Comparative Analysis of Deep Learning Models for Fake Tweet Detection

Article · February 2025

Felix Chad
Ekiti State University

Comparative Analysis of Deep Learning Models for Fake Tweet Detection

Author: Felix Chad

Date: 26^{TH} Feb 2025

Abstract:

With the rise of social media platforms, the spread of misinformation, including fake news and malicious content, has become a significant concern. Twitter, as one of the largest microblogging platforms, is a prime target for the dissemination of fake tweets. Detecting such content is a complex task that requires effective methods for classifying tweets based on their authenticity. This paper presents a comparative analysis of various deep learning models for fake tweet detection, focusing on their performance in terms of accuracy, precision, recall, and F1 score. The study evaluates models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models, specifically BERT and RoBERTa. A detailed examination of the data preprocessing, feature extraction, and model architectures is conducted to identify key factors influencing the detection accuracy. The results demonstrate the strengths and weaknesses of each model, with Transformer-based models showing superior performance in capturing contextual relationships within tweet text. The paper also explores the challenges of handling imbalanced datasets, varying tweet styles, and the rapid evolution of language used in fake content. Finally, recommendations are provided for future research directions, including the integration of multimodal data (text, images, and metadata) and the development of more robust, real-time detection systems.

1. Introduction

A. Background

The proliferation of fake news and misinformation on social media platforms, particularly Twitter, has emerged as a critical global issue.

These platforms have become key sources for information, but they also serve as vehicles for spreading misleading or false content. Fake tweets, whether they involve fabricated news stories, manipulated images, or deceptive narratives, can have far-reaching effects on public opinion, political outcomes, and societal trust. With over 300 million active users globally, Twitter's real-time nature makes it particularly vulnerable to the rapid spread of such harmful content. As the volume of tweets grows exponentially, manual detection becomes increasingly inefficient, leading to the need for automated systems capable of identifying fake content in real time.

Deep learning has shown significant promise in addressing the problem of fake tweet detection due to its ability to learn complex patterns from large datasets and its adaptability to various forms of textual input. Recent advancements in natural language processing (NLP) and deep learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models, have led to improved accuracy in classifying text-based content. However, challenges such as contextual understanding, domain-specific language, and data imbalance persist, making the detection of fake tweets a highly complex and ongoing area of research.

B. Objective

The primary objective of this study is to conduct a comprehensive comparative analysis of deep learning models for fake tweet detection. The study aims to evaluate the effectiveness of various models, including CNNs, RNNs, LSTMs, and Transformer-based architectures like BERT and RoBERTa, in accurately identifying fake content on Twitter. By assessing model performance in terms of accuracy, precision, recall, and F1 score, the paper seeks to highlight the strengths and weaknesses of each approach. Furthermore, the study explores the role of data preprocessing, feature extraction, and model architecture in influencing the detection outcomes. Ultimately, this research aims to contribute to the ongoing efforts in developing robust, scalable, and real-time systems for combating misinformation on social media platforms.

2. Overview of Fake Tweet Detection

A. Definition of Fake Tweets

Fake tweets refer to misleading or fabricated information shared on Twitter, which may include false news, deceptive claims, manipulated images, or malicious narratives. These tweets are often designed to mislead, influence public opinion, or disrupt social discourse.

B. Importance of Fake Tweet Detection

Detecting fake tweets is crucial to maintaining the integrity of information on social media. The rapid spread of false information can lead to harmful societal effects, including public panic, misinformation, and manipulation of political processes. Effective detection helps combat these issues by ensuring that users are exposed to reliable and accurate content.

3. Deep Learning Models for Fake Tweet Detection

A. Convolutional Neural Networks (CNNs)

CNNs are primarily used in image processing but have been successfully applied to text classification tasks, including fake tweet detection. By treating text as a sequence of word embeddings, CNNs learn local patterns and structures in the text, such as phrases or word combinations that may indicate fake content. CNNs are fast and effective for extracting features but may struggle with long-range dependencies in text.

B. Recurrent Neural Networks (RNNs)

RNNs are designed to handle sequential data and are particularly effective for text where the order of words matters. In fake tweet detection, RNNs process the tweet's text one word at a time, maintaining a memory of previous words. While RNNs can capture contextual information, they often face challenges with long-term dependencies due to issues like vanishing gradients.

C. Long Short-Term Memory (LSTM)

LSTMs are an advanced version of RNNs that address the long-term dependency issue by using special gating mechanisms. LSTMs have been widely used in fake tweet detection as they can remember and utilize information from longer text sequences, making them more capable of understanding complex tweet structures and detecting misleading content.

D. Gated Recurrent Units (GRUs)

GRUs are a simplified version of LSTMs, offering similar performance but with fewer parameters, making them more computationally efficient. GRUs use gates to control the flow of information, allowing the model to focus on relevant parts of the input text while mitigating the vanishing gradient problem. GRUs are effective for fake tweet detection, especially in cases with limited data or computational resources.

4. Comparative Performance Evaluation

A. Accuracy

Accuracy measures the overall proportion of correctly classified tweets. While accuracy is an essential metric, it can be misleading when dealing with imbalanced datasets, where fake tweets are less frequent than real ones. Transformer-based models such as BERT and RoBERTa often achieve higher accuracy due to their advanced language understanding and ability to capture contextual nuances, making them ideal for complex fake tweet detection tasks.

