National College of Ireland

A Comparative Study on the Impact of Input Length on Transformer Model Performance in Misinformation Classification

MSc Research Project
MSc in Artificial Intelligence

Mohammad Shehnaj
Student ID: 23305762

School of Computing
National College of Ireland

Supervisor: Abdul Razzaq

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Mohammad Shehnaj
Student ID:	23305762
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Abdul Razzaq
Submission Due Date:	15/09/2025
Project Title:	A Comparative Study on the Impact of Input Length on Transformer Model Performance in Misinformation Classification
Word Count:	8579 words
Page Count:	29

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mohammad Shehnaj
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study on the Impact of Input Length on Transformer Model Performance in Misinformation Classification

Mohammad Shehnaj
23305762

Abstract

Context – The rise of digital platforms has accelerated the spread of fake news undermining trust and influencing public opinion. Transformer-based models are widely used for detection, yet the impact input text length ranging from short headlines to full articles in shaping their performance has only been partially investigated in prior studies.

Objective — This study evaluates how varying input lengths (short, medium, long, hybrid) affect the classification performance and computational efficiency of five transformer models: BERT, RoBERTa, Longformer, BigBird and LLaMA to establish benchmarks that balance accuracy, resource use and inference speed for real-world misinformation detection.

Method – Using a balanced dataset of 157,690 news items from multiple benchmark sources, all samples were preprocessed, tokenized and categorized into length bins before fine-tuning each model separately.

Results – Input length affected performance, repeated-measures ANOVA showed significant effects for RoBERTa, BERT and BigBird ($p < 0.05$) while most pairwise tests were non-significant. Longer inputs generally achieved higher F1-scores with RoBERTa, BigBird and Longformer exceeding 0.98 while shorter inputs lowered accuracy for most models mainly LLaMA ($F1 < 0.71$). Medium and hybrid lengths offered a balanced trade-off with BERT delivering competitive accuracy alongside the fastest inference time (15.69 ms per sample) and BigBird with lowest memory usage. RoBERTa maintained strong performance across all lengths whereas LLaMA consistently underperformed indicating limited ability to leverage extended context. These results highlight input length as a critical factor in balancing accuracy and computational efficiency in transformer-based fake news detection.

Conclusion – These findings provide length-aware benchmarks that guide the selection of transformer architectures and input strategies enabling a balanced trade-off between accuracy, efficiency and deployment feasibility in real-world misinformation detection systems. These findings are supported with a token attribution analysis that highlights how predictive cues concentrate in short texts and diffuse in longer articles.

1 Introduction

The rapid expansion of digital platforms and social media has dramatically altered how information is produced, consumed and circulated enabling instant communication but

also accelerating the dissemination of misinformation Zhou and Zafarani (2020). This transformation has improved accessibility and speed of communication and has also enabled the widespread dissemination of fake news intentionally that undermines public trust, manipulates opinion and threatens democratic processes lazer2018science. Empirical evidence shows that fake news can spread more rapidly and widely than factual information, amplifying its potential impact on politics, health and societal trust vosoughi2018spread,shu2017fake,allcott2017social. Consequently, recent state-of-the-art surveys have highlighted the necessity for developing automated, scalable and accurate fake news detection systems to address the growing challenges of misinformation in contemporary information ecosystems Zhou2019,Shu2019.

In recent years, transformer-based models like BERT Devlin et al. (2019), RoBERTaLiu et al. (2019) and LLaMAChen et al. (2023) have revolutionized natural language processing (NLP) by capturing deep contextual relationships through self-attention mechanisms. These models have been demonstrated to achieve state-of-the-art performance in fake news detection across various benchmark datasets. BERT-based frameworks have outperformed traditional classifiers on LIAR and FNC datasets Yang et al. (2022), RoBERTa variants have achieved superior accuracy in large-scale misinformation detection tasks Rodrigues et al. (2024) and LightGBM transformer hybrids have shown enhanced classification precision and generalization Abdulaziz et al. (2024). However, most existing studies focus primarily on model architecture or dataset type neglecting a critical variable, the length of the input text despite its potential influence on model performance Zhou and Zafarani (2020); Bozic and Anastasopoulos (2023). Although some studies have incorporated varying textual granularities such as headlines, summaries or full articles as part of their dataset structure Shu et al. (2020); Zhou and Zafarani (2020), these works typically did not conduct a systematic investigation into how input content length impacts classification accuracy, computational resource usage or model decision behavior. Long inputs can carry more context, they also impose computational burdens and may introduce noise. Short inputs are efficient but may lack discriminative features. This trade-off remains partially investigated in prior studies of fake news detection systems.

