Leveraging Deep Learning and Natural Language Processing for the Identification of Deceptive Online Content

Kamran Dahri[1], Faheem Ahmed[1], Muhammad Aquib[1], Mohib Ali Khan[1], Muhammad Yaqoob Koondhar[2]
[1]Department of Information Technology, University of Sindh, Jamshoro, Sindh, Pakistan.
[2]Information Technology Centre, Sindh Agriculture University.
*Correspondence: faheem.abbasi@usindh.edu.pk

Citation | Dahri. K, Ahmed. F, Aquib. M, Khan. M. A, Koondhar. M. Y, "Leveraging Deep Learning and Natural Language Processing for the Identification of Deceptive Online Content", IJIST, Vol. 08 Issue. 01 pp 429-436, February 2026
Received| January 09, 2026 Revised| February 09, 2026 Accepted| February 12, 2026 Published| February 16, 2026.

The digital world has a problem with news spreading really fast. Fake news is an issue in the media and news networks. It is a problem that threatens the way we share information around the world. This research paper is about finding a way to automatically detect content. We used machine learning and linguistics, and advanced neural networks to do this. We tried methods to see what works best. We used ways like Logistic Regression and more advanced ways like BERT. We tested our method using the Fake News Net and LIAR datasets. We got good results. Traditional ways, like Support Vector Machines and Random Forests, are still good. Deep learning models work even better. The The BERT model performed well, achieving 92.5% accuracy in detecting fake news. This paper also talks about the problems we face when dealing with news. There are biases in the algorithms. It is hard to understand how the detectors work. We found out that automated detectors are good at what they do. But making them better by improving these models by incorporating contextual understanding remains a challenging problem. If we use what we know about the context, we can help keep conversations safe and make people trust the media again. The digital world has a problem with news spreading really fast. Fake news is an issue in the media and news networks. It is a problem that threatens the way we share information around the world. This research paper is about finding a way to automatically detect content. We used news detection methods and advanced neural networks to do this. We tried fake news detection methods to see what works best. We used ways like Logistic Regression and more advanced ways like BERT to detect fake news. We tested our news detection method using the Fake News Net and LIAR datasets, and we got good results. Traditional machine learning methods such as Support Vector Machines and Random Forests remain competitive at detecting news. Deep learning models work even better at detecting fake news. The BERT model was very good at detecting news stories. It was 92.5 percent of the time when detecting fake news. This paper also talks about the problems we face when dealing with news. There are biases in the news detection algorithms, and it is hard to understand how the fake news detectors work. We found out that automated fake news detectors are good at what they do. But making them better by adding a touch to fake news detection is a hard problem. If we use what we know about the context, we can help keep conversations safe from fake news and make people trust the media again when it comes to fake news.

Keywords: Fake News Detection, Misinformation, Bidirectional LSTM (Bi-LSTM), Support Vector Machine (SVM), Text Preprocessing, Tokenization, Contextual Embeddings, Deepfake Detection

Introduction:

The information explosion has re-engineered the social landscape and political discourse to a great extent. Online channels have become the dominant medium as news consumption has grown exponentially, in terms of the rate and quantity of information passed. This is the democratization of content production, which inadvertently has created the concept of fake news, the word of which we can just say, speaking against the wind, a contrived distortion of information published, to announce the facts as broadcasting. The political-social effects of this phenomenon are radical; the purpose of their application, such as the recent scholarly discourse would have it, is to do what misinformation within democracy, social upheavals, and even the electoral results have been influenced and assisted by a general mistrust of institutional journalism.

Given that the amount of digital content is more than the human brain can process, the more qualitatively superior, yet less scalable, traditional fact-checking by human labor fails to meet the requirement of being scalable. Consequently, scaled, powerful, and automated detection systems have been built, pushed toward a hypothetically desired objective into an information security requirement. This research satisfies this important need by modelling an information-intensive solution that extrapolates the overlaps of Deep learning, machine learning (ML), and Natural Language Processing (NLP). The study aims to make the study on human intuition transition to a paradigm of computational modelling to offer a technical background to detect lie content on a large scale. Our objective is to experiment with the consequences of different architectural constructions of the high-dimensional mapping of features. Machine learning Support Vector Machines to sequential memory of Bidirectional LSTMs-to deception marks of a hoaxed news, therefore of a full-sized one, a difficult linguistic.

