UDC 004.056.55:004.312.2 **doi: 10.26906/SUNZ.2025.3.72-75**

S. Datsenko

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

NEURAL ARCHITECTURE COMPARISON FOR FACT VERIFICATION ON FEVER DATASET

Abstract. The exponential growth of misinformation and fake news across digital platforms poses unprecedented challenges to information integrity, requiring sophisticated automated fact-checking systems capable of verifying claims against reliable evidence sources with high accuracy and computational efficiency. This study **aims** to evaluate and compare four hybrid neural architectures (BiLSTM-CNN, BiLSTM-RNN, BiLSTM-GRU, and BiLSTM-GNN) for automated fact verification using the FEVER dataset, investigating their effectiveness in claim-evidence verification under GPU memory constraints while analyzing training dynamics and generalization capabilities. The following **results** are obtained: The BiLSTM-CNN architecture achieved optimal performance with 79.5% accuracy, 79.5% recall, 77.9% F1-score, and 93.4% AUC-ROC, followed by BiLSTM-GNN (78.9% accuracy, 93.3% AUC-ROC) and BiLSTM-GRU (77.9% accuracy, 92.2% AUC-ROC), while BiLSTM-RNN exhibited catastrophic failure (33.3% accuracy). All successful architectures demonstrated significant overfitting with 15-17% train-validation accuracy gaps, indicating systematic generalization challenges with limited training data (40,000 samples). **Conclusion**. Multi-kernel convolutional feature extraction proves most effective for local pattern recognition in fact verification, while graph-inspired approaches show promising potential for relational reasoning. The consistent overfitting across architectures highlights the critical need for enhanced regularization, data augmentation, and ensemble methods to achieve robust performance in automated fact-checking systems under computational constraints.

Keywords: fact verification, neural networks, FEVER dataset, hybrid architectures, graph neural networks, bidirectional LSTM.

Introduction

The proliferation of misinformation and fake news across digital platforms poses significant threats to public discourse, democratic processes, and societal well-being. Automated fact-checking systems have emerged as critical tools for combating this challenge, employing various computational approaches to verify claims against reliable evidence sources. Current fake news detection methodologies encompass content-based analysis examining linguistic patterns and semantic inconsistencies, social context-based approaches analyzing propagation patterns and user behavior, and hybrid methods combining textual and multimedia evidence. Deep learning architectures, particularly recurrent neural networks (RNNs), convolutional neural networks (CNNs), and graph neural networks (GNNs), have demonstrated substantial effectiveness in identifying deceptive content through sophisticated pattern recognition and feature extraction capabilities. The FEVER (Fact Extraction and VERification) dataset represents a pivotal benchmark in this domain, providing a standardized framework for evaluating automated verification systems against evidence retrieved from Wikipedia. This work contributes to the advancement of fact verification research by conducting a comprehensive comparison of hybrid neural architectures, combining bidirectional LSTM processing with CNN, RNN, GRU, and graph-based components, thereby exploring the effectiveness of different neural paradigms for claim verification tasks under computational constraints.

Recent advances in automated fact verification have been driven by large-scale datasets like FEVER [1] and sophisticated neural architectures. The FEVER dataset, introduced by Thorne et al. [1], established a benchmark containing 185,445 claims verified against Wikipedia, achieving 0.6841 Fleiss κ inter-annotator agreement. Early baseline systems achieved only 31.87% accuracy with correct evidence and 50.91% without evidence selection [1], highlighting the task's complexity.

Neural Semantic Matching Approaches. Nie et al. [2] introduced neural semantic matching networks that jointly conduct document retrieval, sentence selection, and claim verification, achieving state-of-the-art results by integrating semantic relatedness scores and WordNet features. This work demonstrated the importance of end-to-end optimization across all verification stages.

RNN and CNN-Based Approaches. Recurrent and convolutional neural networks have formed the backbone of many fact verification systems. Research shows that bidirectional LSTM models achieve superior performance over unidirectional approaches and vanilla RNNs [3, 4]. Sastrawan et al. [4] demonstrated that Bidirectional LSTM outperformed CNN and ResNet architectures across multiple fake news datasets. The effectiveness of CNN-LSTM hybrid models has been validated across different domains [5, 6], with architectures combining convolutional feature extraction and LSTM temporal modeling achieving high accuracy rates (96-98%) in news verification tasks.