This research addresses this gap by evaluating the impact of input text length on transformer-based models for fake news detection. The study categorizes inputs into four distinct groups - short, medium, long and hybrid and assesses model performance across these categories. It includes both standard models (BERT, RoBERTa) and long-sequence models (Longformer, BigBird, LLaMA) to offer a holistic comparison. The objective of this study is not to introduce a novel model but to present an evidence-based evaluation of how different input content lengths influence model performance and computational efficiency across varied real-world scenarios.

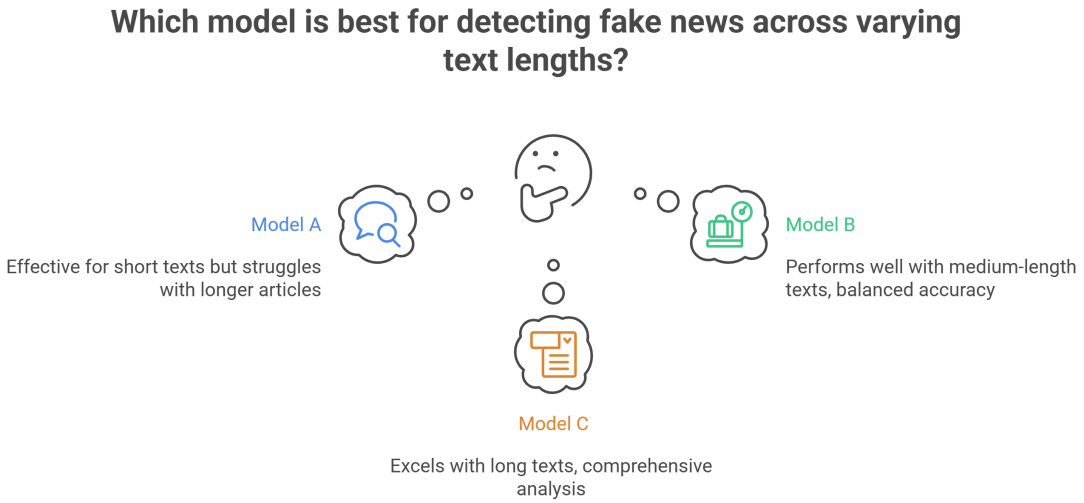


Figure 1: Fake news detection problem across varying input text lengths.

The primary objective of this research is to determine how varying input lengths affect both the *classification effectiveness* and *computational efficiency* of transformer-based models. Accordingly, the research addresses the following questions:

- **Research Question 1:** How does input text length (short, medium, long and hybrid) influence the fake news detection performance of transformer-based models? This study assumes that input characteristics which is text length play a significant role in model effectiveness. Short inputs may lack sufficient context for nuanced classification, medium-length summaries might offer an optimal balance, long inputs could introduce noise despite providing richer context and hybrid inputs may combine the strengths and weaknesses of all types.
- **Research Question 2:** What trade-offs exist between classification accuracy and computational cost across different models and input lengths? Models delivering higher accuracy are anticipated to demand more computational resources and more efficient architectures may offer faster processing at the expense of detection quality. Identifying this balance is critical for selecting an appropriate model in real-world applications.

Key Contributions:

- Empirically demonstrate that input text length significantly influences fake news detection performance and computational efficiency.
- Reveal the trade-offs between classification effectiveness and computational efficiency showing that models with higher accuracy often incur higher resource costs.
- Provide statistically validated evidence that input length selection can be as impactful as model choice highlighting a critical yet partailly explored factor in transformer-based fake news detection systems.

This thesis proceeds with a review of Section 2 related work in transformer-based fake news detection, Section 3 details the methodology and system design, Section 4 outlines the implementation process, Section 5 presents experimental results with comparative analysis and Section 6 concludes with key findings, limitations and future research directions.

2 Related Work

The proliferation of fake news poses a substantial threat to public trust mainly in critical domains such as health, politics and global security. Although recent advancements in transformer-based language models have enabled remarkable progress in fake news detection, several persistent challenges remain most notably in the scalability of these models across varying input lengths. This section presents a critical and comparative review of existing techniques, highlighting their strengths, limitations and relevance to our proposed research focus on input-length-aware evaluation.