Literature Review:

The academic field of detecting fake news has progressed from simple heuristic-based classifier systems to more complex multimodal and transformer systems. This section synthesizes the prevailing research and provides a thematic overview of the field's technical trajectory.

Content-Based vs. Context-Based Detection:

The difference between finding news by looking at what is inside the news and finding fake news by looking at what is outside the news is a big deal. Fake news can be found by looking at the words and sentences in the news. This is what [1][2] and others stated in their studies. . They said we should look at the vocabulary and the way the words are put together in the news. Either remove phrase or complete contrast structure to find out if it is fake, like how people are talking about it and if they are sharing it on social media. Credibility of the author based on historical truthfulness should also be considered [3] and other people found out in 2018 that fake news often uses words that try to make people feel something rather than think about what is real. They said that fake news often has headlines that are like "clickbait," which means they are trying to get people to click on the news rather than really telling them what is going on. Fake news also often uses words that are very emotional and tries to make people feel a certain way rather than just telling them the facts. This is what Pérez-Rosas and other people found out. It can help us find out what is real and what is not. We can use this information to find news by looking at the news itself and by looking at what is going on around the news.

Hybrid and Multi-Modal Approaches: The Integration of Behavior:

Researchers know that fake information is not usually posted on its own, so they are now using a mix of methods that combine all sorts of data. The CSI (Capture, Score, and Integrate) model [4] is significant because it actually does what some other people, like Shu, said should be done. The CSI model looks at what people are saying, figures out if the people

using the site are being honest because fake accounts and bad people do things in a way that is different from real people, and then it puts all of this information together to decide what is real and what is not. This way of looking at everything is important because it considers what the words actually mean and also what is going on with the people using the site, which is necessary to find fake information and be sure it is really fake. The CSI model is an example of how this can be done by looking at the content and the people using the site at the same time.

The Evolution of Benchmarking: From Long-Form to Short-Text:

The creation of datasets has really helped innovation move forward.[5] introduced the LIAR dataset, which pivoted the research focus toward short-text political statements. The LIAR dataset showed how hard it is to work with texts, where computer models cannot rely on a lot of extra information. After that, the [6] dataset was created in 2018. This was a change towards checking facts using evidence. It meant that computer models had to do more than just say if a claim is true or not. The FEVER dataset required models to find and look at evidence from sources to back up their claims. This is how we got the fact-checking systems we use today. Merge properly: The FEVER dataset is important for fact-checking systems.

Transformer Architectures and Contextual Embeddings:

In recent years, the largest difference compared to previous years has been going from static word embeddings (e.g., Word2Vec) to creating dynamic architectural models using transformers. There are multiple papers like [7][8] about how models such as RoBERTa and XLNet are setting new benchmarks for models in many predictive tasks. Unlike prior models, RoBERTa and XLNet are not based on performing tasks on text input using simple sequential/one-dimensional inputs (horizontally or vertically), but will utilize two-dimensional actions (input). This distinction is significant when attempting to identify differences between words and phrases, as well as between very small subtleties that are typically found in instances of fake news (e.g., "he was killed" as opposed to "he committed suicide"). RoBERTa and XLNet can accomplish this by having a contextual understanding of words, which is critical to recognizing discrepancies and subtleties associated with the delivery of misinformation. [7][8] research provide ample evidence that RoBERTa and XLNet perform quite well at this task.

Contemporary Developments in Forgery Detection (2023–2024):

People are now looking into deception that uses forms of media and can create things. Some researchers, like [9][10], found a way to use changes in frequency to find deepfakes that're hard to spot. Then, The ISTVT [11][12] use a kind of architecture to look at videos and find things that are not quite right with the faces in them. These things are so small that people cannot see them. Other researchers [13][14] are looking into finding things in videos that have many people in them or videos that are different sizes. They are doing this in 2024. There is also wok called MMNet [15] which focuses on looking at things in space and time. This means that the next systems we build to find things need to be able to look at many things at once and look at them in a special order. This is because fake things made by computers are getting very good, and we need to be able to stop them.

Methodology:

This study uses a stage technical process to test how well different computer systems work in a controlled environment. The study uses this process to check how good various computational architectures are.

Dataset Analysis and Feature Engineering:

To make sure we are doing an evaluation, we picked two main groups of data to compare:

Fake Newsnet: This group of data is really important because it brings together what people write about the situation. These sources allow us to better understand how information

spreads through social media platforms (e.g., fake news) and ultimately give us insights about the way individuals will react to current events.