Hybrid Neural Architectures. Recent work has focused on combining multiple neural paradigms for improved performance. Hybrid CNN-RNN models [7] demonstrate superior results by leveraging CNNs' parallel processing capabilities for feature extraction alongside RNNs' sequential modeling strength. Research indicates that ensemble methods combining different neural architectures can significantly outperform individual models [8, 9], with voting mechanisms and stacking approaches achieving substantial accuracy improvements.

Graph-Based Reasoning. Graph neural networks have emerged as powerful tools for fact verification.

Velickovic et al. [10] introduced Graph Attention Networks (GATs), enabling nodes to attend over neighborhood features without costly matrix operations. Recent advances include GATv2 [11], which addresses static attention limitations in original GAT architectures. Applied to fact verification, Wu et al. [12] developed evidence-aware hierarchical interactive attention networks, while Xu et al. [13] proposed GET framework using graph-based semantic structure mining to address scattered information integration challenges.

Attention Mechanism Evolution: The evolution of attention mechanisms has significantly impacted fact verification. Comprehensive surveys [14, 15] reveal three developmental stages: graph recurrent attention networks, graph attention networks, and graph transformers. Recent work has focused on improving attention expressiveness [11] and developing multi-modal attention for claim verification [16].

Given the promising developments in hybrid neural architectures for fact verification, this research aims to systematically evaluate and compare four distinct hybrid approaches that combine bidirectional LSTM with different complementary components: CNN, RNN, GRU, and GNN. The study seeks to identify the most effective architectural combinations for automated fact verification under computational constraints, while providing insights into training dynamics, generalization capabilities, and the underlying factors that contribute to successful claim-evidence verification performance.

Task Solution

FEVER Dataset. The Fact Extraction and VERification (FEVER) dataset [1] provides a comprehensive benchmark for automated fact-checking, containing human-generated claims paired with Wikipedia evidence. Claims are annotated with three labels: SUPPORTS, REFUTES, or NOT ENOUGH INFO. Our preprocessing combines claims and evidence using '[SEP]' separator, creating unified sequences processed by Keras Tokenizer with 8,000-word vocabulary and 100-token maximum length. Training data is stratified-sampled to 40,000 examples for 4GB GPU constraints, with balanced class weighting addressing label imbalance. Research indicates that 31.75% of claims require multiple evidence sentences, while 16.82% need multi-sentence composition and 12.15% require cross-document evidence [1]. This complexity necessitates sophisticated reasoning architectures capable of handling multi-hop inference and cross-document relationships.

BiLSTM-CNN Architecture. This hybrid design integrates bidirectional sequential processing with multi-scale convolutional feature extraction, inspired by successful CNN-LSTM combinations in fake news detection [5, 6]. The architecture begins with 128-dimensional embeddings, followed by bidirectional LSTM (64×2 units) with dropout regularization (0.2). Three parallel 1D convolutional branches (kernel sizes 3, 4, 5, each 64 filters) extract unigram, bigram, and trigram patterns essential for factual inconsistency detection. Features are concatenated and processed through global max pooling before classification via dense layers (128→64→3 units).

BiLSTM-RNN Architecture. A dual-recurrent design combining bidirectional LSTM with SimpleRNN for computational efficiency. Following identical embedding and BiLSTM processing, a SimpleRNN layer (64 units) provides additional sequential modeling with reduced overhead. Batch normalization layers stabilize training between recurrent components. This streamlined approach offers faster computation while maintaining temporal dependency modeling, following principles from ensemble neural networks.

BiLSTM-GRU Architecture. This architecture leverages complementary strengths of bidirectional LSTM and GRU for optimized sequence processing. The GRU component (64 units) employs simplified gating mechanisms combining forget and input gates, reducing parameters while maintaining gradient flow. The reset and update gates enable selective information retention, addressing vanishing gradients in deep recurrent networks.

BiLSTM-GNN Architecture. A novel graph-inspired approach treating sequences as graphs where words represent nodes, motivated by recent advances in graph attention networks [10,11]. Following BiLSTM processing, node feature transformation (1D conv, 64 filters, kernel=1) creates graph abstractions. Three parallel aggregation operations (kernels 1, 3, 5, each 32 filters) simulate GNN message-passing at multiple scales, inspired by evidence-aware graph networks [12, 13]. Concatenated features undergo global max pooling for graph-level representation, enabling complex relational reasoning for claim-evidence verification.