2.1 Terminology and Taxonomy

To contextualize the work, we define and classify key concepts related to fake news detection using transformer-based models:

- **Fake News** refers to intentionally false or misleading information presented as legitimate news.
- **Input Length** in this study is categorised as short ($\leq$145 tokens), medium (146–387), long (>387) and hybrid (mixed lengths).
- **Transformer Models** such as BERT, RoBERTa, Longformer, BigBird and LLaMA leverage self-attention to capture contextual relationships in sequences.
- **Explainability** refers to the interpretability of model decisions enabling stakeholders to understand and trust the outputs. This work focuses solely on textual inputs.

This taxonomy lays the foundation for analyzing the strengths and limitations of existing methods discussed in subsequent sections.

2.2 Overview of Fake News Detection Techniques

The proliferation of fake news has inspired significant research into automated detection techniques leveraging Natural Language Processing (NLP) and Deep Learning (DL) methods. Traditional ML-based classifiers like Naive Bayes and SVM initially showed promise on short texts (Zhou and Zafarani; 2020) but their reliance on hand-crafted features limited scalability across domains. Early methods often relied on traditional machine learning or deep learning architectures such as CNNs and LSTMs (Vijayaraghavan and Vosoughi; 2017) before the advent of transformer-based models. Deep learning architectures such as CNNs and RNNs brought notable performance improvements due to their hierarchical feature learning capabilities (Wang; 2017) and hybrid variants combining LSTM and BERT have further enhanced detection accuracy in certain datasets

(Singh et al.; 2021). However, these models often struggle with long-range dependencies and context preservation especially when dealing with full-length news articles or posts (Ruchansky et al.; 2017).

The advent of transformer-based models has revolutionized fake news detection by enabling attention mechanisms that capture contextual relationships across entire sequences. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al.; 2019), pre-trained on large corpora has become a cornerstone of many fake news detection systems achieving F1-scores close to 88–90% across multiple datasets (Kaliyar et al.; 2020). RoBERTa (Liu et al.; 2019), a robustly optimized variant further improved pretraining dynamics and generalization. Standard transformer models are often limited to short input sequences (typically up to 512 tokens), which poses a challenge for processing long or hybrid texts.

2.3 Input Length and Its Influence on Model Performance

Several studies have explored the correlation between input text length and detection accuracy. The paper 'Same Task, More Tokens' (Bozic and Anastasopoulos; 2023) provides empirical evidence that as input length increases, models trained on shorter sequences tend to underperform due to truncation and incomplete context. This is particularly critical in fake news where factual consistency often requires evaluating content across titles, body and metadata (Shu et al.; 2020).

Models such as Longformer (Beltagy et al.; 2020) and BigBird (Zaheer et al.; 2020) were introduced to address this limitation utilizing sparse attention mechanisms to scale up to 4k+ tokens. These architectures are ideal for processing full articles but come with increased computational demands and complex fine-tuning requirements (Sun et al.; 2022). Models like Dual-BERT (Singhal et al.; 2021) and HyProBERT (Chen et al.; 2023) take a hybrid approach by combining title and content streams fusing deep contextual layers to maintain both brevity and semantics. These models have demonstrated improved F1 scores ($\sim$90–91%) on medium-length inputs without requiring substantial architectural tune-ups.

Despite these innovations, the literature still lacks a structured comparative analysis of how input length categories (short, medium, long and hybrid) directly impact classification performance. (Guo et al.; 2021) proposed fusing social context with textual features demonstrating improved robustness across varying news domains. Most works optimize architecture or loss functions but do not measure the role of input content length as a variable of interest. This motivates our investigation.

2.4 Transformer Models Applied in Fake News Detection

Among the earliest and most influential transformer-based models, BERT (Devlin et al.; 2019) has been widely adapted due to its bidirectional attention and generalization power. It was extended by RoBERTa (Liu et al.; 2019) which achieved better accuracy through longer training and removal of the next sentence prediction task. HyProBERT (Kaliyar et al.; 2020), a recent innovation enhances semantic richness using hyper-contextual layers improving explainability and performance.

Dual-BERT (Singhal et al.; 2021) effectively processes headline and article pairs using dual encoders and achieves strong results especially on datasets where headline-body stance is important like FNC-1. BERT-LSTM and CNN-BERT hybrids (Shu et al.;

2020; Sanh et al.; 2021) attempt to fuse sequential or spatial features into transformer embeddings yielding mixed success depending on input structure.