LIAR Dataset: This group of data is about what politicians say. The LIAR Dataset has 12,836 statements from them. What makes the LIAR Dataset hard to work with is that it uses a system with six levels to label what politicians say: True, Mostly True, Half Barely True, False, and Pants on Fire. The good thing about the LIAR Dataset is that it has a lot of information: the persons job, what party they belong to and a record of how honest they have been in the past which's like a history of how many true and false statements they have made and this helps the model think logically about what is true and what is not, in the LIAR Dataset.

Data Preprocessing Pipeline:

To transform raw, unstructured text into a machine-readable format, we implemented a rigorous preprocessing pipeline using the Natural Language Toolkit (NLTK):

Noise Removal: Systematic elimination of punctuation, numerical noise, and URLs to prevent the model from learning from non-informative features.

Stopword Removal: Filtering out high-frequency words (e.g., "the", "is", "at") that do not contribute to the semantic uniqueness of a document.

Tokenisation and Stemming: Segmenting sentences into individual tokens and reducing words to their root forms (e.g., "fabricated" to "fabricate") to standardize the vocabulary and reduce dimensionality.

TF-IDF Vectorisation: The application of Term Frequency-Inverse Document Frequency (TF-IDF) to convert text into numerical vectors. This statistical measure weights terms based on their local importance (within a document) relative to their global frequency (across the corpus), effectively identifying keywords that characterize deceptive vs. legitimate content.

Technical Model Specifications:

Logistic Regression:

We are going to begin with Logistic Regression as our model. Logistic Regression is a way to find out how likely something is to happen. It does this by using a function that looks at every word in the text. This function then decides if something is probably true or probably not true. Logistic Regression is a simple thing, but it is actually very good at working with large sets of data that do not have a lot of details. This is why Logistic Regression is a starting point to compare with other models that are more complicated, like Logistic Regression.

Support Vector Machines (SVM):

The Support Vector Machine architecture is really good at dealing with the dimensional feature spaces that come from Term Frequency-Inverse Document Frequency. It uses a method to find the best hyperplane that makes the most space between classes. This method is called a kernel trick. It uses either a Radial Basis Function or a Linear kernel. The Support Vector Machine uses this to find the hyperplane that makes the most space between classes. When we are talking about news, this helps the Support Vector Machine find the best way to separate things into groups, even when the groups are not simple and cannot be separated by a straight line. The Support Vector Machine is good at finding the way to separate fake news from real news, even when it is hard to tell them apart.

Random Forest:

The Random Forest method is a way of learning that creates a group of decision trees that are not connected to each other. When the Random Forest method is being trained, it uses something called bagging to make sure all the trees are different from one another. The Random Forest method is used for classification. It gives us the most common class. The Random Forest method is good because it stops the model from overfitting, and it helps us see which features are important. This is really helpful because it tells us which words or phrases are most likely to show that someone is not telling the truth. The Random Forest method is very useful for understanding what makes the Random Forest method work. It

helps us with the Random Forest method by showing us which linguistic markers are most predictive of deception using the Random Forest method.

Bidirectional LSTM (Bi-LSTM):

The LSTM, a type of Recurrent Neural Network, is made to fix the issue of gradients getting smaller and smaller as the sequence gets longer. This issue is called the "vanishing gradient" problem. It stops the model from learning things that are far apart in a sequence. The Bi-LSTM is a version of the LSTM.

It works by looking at the input sequence in two ways:

from the start to the end

and from the end to the start.

This two-way approach helps the model understand the context of a word. It does this by considering both what comes before and after a word. This solves the problem of not having context, which is a limitation of models that only look one way.

The Bi-LSTM helps the model learn from the sequence, not just one part of it.

It captures the contextual environment of a word. This is how Bi-LSTM works to improve the understanding of sequences.

BERT (Bidirectional Encoder Representations from Transformers):

BERT is today's method to process natural language. Unlike models that look at a single word within text, the underlying architecture of BERT is based on using Multi-Head Self-Attention, which allows each word in the sentence access to all other words in that same sentence simultaneously, regardless of their position [16].