Results Analysis

Performance Overview. The BiLSTM-CNN architecture achieved optimal performance with 79.5% accuracy, 79.5% recall, 77.9% F1-score, and 93.4% AUC-ROC. BiLSTM-GNN demonstrated competitive results (78.9% accuracy, 93.3% AUC-ROC), while BiLSTM-GRU achieved respectable performance (77.9% accuracy, 92.2% AUC-ROC). BiLSTM-RNN exhibited catastrophic failure (33.3% accuracy), performing at random chance levels. These results align with findings from hybrid neural approaches [7], which demonstrate that combining multiple neural paradigms can achieve superior performance compared to single-architecture models. However, our results also confirm that not all hybrid combinations are beneficial, as evidenced by the BiLSTM-RNN failure.

Training Dynamics. BiLSTM-CNN showed smooth convergence with training accuracy improving from 85% to 95% while validation remained stable around 80%. BiLSTM-GNN exhibited similar robust optimization. BiLSTM-GRU displayed initial instability with dramatic accuracy jumps around epoch 2, suggesting initialization sensitivity. BiLSTM-RNN's flat curves confirmed complete optimization failure.

Architectural Insights. CNN's superior performance validates multi-kernel convolution's importance for local pattern recognition in fact verification, consistent with hybrid CNN-LSTM approaches [5, 6]. The competitive GNN performance demonstrates graph-inspired

neighborhood aggregation's potential for textual reasoning, aligning with recent graph-based fact verification systems [12, 13] (Fig. 1). GRU's reasonable results confirm sophisticated gating benefits, though combination with bidirectional LSTM introduces optimization complexity.

Overfitting Evidence. All successful architectures exhibit substantial train-validation gaps (15-17%): BiLSTM-CNN (95% vs 80%), BiLSTM-GRU (93% vs 78%), BiLSTM-GNN (96% vs 80%). This consistent disparity indicates systematic overfitting across architectures, a common challenge in fact verification systems with limited training data [3, 4] (Fig. 2).

Root Causes are:

1) Dataset Limitations. The 40,000-sample constraint severely limits pattern diversity for complex architectures with 650K-850K parameters. Models memorize specific claim-evidence combinations rather than learning generalizable verification reasoning, similar to challenges identified in CNN-RNN hybrid approaches [7].

2) Model Complexity. High parameter-to-sample ratios enable perfect training data fitting while failing validation generalization. Current regularization (dropout 0.2-0.5, batch normalization) proves insufficient for model complexity.