More recent works employ hybrid models that combine BERT with external classifiers such as LightGBM (Kaliyar et al.; 2020), or with retrieval augmentation as VeraCT (Nguyen et al.; 2020; Fang et al.; 2021) which supports explainability by tracing evidence. Multimodal and continual learning approaches have also emerged (Sun et al.; 2022) but these are beyond the scope of this thesis focused on text-only inputs.

2.5 Handling Long and Hybrid Inputs with Advanced Transformers

Models such as Longformer (Beltagy et al.; 2020) and BigBird (Zaheer et al.; 2020) have extended the capability of transformers to long inputs using global and sparse attention supporting sequences up to 4k–8k tokens. These architectures are particularly suited for full-article analysis achieving F1 scores up to ∼92% (Sun et al.; 2022). However, they remain resource-intensive and require significant adaptation to fit fake news datasets which are typically optimized for BERT-style models.

LLaMA (Chen et al.; 2023), a decoder-only model by Meta, though primarily known for generative tasks has demonstrated strong performance on classification through fine-tuning and quantized variants. Recent studies show that LLaMA-7B and its distilled versions can match BERT-family accuracy while supporting longer contexts and being more memory-efficient (Chen et al.; 2023). These properties make it a suitable candidate for our comparison especially in the hybrid input category.

Multiscale Transformers (Sanh et al.; 2021), trained to handle variable-length inputs and Ensemble Architectures (Zhou et al.; 2020; Kim and Lee; 2023) using CNN, LSTM and transformer fusion have also shown promise. Nevertheless, their generalization to multiple input lengths hasn't been explicitly benchmarked.

2.6 Critical Gaps and Justification for This Study

While substantial advancements have been made several critical gaps remain. First, there is a lack of comparative studies that control for input length across transformer-based architectures. Most studies test models on default token limits or preprocessed datasets without evaluating truncation effects or full-input utilization. Second, long-sequence transformers are partially explored in fake news classification with few works using Longformer or BigBird in this domain due to infrastructure demands. Third, models like LLaMA which are powerful have not been comparatively evaluated against BERT-family models on real-world fake news datasets using controlled input length categories.

Moreover, explainability and evaluation metrics are often overlooked in favor of raw performance scores. Recent work has also demonstrated how transformer architectures can be adapted for explainable fake news detection, balancing transparency with predictive performance (Zhang et al.; 2021). Our study addresses these gaps by systematically evaluating models across short (≤145 tokens), medium (146–387 tokens), long (¿ 387 tokens) and hybrid (mixed lengths) inputs using both classical and recent transformer models including BERT, RoBERTa, HyProBERT, Longformer, BigBird and LLaMA-7B.

2.7 Comparative Summary of Existing Techniques

Table 1: Comparative Summary of Empirical Studies on Transformer-Based Fake News Detection Techniques

Model Type)	Input Length	Architecture	F1 Score	Explainability	Notable Feature
BERT (Devlin et al.; 2019)	Short	Transformer	∼88%	Low	Pretrained contextual embeddings
RoBERTa (Liu et al.; 2019)	Short	Optimized BERT	∼90%	Low	Robust pretraining corpus
HyProBERT (Kaliyar et al.; 2021a)	Short	Deep Contextual	∼91%	Medium	Hypercontext layers
Dual BERT (Singhal et al.; 2021)	Medium	Dual-Input BERT	∼90%	Medium	Title + article joint modeling
Longformer (Beltagy et al.; 2020)	Long	Sparse Transformer	∼92%	Low	Long input support (4K tokens)
BigBird (Zaheer et al.; 2020)	Long	Sparse Transformer	∼91%	Low	Global/random attention
LLaMA (AI, 2023; Chen et al.; 2023)	Long	Decoder-only	∼90%	Medium	Efficient with quantization
BERT + LightGBM (Kaliyar et al.; 2021b)	Hybrid	Hybrid	∼88%	Medium	ML classifier over embeddings
Multiscale Trans-former (Sanh et al.; 2021)	Hybrid	Transformer	∼89%	Low	Handles mixed-length inputs
VeraCT (RAG) (Nguyen et al.; 2020)	Hybrid	RAG Transformer	∼87%	High	Retrieval with justification

2.8 Summary and Research Motivation

The literature provides a strong foundation for transformer-based fake news detection with numerous innovations targeting architecture, attention mechanisms and training paradigms. However, the specific influence of input content length remains an underex-

plored dimension. Most studies benchmark models on static or truncated inputs without systematically comparing how models behave across length categories. Furthermore, LLaMA, Longformer and BigBird, despite their capacity for longer sequences are rarely compared against classical transformers in this context.

