for which we have the two branches (768). These two outputs are now combined and passed through a linear layer. The performance probabilities are calculated using two logits and a ‘softmax’ activation.

Both branches of this architecture have the same weights between them. This method is used to leverage the additional information we have, in this case the ‘justifications’.

The binary classification accuracy of this system as 65.4 percent.

Diagram

Description automatically generated

Fig 4: Siamese Network Block Diagram.

**Triple branch Siamese network with BERT as the base network:**

This design is identical to the previous case, but we have added one more branch with BERT as the base network to the previous case Siamese network, making it a triple branch Siamese network. The input to this additional branch will be the remaining meta data available like speaker, source, affiliation etc. apart from the justification.

Diagram

Description automatically generated

Fig 5: Triple Branch Siamese Network Block Diagram.

The second change/addition here is to consider the publisher's integrity. We have created a feature called " Credit Score " for this reason. According to the dataset documentation, the columns 9-13 in the dataset refer to the number of barely true counts, fake counts, half true counts, mostly true counts, and pants on fire counts made by the news source.

So, the Credit score is calculated as,

A = [(mostly true counts) \* 0.2 + (half true counts) \* 0.5 + (barely true counts) \* 0.75 + (false counts) \* 0.9 + (pants on fire counts) \* 1]

B = [mostly true counts + half true counts + barely true counts + false counts + pants on fire counts]

Credit\_score = A / B

The credit score indicates how true or false the information provided by the author or source is on average. The tensor arising from the concatenation of the credit score is multiplied by this credit score from all the three branches of the Siamese network output tensors and then multiplied 1D tensor is passed through a fully connected layers to get logits as output. The reason for using this credit score is to increase the relative gap between performance activations in the false and real cases (as the credit score for the publisher who publishes fake news would be higher than for someone who publishes real news)

The model was unable to learn anything from the training data when it came to binary classification. Throughput the preparation, there is also a persistent loss. This may be as a result of the fact. There are a lot of moving parts here, including credit score integration, meta data integration, and so on. As a result, fine-tuning the network and learning parameters became challenging. It was also unable to adequately modify various parameters due to the limited computational resources available to us and the lengthy training times the network needs. It was difficult to experiment with various methods for integrating meta data and news.

Two further improvements to this approach have been made, producing improved performance. They are,

**Modification 1:** Instead of multiplication, the credit scores were added to the contribution of the concatenation layer. In addition, the learning rate was reduced by five times.

**Modification 2:** Instead of giving all three branches the same input sequence size (128), the input sequence size is varied based on the form of data and the average number of words in each. Since there are only 5-10 input sequences of more than 64 terms, the sequence size for the branch that uses news statements as input is 64. Since certain justifications have 128 to 264 words and there are only about 10 input sequences of more than 264 words, the sequence size for the branch that takes justifications as input is 256. The same is true for metadata input, where the input sequence size is set to 32 so there are no inputs that have more than 32 words.

|  |  |
| --- | --- |
| Testing Set | Binary Classification |
| LIAR-PLUS Test Set | 75.2% |

Table 2: Highest Accuracy Achieved

Chart, line chart

Description automatically generated

Fig 6: Graph Diagram of Model Accuracy.

Chart, line chart

Description automatically generated

Fig 7: Graph Diagram of Model Loss.

# CONCLUSION

In recent years, fake news detection has an important role in business, law enforcement, national security, political due to the potential impact fake reviews can have on consumer behavior and purchasing decisions. Researchers used deep learning to improve learning and hence get the best results by using word embedding to derive features or clues that differentiate syntactic and semantic relationships between words. In this research, we have demonstrated the performance of our proposed model for fake news detection. We have used three architectures in this model. Finetuning, Siamese network with BERT as the base network and triple branch siamese network with BERT as the base network. The binary classification accuracy is around 60%, 65% and 75% respectively for the above three methods. We used a feature called credit\_score for triple branch siamese network which tells us about how false or fake the news published by the author or the source is on average.

In future work, we would like to develop a hybrid approach that applies to both the binary and multi-class real-world fake news datasets, combining content, meaning, and temporal level information from news articles. For multi-label datasets that propagate in a graph, this hybrid technique may be useful for detecting instances of fake news. Furthermore, we would like to extend our dataset to contain not only text but also photographs and videos.

##### References

1. Langin, K.: http://www.sciencemag.org. Fake news spreads faster than true news on Twitter—thanks to people, not bots (2018)
2. Allcott, H., & Gentzkow, M. (2017). *Social Media and Fake News in the 2016 Election.*doi:10.3386/w23089
3. Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1 : No. 3 , Article 10.
4. Zhang, J., Dong, B., and Yu, P. S., “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, <i>arXiv e-prints</i>, 2018.
5. William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
6. Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaud- huri, Christian M Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance detection task. *arXiv preprint arXiv:1806.05180*.
7. Yoon Kim. 2014. Convolutional neural net- works for sentence classification. *arXiv preprint arXiv:1408.5882*.
8. Vlachos, Andreas & Riedel, Sebastian. (2014). Fact Checking: Task definition and dataset construction. 18-22. 10.3115/v1/W14-2508.
9. S. Girgis, E. Amer and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2018, pp. 93-97, doi: 10.1109/ICCES.2018.8639198.
10. G. K. W. Huang and J. C. Lee, "Hyperpartisan News and Articles Detection Using BERT and ELMo," 2019 International Conference on Computer and Drone Applications (IConDA), Kuching, Malaysia, 2019, pp. 29-32, doi: 10.1109/IConDA47345.2019.9034917.
11. Kaliyar, R.K., Goswami, A. & Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimed Tools Appl* (2021). https://doi.org/10.1007/s11042-020-10183-2
12. B. Bhutani, N. Rastogi, P. Sehgal and A. Purwar, "Fake News Detection Using Sentiment Analysis," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844880.
13. Alhindi, Tariq & Petridis, Savvas & Muresan, Smaranda. (2018). Where is Your Evidence: Improving Fact-checking by Justification Modeling. 85-90. 10.18653/v1/W18-5513.
14. Ahmed H, Traore I, Saad 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. Springer, Cham, pp 127–138
15. Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inform Sci 497:38–55