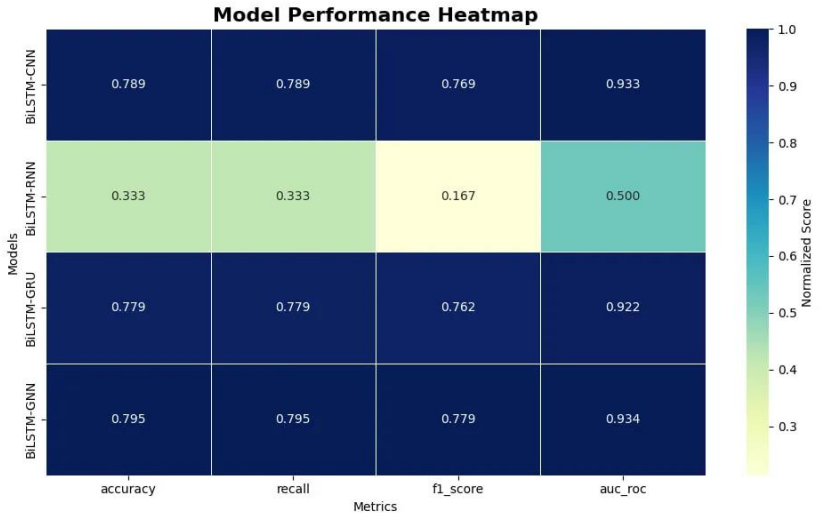

Fig. 1. Performance heatmap of all models

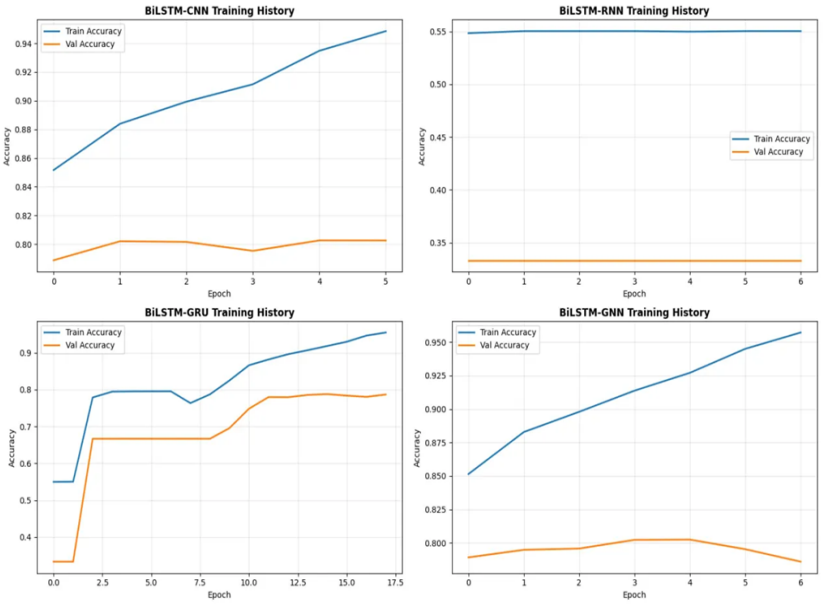

Fig. 2. Training graphs of all models

Conclusions

The consistent 15-17% overfitting gap suggests fact verification under hardware constraints requires fundamentally different approaches: few-shot learning methods, meta-learning for rapid adaptation, or knowledge-augmented models reducing training data dependence. Recent hybrid approaches [7] achieve superior performance through architectural diversity, suggesting future work should prioritize ensemble strategies.

The superior CNN performance validates local pattern recognition importance, while competitive GNN results demonstrate graph-inspired reasoning potential, consistent with recent evidence-aware approaches [12, 13]. Future research should bridge the performance gap through ensemble methods [8,9] and advanced hybrid architectures while exploring graph-based reasoning capabilities for automated fact-checking systems.

Several advanced approaches can be implemented for future work. Ensemble Methods: Implement ensemble learning approaches [8, 9] combining predictions from multiple architectures. Voting mechanisms, stacking, or bagging could leverage the complementary strengths of different neural paradigms while reducing variance. Enhanced Graph Modeling: Extend GNN approaches with explicit graph construction incorporating named entity relationships, following evidence-aware graph networks [12, 13]. Multi-modal graph attention [16] could integrate diverse evidence types.

Hybrid Architecture Optimization: Explore alternative hybrid combinations inspired by successful CNN-LSTM models [5,6], potentially investigating CNN-GRU or GNN-CNN combinations for improved performance.

REFERENCES

1. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, New Orleans, Louisiana, June 2018; pp. 809–819. doi: https://doi.org/10.18653/v1/N18-1074.
2. Nie, Y.; Chen, H.; Bansal, M. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, July 2019; pp. 6859–6866. doi: https://doi.org/10.1609/aaai.v33i01.33016859.
3. Abualigah, L.; Al-Ajlouni, Y.Y.; Daoud, M.S.; Altalhi, M.; Migdady, H. Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe. Soc. Netw. Anal. Min. 2024, 14, 40. doi: https://doi.org/10.1007/s13278-024-01198-w.
4. Sastrawan, I.K.; Bayupati, I.P.A.; Arsa, D.M.S. Detection of fake news using deep learning CNN–RNN based methods. ICT Express 2021, 8, 396–408. doi: https://doi.org/10.1016/j.icte.2021.10.003.
5. Dev, D. G.; Bhatnagar, V.; Bhati, B. S.; Gupta, M.; Nanthaamornphong, A. LSTMCNN: A hybrid machine learning model to unmask fake news. Heliyon. 2024, 3, 10. doi: https://doi.org/10.1016/j.heliyon.2024.e25244.
6. Ajik, E. D.; Obuandike, G. N.; Echobu, F. O. Fake News Detection Using Optimized CNN and LSTM Techniques. Journal of Information Systems and Informatics. 2023, 3, 5, 1044-1057. doi: https://doi.org/10.51519/journalisi.v5i3.548.
7. Nasir, J. A.; Khan, O. S.; Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights. 2021, 1, 1, 100007. doi: https://doi.org/10.1016/j.jjimei.2020.100007.
8. Almandouh, E.; Alrahmawy, M.F.; Eisa, M. et al. Ensemble based high performance deep learning models for fake news detection. Sci. Rep. 2024, 14, 26591. doi: https://doi.org/10.1038/s41598-024-76286-0.
9. Mohammed, A.; Kora R. A comprehensive review on ensemble deep learning: Opportunities and challenges. J. King Saud Univ.-Comput. Inf. Sci. 2023, 2, 35, 757–774. doi: https://doi.org/10.1016/j.jksuci.2023.01.014.
10. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning Representations, 2018; pp. 1–12. url: https://arxiv.org/abs/1710.10903v3.
11. Brody, S.; Alon, U.; Yahav, E. How Attentive are Graph Attention Networks? In Proceedings of the International Conference on Learning Representations, May 2021. url: https://arxiv.org/abs/2105.14491.
12. Wu, L.; Rao, Y.; Yang, X.; Wang, W.; Nazir, A. Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, July 2020; pp. 1388–1394. doi: https://doi.org/10.24963/ijcai.2020/193.
13. Xu, W.; Wu, J.; Liu, Q.; Wu, S.; Wang, L. Evidence-aware Fake News Detection with Graph Neural Networks. In Proceedings of the ACM Web Conference 2022, April 2022; pp. 2501–2510. doi: https://doi.org/10.1145/3485447.3512122.
14. Sun, C.; Li, C.; Lin, X. et al. Attention-based Graph Neural Networks: A Survey. Artif. Intell. Rev. 2023, 56, 2263–2310. doi: https://doi.org/10.1007/s10462-023-10577-2.
15. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. IEEE Trans. Neural Netw. Learn. Syst. 2021, 32, 4–24. doi: https://doi.org/10.1109/TNNLS.2020.2978386.
16. Luu, S. T.; Vo, T.; Nguyen, LM. MCVE: Multimodal Claim Verification and Explanation Framework for Fact-Checking System. Multimed. Syst. 2025, 31, 242. doi: https://doi.org/10.1007/s00530-025-01804-7.