This thesis addresses this gap by proposing a structured, length-aware evaluation across multiple transformer models, assessing classification performance, computational efficiency and explainability. The findings aim to inform future research on designing fake news detectors that are robust to content length variability.

3 Methodology

This study adopts a structured and comparative empirical methodology to *investigate how input text length affects the performance of transformer-based models in fake news detection*. Input text length is defined as the number of tokens computed after preprocessing and tokenization using each model's respective tokenizer. The experimental pipeline is standardized across all models to ensure fair comparison.

The methodology consists of five phases: (A) Dataset Selection and Characteristics (B) Data Preprocessing and Length Binning (C) Transformer Model Selection (D) Training and Optimization Strategy and (E) Evaluation Framework. Each step is designed to isolate the effect of input length across multiple transformer architectures and input types.

3.1 Datasets Overview

To support reproducibility and meaningful comparison, this study selects four benchmark fake news datasets that have been extensively used in prior research. These datasets offer a representative range of input lengths spanning short claims, full articles and hybrid formats enabling a structured evaluation of model performance across varying content lengths. This selection was also influenced by prior studies reviewed in Section 2, where dataset variety and length diversity were shown to impact model generalization. A summary of dataset characteristics is presented in Table 2.

Table 2: Datasets used in this study

Dataset	**Key Characteristics**
US Elections 2024[1]	25K full-length political news articles. Suitable for evaluating long-text performance.
WELFake Dataset[2]	72K fact-checked articles. Includes a range of short and long texts.
LIAR Dataset[3]	12.8K short political statements, binarized from original 6-label schema.
FNC-1 Dataset[4]	49.9K headline–body pairs. Converted to binary format. Ideal for hybrid-length scenarios.

[1] `https://huggingface.co/datasets/newsmediabias/fake_news_elections2024`
[2] `https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification`
[3] `https://huggingface.co/datasets/ucsbnlp/liar`
[4] `https://github.com/FakeNewsChallenge/fnc-1/tree/master`

Figure 2: End-to-end Structured Methodology Flow from Dataset to Evaluation across Input Lengths

All datasets were cleaned, deduplicated and relabeled (where required) to ensure binary target consistency. The final unified dataset contains 157,690 balanced samples across the two target labels.

3.2 Preprocessing and Length Binning

To ensure consistency across all datasets, a standardized preprocessing pipeline was applied. Each text sample underwent the following steps:

- Conversion of all text to lowercase
- Removal of HTML tags, URLs, emojis, special characters and excess whitespace
- Concatenation of available fields such as title, headline, body and summary into a single input
- Mapping of multi-class labels into binary categories: real or fake
- Tokenization using Hugging Face AutoTokenizer appropriate for each model architecture such as BertTokenizer, RobertaTokenizer and more

The thresholds for length-based segmentation were empirically derived by analyzing the token length distribution across the full dataset. Token count percentiles were examined across the unified dataset to derive meaningful bins, balancing the need for real-world representativeness and training stability. This categorization enables structured evaluation of model sensitivity to input size. After tokenization, the number of tokens was computed for each input and samples were grouped into the following four categories:

- **Short** ($\leq$145 tokens): Typically includes headlines and concise statements
- **Medium** (146–387 tokens): Represents compact articles or partial body content
- **Long** (>387 tokens): Full-length articles or detailed reports
- **Hybrid**: Formed by sampling an equal proportion of short, medium and long samples to simulate real-world variability

All datasets were split into training, validation and test sets using a 70:15:15 ratio while maintaining label balance within each length bin to ensure unbiased performance evaluation.

3.3 Transformer Model Selection

To enable a structured investigation of how input length influences fake news detection, this study evaluates a selection of transformer-based models categorized into two groups: (i) **baseline models** which are widely used in prior studies and constrained to an input limit of 512 tokens and (ii) **long-context models** which are capable of processing extended sequences beyond standard limitations. This categorization ensures coverage of key architectural approaches including encoder-only, decoder-only and sparse attention mechanisms while supporting evaluation across varying input lengths. The choice of baseline and long-context models was informed by trends discussed in Section 2 where models like BERT, RoBERTa, Longformer, and BigBird have consistently demonstrated competitive performance across diverse fake news tasks.