Technical Depth: There are two main models that BERT learns from. The first is called Masked Language Modelling (MLM) where BERT learns to predict the value of a word by looking at the other words within the sentence positionally adjacent to researched word; thus the second task is known as Next Sentence Prediction (NSP), where BERT learns to determine whether or not two sentences are semantically related in meaning to one another - this gives BERT a solid understanding of language. As part of this study we fine-tune a pre-trained BERT model called bert-base-uncased further utilize this pre-trained model provides with an already significant level of computational power, consisting of twelve (12) layers and 110 Million (110M) parameters; we utilize this excess computational power to find language inconsistencies, i.e., phrases/words that appear correct at face-value but do not align with reality when further evaluated within their contextual limits for semantic understanding; and we can identify these types of inconsistencies with BERTs help.

Results and Analysis:

The empirical effectiveness of the models that had been implemented was assessed in a standardized set of metrics, which guaranteed the multi-dimensional assessment of their predictive capabilities.

Table 1. Performance Metrics Table

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.4%	0.81	0.82	0.815
SVM	85.2%	0.84	0.85	0.845
Random Forest	87.6%	0.86	0.87	0.865
Bi-LSTM	89.1%	0.89	0.89	0.890
BERT	92.5%	0.93	0.92	0.925

Comparative Analysis: Table 1 shows, the findings highlight that there are distinct proportions of performance based on the progress in deep contextualizing linear feature mapping through our models. Starting with a basic linguistic frequency (using a TF-IDF), there was a baseline Logistic Regression accuracy of 83.4%, indicating that basic frequencies can work very well to identify clear misinformation. And yet, the small gains seen by SVM (85.2)

and Random Forest (87.6) indicate that there is real value in using non-parametric ensemble techniques to model the high-dimensional/non-linear boundaries of generated text.

The greatest performance improvement is observed in deep learning models, achieving an accuracy of 89.1% using a bi-LSTM, which lends support to the notion of temporal context being an essential factor for successful detections. BERT achieved the highest accuracy at 92.5%, 9.1% above our base case, with the clear distinction being the use of dense and contextualized representations of all words instead of sparse and discrete representations with TF-IDF. While traditional models may struggle to separate words found in satirical contexts to describe what they represent, BERT can easily identify the difference based on its understanding of word usage with respect to time.

BERT achieved an F1-score of 0.925 on the dataset. When we refer to the cost of a false positive, we are also referring to the cost of built-in user confidence as well.

Challenges and Limitations:

These systems are really good at getting things right. There are still some big problems that stop them from working well for everyone. The thing is, these systems have a lot of trouble with things, and that means they do not work as well as they should for everybody. These systems are what we are talking about here.

Data Bias and Fairness:

Model training relies on human-annotated datasets such as LIAR or FakeNewsNet like LIAR or FakeNewsNet. These labels can have biases based on the values of those who label them. If the training data has a bias towards a political party or geographic region, the result will show that bias. This means honest opinions from communities might be misclassified. A lie from a community might not be detected if it is well-crafted. The labels from annotators with their set of values can affect the model's output. So, we have to be careful when using these datasets to train our models. The model's result will reflect the values in the training data. If the data is not diverse, the model may not work fairly.

The Black-Box Problem (Explainability):

The Black-Box Problem is one of the significant problems related to deep learning systems such as BERT. These models are capable of giving very precise results, but they do not offer any form of explanation as to the way they got them. There are millions of weight parameters and attention heads in each of the models. The general logic applied in coming up with predictions cannot be understood by a human being. This black-box property leads to the mistrust of journalists and fact-checkers, as it happens with editorial consumption of the model output.

Adversarial Evolution and Evolving Narratives:

News is a harsh environment. It resembles a battle between people who create news and people who try to catch them. As we become more successful at locating news, those who compose it discover new ways to trick us. They cease using self-evident words, such as those that attempt to prompt us to click on something. This makes it very difficult to cope with them. They never cease inventing tales and novel technologies to fabricate fake news. We even have computers that are able to write news and videos that look real. This implies that we are forced to continue teaching our computers to locate the news, or they will not be in a position to work. We must continue updating them so that they can be aware of the tricks through which people are still making fake news. The problem of fake news is increasing and getting more difficult to solve since the people who constitute it are never sticking to the same strategies.

Future Work and Conclusion:

Strategic Research Directions:

To address the aforementioned challenges, we propose the following research priorities:

Cross-lingual and Multimodal Models: Expanding detection capabilities beyond English-centric datasets and integrating visual/audio cues. as seen in the workauthers [17][18][7] in year 2024 to combat deepfakes.