Received (Надійшла) 15.05.2025
Accepted for publication (Прийнята до друку) 06.08.2025

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Даценко Сергій Сергійович – аспірант кафедри комп'ютерної інженерії, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна;
Serhii Datsenko – PhD student, Department Computer Engineering, National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine;
e-mail: sergdacenko@gmail.com; ORCID Author ID: https://orcid.org/0000-0001-9514-0433;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57218596147&origin=resultslist.

Порівняння нейронних архітектур для перевірки фактів на наборі даних FEVER

С. С. Даценко

Анотація. Експоненціальне зростання дезінформації та фейкових новин на цифрових платформах ставить безпрецедентні виклики перед цілісністю інформації, що вимагає створення складних автоматизованих систем перевірки фактів, здатних з високою точністю та обчислювальною ефективністю перевіряти твердження на основі надійних джерел доказів. Це дослідження має на меті оцінити та порівняти чотири гібридні нейронні архітектури (BiLSTM-CNN, BiLSTM-RNN, BiLSTM-GRU та BiLSTM-GNN) для автоматизованої перевірки фактів за допомогою набору даних FEVER, досліджуючи їх ефективність у перевірці тверджень та доказів в умовах обмеженої пам'яті GPU, одночасно аналізуючи динаміку навчання та можливості узагальнення. Отримано наступні результати: архітектура BiLSTM-CNN досягла оптимальної продуктивності з точністю 79,5%, відтворюваністю 79,5%, 77,9% F1-показником та 93,4% AUC-ROC, за нею йдуть BiLSTM-GNN (78,9% точність, 93,3% AUC-ROC) та BiLSTM-GRU (77,9% точність, 92,2% AUC-ROC), тоді як BiLSTM-RNN продемонструвала катастрофічну невдачу (33,3% точність). Усі успішні архітектури продемонстрували значне перенавчання з розривом у точності тренування-валідації 15-17%, що вказує на системні проблеми узагальнення з обмеженими даними для навчання (40 000 зразків). **Висновок.** Витяг багатоядерних конволюційних ознак виявляється найефективнішим для розпізнавання локальних візерунків у перевірці фактів, тоді як підходи, натхненні графами, демонструють багатообіцяючий потенціал для реляційного міркування. Постійне перенавчання в різних архітектурах підкреслює гостру необхідність вдосконалення методів регуляризації, збільшення обсягу даних та ансамблевих методів для досягнення надійної роботи автоматизованих систем перевірки фактів в умовах обчислювальних обмежень.

Ключові слова: перевірка фактів, нейронні мережі, набір даних FEVER, гібридні архітектури, графічні нейронні мережі, двонаправлений LSTM.