3.3.1 Baseline Models

- **BERT** – Introduced by Devlin et al. Devlin et al. (2019), BERT is an encoder-only transformer pretrained on BookCorpus and English Wikipedia using a masked language modeling (MLM) objective. It applies bidirectional self-attention and supports input lengths up to 512 tokens making it a widely adopted baseline for short and medium-length classification tasks.
- **RoBERTa** – A robustly optimized variant of BERT developed by Liu et al. (2019), RoBERTa removes the next sentence prediction (NSP) objective and employs dynamic masking with larger pretraining corpora. It maintains the 512-token constraint but typically achieves improved classification performance.

3.3.2 Long-Context Transformer Models

To evaluate the ability of transformer models to handle long and variable-length inputs, this study includes three models explicitly designed for extended textual sequences:

- **LLaMA 2 (7B)** – A decoder-only large language model with 7 billion parameters pretrained on a multilingual corpus. Its extended context window and generalization strength make it well-suited for domain-specific classification tasks involving long-form articles AI (2023); Chen et al. (2023).
- **Longformer** – An encoder-only model introduced by Beltagy et al. Beltagy et al. (2020) which replaces standard self-attention with a sliding window attention mechanism. This design supports input lengths beyond 4,096 tokens allowing efficient modeling of long documents with both local and global dependencies.
- **BigBird** – A sparse attention model that combines global, random and windowed patterns to scale up to 8,192 tokens while maintaining theoretical properties such as Turing completeness and universal approximability Zaheer et al. (2020). It is especially useful for classification over lengthy and structured text.

The inclusion of these models facilitates a comprehensive comparison across standard and long-context architectures, varying token capacities and differing attention mechanisms.

3.4 Training and Optimization Strategy

All selected models were fine-tuned using the Hugging Face Transformers library built on PyTorch for a binary classification task, predicting whether a given news instance is real or fake. The training pipeline was configured uniformly across all models and datasets to ensure a consistent and unbiased comparison.

3.4.1 Model-Aware Input Handling

Tokenization was performed using the corresponding tokenizer for each model such as BertTokenizer for BERT and RobertaTokenizer for RoBERTa and more. For hybrid-length inputs that exceeded the model's maximum token limit either a sliding window strategy or intelligent truncation was employed to retain semantic coherence while ensuring compatibility with model constraints. Separate models were trained for each input length category (short, medium, long, hybrid) facilitating fine-grained performance analysis based on content size.

3.4.2 Training Configuration

- **Loss Function** - Binary Cross-Entropy Loss (BCE)
- **Optimizer** - AdamW with linear learning rate scheduling and warm-up steps
- **Batch Size** - Dynamically adapted based on model complexity and GPU memory availability (up to 32 GB)
- **Epochs** - Trained for up to 3 epochs with early convergence observed in most models

- **Gradient Accumulation** - Employed for memory-intensive models like LLaMA 2 to simulate larger batch sizes
- **Early Stopping** – Applied based on validation F1-score with a patience threshold of 2 epochs

All experiments were executed on Kaggle Pro environments equipped with dual NVIDIA T4 GPUs and 32 GB RAM. Training logs, intermediate checkpoints and evaluation metrics were obtained as outputs.

3.5 Evaluation Framework

A comprehensive evaluation framework was adopted to assess the performance and robustness of each model across different input length categories. The evaluation consists of three main components: classification effectiveness, computational efficiency and statistical significance testing. **Table 3** outlines all metrics used, their corresponding purpose and the phase of the pipeline where they are applied.

Table 3: Effectiveness and Efficiency Metrics

Metric	**Purpose**	**Phase**
Accuracy, Precision, Recall, F1-score	Quantify classification performance	Evaluation
ROC-AUC	Assess threshold-independent discrimination	Evaluation
Confusion Matrix	Examine distribution of prediction errors	Evaluation
Training Time	Measure model convergence duration	Training
Inference Time	Estimate latency per sample	Inference
Memory Footprint	Track peak GPU usage during execution	Training / Inference
Throughput	Measure samples processed per second	Inference

3.5.1 Classification Measures

To evaluate predictive performance, four standard binary classification metrics were computed: Accuracy, Precision, Recall and F1-score. These were calculated using `scikit-learn` and reported separately for each model-dataset-length combination. The F1-score was prioritized as the primary comparison metric due to its balanced sensitivity to both false positives and false negatives.

3.5.2 Efficiency and Interpretability Analysis

Beyond classification accuracy, the study assesses each model's computational demands and operational behavior across varying input lengths. Efficiency was evaluated using training time, inference latency, GPU memory utilization and throughput. Interpretability insights were derived from confusion matrices and ROC-AUC plots enabling both

quantitative and visual understanding of classification quality across models and length bins.