B. F1 Score, Precision, and Recall

In the context of fake tweet detection, F1 score, precision, and recall are crucial metrics because they balance the trade-off between correctly identifying fake tweets and minimizing false positives and false negatives:

Precision measures the percentage of fake tweets correctly identified by the model. Higher precision means fewer false positives.
Recall reflects the ability of the model to correctly identify all fake tweets. High recall indicates that most fake tweets are captured, even at the cost of some false positives.
F1 Score is the harmonic mean of precision and recall, offering a balanced evaluation of the model’s performance. Transformer-based models often achieve superior F1 scores due to their ability to understand context and relationships between words, while simpler models like CNNs and RNNs may have lower F1 scores, especially when the dataset is highly imbalanced.

C. Model Training Time and Efficiency

Training time and computational efficiency are key considerations when selecting a model for fake tweet detection. CNNs and GRUs are more computationally efficient and faster to train, as they have fewer parameters and simpler architectures. In contrast, models like LSTMs and Transformers (e.g., BERT and RoBERTa) require more time and computational resources to train due to their complexity and large parameter sets. While Transformer-based models deliver higher accuracy and performance, their resource demands make them less suitable for real-time, resource-constrained environments.

D. Scalability and Generalization

Scalability refers to a model's ability to handle increasing volumes of data, while generalization evaluates how well the model performs on unseen data. Transformer-based models excel at generalization, as they are pre-trained on vast amounts of data and fine-tuned for specific tasks, making them robust to diverse and evolving tweet content. However, their scalability can be challenging due to high computational costs. On the other hand, simpler models like CNNs and GRUs are more scalable and easier to deploy in large-scale, real-time systems, but they may struggle with generalization to varied or unseen content.

5. Results and Discussion

A. Summary of the Comparative Performance of Models

Transformer-based models (BERT, RoBERTa) consistently outperform other deep learning models in terms of accuracy, F1 score, precision, and recall, due to their superior ability to understand context. CNNs and GRUs, while faster and more computationally efficient, show lower performance in handling complex patterns. LSTMs offer a balance, performing well on sequential data but require more resources compared to CNNs and GRUs.

B. Discussion on the Strengths and Weaknesses of Each Model in Real-World Applications

Transformer-based models excel in accuracy and generalization, making them ideal for complex fake tweet detection tasks. However, their high computational demands limit their scalability in real-time applications. CNNs and GRUs are more efficient and scalable but may struggle with nuanced language or longer dependencies. LSTMs are effective for sequential data but also face higher resource demands. In real-world scenarios, the choice of model depends on the trade-off between accuracy and computational efficiency, especially when dealing with large-scale, real-time data.

6. Conclusion

A. Summary of Key Findings

This study highlights the effectiveness of Transformer-based models (BERT, RoBERTa) for fake tweet detection, demonstrating superior performance in accuracy, precision, recall, and F1 score. While CNNs and GRUs are more efficient and scalable, they lag in capturing complex language patterns. LSTMs offer a middle ground but still require significant computational resources. Overall, Transformer models are ideal for high-accuracy tasks, though they come with scalability challenges.

B. Future Directions for Improving Fake Tweet Detection

Future research could focus on developing hybrid models that combine the strengths of different architectures, such as combining CNNs for feature extraction with Transformer-based models for contextual understanding. Transfer learning, where models are pre-trained on large datasets and fine-tuned for fake tweet detection, could improve performance with limited data. Additionally, exploring more robust evaluation metrics that account for the real-world impact of false positives and negatives, such as user trust or social influence, could lead to more effective and reliable detection systems.

Reference:

1. Subramaniam, E. V. D., Srinivasan, K., Qaisar, S. M., & Pławiak, P. (2023). Interoperable IoMT Approach for Remote Diagnosis with Privacy-Preservation Perspective in Edge Systems. *Sensors*, *23*(17), 7474. https://doi.org/10.3390/s23177474

2. Dinesh, S. E. V., & Valarmathi, K. (2020). A novel energy estimation model for constraint based task offloading in mobile cloud computing. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 5477–5486. https://doi.org/10.1007/s12652-020-01903-5

3. Subramaniam, E. V. D., & Krishnasamy, V. (2023). Hybrid Optimal Ensemble SVM Forest Classifier for task offloading in mobile cloud computing. *The Computer Journal*, *67*(4), 1286–1297. https://doi.org/10.1093/comjnl/bxad059

4. Subramaniam, E. V. D., & Krishnasamy, V. (2024). ABES: attention bi-directional ensemble SVM for early detection of brain tumors. *Neural Computing and Applications*, *36*(26), 16179–16193. https://doi.org/10.1007/s00521-024-09688-w

5. Mareeswari, G., & Dinesh, E. V. (2023, March). Deep neural networks based detection and analysis of fake tweets. In *2023 4th International Conference on Signal Processing and Communication (ICSPC)* (pp. 56-61). IEEE.

6. Rathinam, V., A, R., K, V., & S, E. V. D. (2025). S3-GHOSTNET: skin disease detection via Sand Cat Swarm Optimization deep GHOSTNET with facial images. *Australian Journal of Electrical & Electronics Engineering*, 1–11. https://doi.org/10.1080/1448837x.2025.2454803