Real-time Scalability: Optimizing the computational overhead of transformer models for real-time deployment on high-velocity social media feeds.

Explainable AI (XAI): Developing attention-based visualization tools that allow human moderators to see which specific phrases or context clues triggered a "fake" classification.

Conclusion Remarks:

The digital era faces rapid dissemination of misinformation. Research shows that the usual ways of using machines to learn and find this false information are a start. Using more advanced methods like deep learning will make it even better at finding this information. These advanced methods look at the context of the information from all sides. The tool called BERT is very good at understanding relationships between words, and it is accurate most of the time. For example, it is 92.5% accurate, which is really good for finding types of fraud. However, BERT still has some issues with being transparent and not being biased. This is a known problem. New developments in tools that process language and architectures that use transformers will help create a system that gives people reliable information from all over the world. This will reduce the risks that false information poses to the community.

References:

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *arXiv:1708.01967*, 2017, [Online]. Available: https://arxiv.org/abs/1708.01967

[2] Y. Yan, P. Zheng, and Y. Wang, "Contrastive Learning-Driven Fake News Detection: Preserving Semantics, Unveiling Distortions," *IEEE Trans. neural networks Learn. Syst.*, vol. PP, 2025, doi: 10.1109/TNNLS.2025.3634147.

[3] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea, "Automatic Detection of Fake News," *arXiv:1708.07104*, 2017, [Online]. Available: https://arxiv.org/abs/1708.07104

[4] Y. L. Natali Ruchansky, Sungyong Seo, "CSI: A Hybrid Deep Model for Fake News Detection," *arXiv:1703.06959*, 2017, [Online]. Available: https://arxiv.org/abs/1703.06959

[5] D. Hosseini and R. Jin, "Graph Neural Network based Approach for Rumor Detection on Social Networks," *2023 Int. Conf. Smart Appl. Commun. Networking, SmartNets 2023*, 2023, doi: 10.1109/SmartNets58706.2023.10215926.

[6] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal, "FEVER: a large-scale dataset for Fact Extraction and VERification," *arXiv:1803.05355*, 2018, [Online]. Available: https://arxiv.org/abs/1803.05355

[7] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020, doi: 10.1145/3395046

[8] Xinyi Zhou, Reza Zafarani, "Network-based Fake News Detection: A Pattern-driven Approach," *arXiv:1906.04210*, 2019, [Online]. Available: https://arxiv.org/abs/1906.04210

[9] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Jongwon Choi, "FrePGAN: Robust Deepfake Detection Using Frequency-level Perturbations," *arXiv:2202.03347*, 2022, [Online]. Available: https://arxiv.org/abs/2202.03347

[10] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, "Constructing New Backbone Networks via Space-Frequency Interactive Convolution for Deepfake Detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 401–413, 2024, doi: 10.1109/TIFS.2023.3324739.

[11] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1335–1348, 2023, doi: 10.1109/TIFS.2023.3239223.

[12] Jumana Jouhar, Anju Pratap, "A Transformer-Based Deep Learning Framework with Semantic Encoding and Syntax-Aware LSTM for Fake Electronic News Detection," *Comput. Mater. Contin.*, vol. 86, no. 1, pp. 1–25, 2025, doi: https://doi.org/10.32604/cmc.2025.069327.

[13] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, Kui Ren, "Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection," *arXiv:2403.01786*, 2024, [Online]. Available: https://arxiv.org/abs/2403.01786

[14] Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, "MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection," *arXiv:2211.10996*, 2022, [Online]. Available: https://arxiv.org/abs/2211.10996

[15] Ruiyang Xia, Decheng Liu, Jie Li, Lin Yuan, Nannan Wang, Xinbo Gao, "MMNet: Multi-Collaboration and Multi-Supervision Network for Sequential Deepfake Detection," *arXiv:2307.02733*, 2023, [Online]. Available: https://arxiv.org/abs/2307.02733

[16] N. Vaswani, A., Shazeer, N., Parmar, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

[17] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, "Where Deepfakes Gaze at? Spatial-Temporal Gaze Inconsistency Analysis for Video Face Forgery Detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 4507–4517, 2024, doi: 10.1109/TIFS.2024.3381823.

[18] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "PVASS-MDD: Predictive Visual-Audio Alignment Self-Supervision for Multimodal Deepfake Detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 6926–6936, 2024, doi: 10.1109/TCSVT.2023.3309899.