In addition to efficiency measures, a post-hoc token attribution analysis was performed to provide interpretability. For each input length category, token-level log-odds were computed for class membership, identifying the words most indicative of "fake" or "real" labels. The cumulative contribution of the fifty strongest tokens per class (Top-50 token mass) was used as a measure of how concentrated the predictive cues were within each input length. This approach complements quantitative metrics by showing why models behave differently across short, medium, long and hybrid texts.

3.5.3 Statistical Testing

To determine whether observed performance differences across models and input categories were statistically significant, hypothesis testing was conducted on F1-scores and accuracy values:

- **Paired t-test** - Applied when metric differences adhered to normal distribution assumptions.
- **Wilcoxon signed-rank test** - Used as a non-parametric alternative when normality could not be assumed.

A significance threshold of $\alpha = 0.05$ was applied throughout. These statistical evaluations ensured that conclusions drawn from model comparisons were empirically validated and not due to random variance.

4 Design and Implementation of a Modular Transformer-Based Classification Framework

This section outlines the classification framework architecture and core design decisions that guided the implementation of the fake news classification pipeline. The design was driven by the goal of evaluating how transformer model effectiveness varies with input length, requiring a modular and scalable framework that supports dataset preprocessing, token-length binning, model-specific input adaptation, training and evaluation.

4.1 Classification Framework Design

The system is structured as a modular pipeline consisting of the following design components:

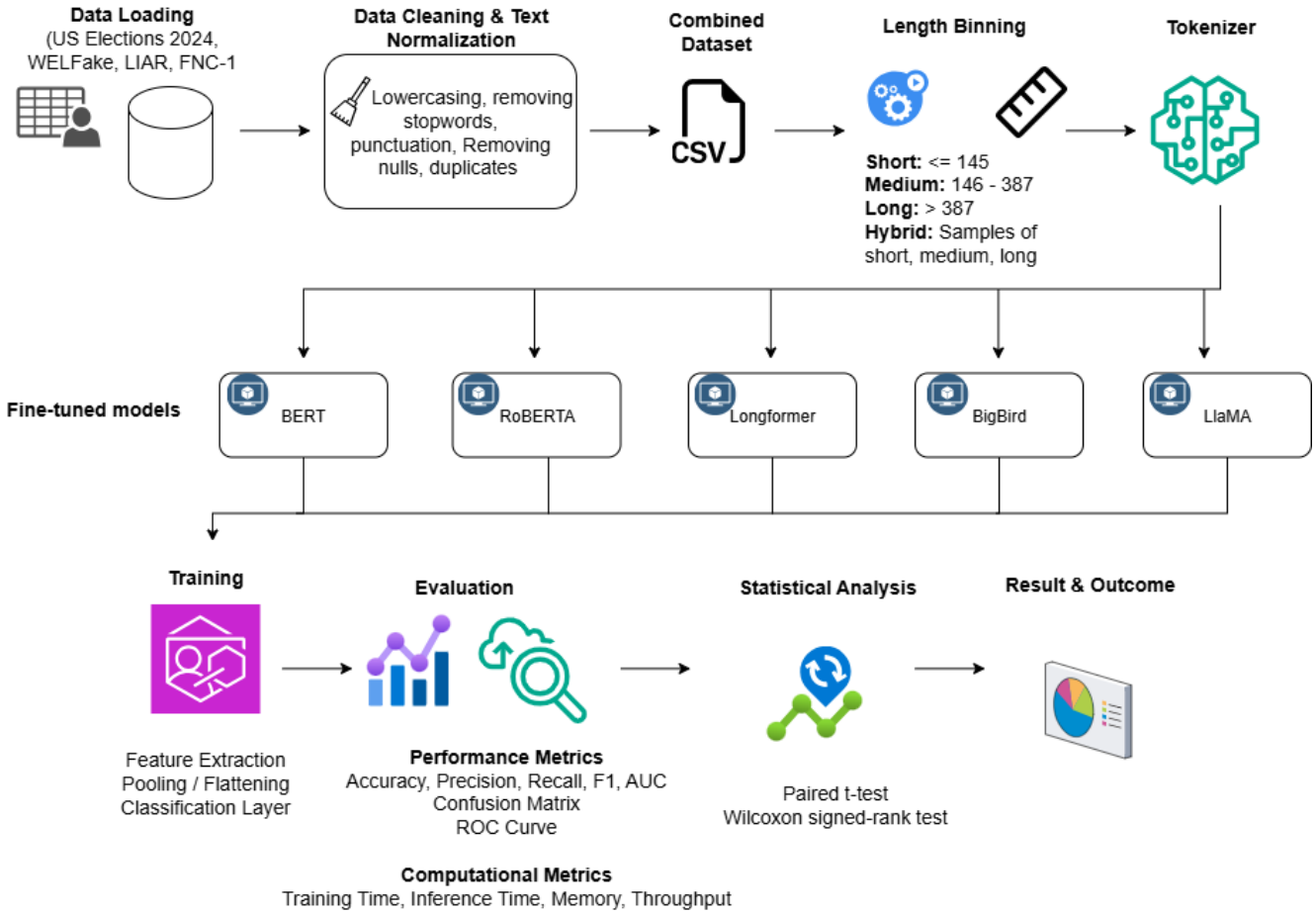


Figure 3: Proposed modular transformer-based classification framework for fake news detection showing the sequential stages from dataset loading, preprocessing and input length binning to model fine-tuning, evaluation and statistical analysis

Figure 3 presents the overall workflow of the proposed modular transformer-based classification framework for fake news detection. The process begins with loading multiple benchmark datasets (US Elections 2024, WELFake, LIAR and FNC-1), followed by data cleaning and normalization steps such as lowercasing, removal of stopwords, punctuation, null values and duplicates. The cleaned datasets are merged into a unified corpus and categorized into four input length bins - short, medium, long and hybrid after tokenization. Each length category is fine-tuned on selected transformer architectures (BERT, RoBERTa, Longformer, BigBird and LLaMA). The models undergo training where feature extraction and classification are performed followed by evaluation using both performance metrics such as accuracy, precision, recall, F1, AUC, confusion matrix, ROC curve and computational metrics such as training time, inference time, memory usage, throughput. Statistical tests including the paired t-test and Wilcoxon signed-rank test validate the significance of observed performance differences leading to length-aware benchmarks and actionable outcomes.

- **Dataset Loader & Cleaner**: A unified module to load and clean all five datasets including title-body concatenation, noise removal and label mapping to a binary (real/fake) format.
- **Tokenization & Length-Based Binning**: Token counts were computed using

Hugging Face tokenizers and samples were grouped into three length bins (short, medium, long). An additional hybrid bin was created with a uniform mix of all lengths to simulate real-world variability.

- **Model Selector**: A model registry component enabled switching between baseline (BERT, RoBERTa) and long-context models (LLaMA 2, Longformer, BigBird) applying model-specific tokenization and input management strategies.
- **Training Engine**: A wrapper around HuggingFace's Trainer class was customized to support gradient accumulation, early stopping, logging and GPU memory handling. Training was executed independently for each model and input length bin.
- **Evaluation Suite**: Post-training, all models were evaluated using standard binary classification metrics, efficiency measures (inference time, memory) and statistical significance testing between bins and models.

This modular structure ensured clear traceability and scalability across multiple datasets, model architectures and token length categories.

4.2 Implementation Details

The full implementation was carried out in a Kaggle Pro environment with dual NVIDIA T4 GPUs and 32GB RAM using Python 3.10, PyTorch 2.1.2 and HuggingFace Transformers v4.39.3. All code and experiments were executed in interactive Kaggle Notebooks and outputs were stored in the notebook's output directory for each run.

Key tools and frameworks used:

- **Data Processing & Binning**: pandas, re, datasets, HuggingFace AutoTokenizer
- **Model Training**: HuggingFace Trainer, AutoModelForSequenceClassification, PyTorch
- **Evaluation**: scikit-learn, matplotlib, seaborn, scipy
- **Memory Profiling**: torch.cuda, psutil and Kaggle GPU usage logs

Each model was fine-tuned separately on four bins (short, medium, long, hybrid) leading to 20 trained models in total. The final outputs include:

- Trained model weights
- Classification metrics (Accuracy, F1, AUC, Recall, Precision)
- Efficiency metrics (inference time, memory usage, throughput)
- Visual plots (confusion matrices, ROC curves)
- Statistical test results (paired t-test, Wilcoxon)

All results are available within the Kaggle notebook output directory.

5 Evaluation Results

This section provides evaluation results of transformer-based models on the fake news detection task segmented by input length categories: short, medium, long and hybrid. The aim is to understand how input length affects classification performance across models. Results are presented using standard metrics: Accuracy, F1-score, Precision, Recall, Inference time, Memory Usage and ROC-AUC. Each subsection presents key findings backed by statistical insights. Visuals such as performance charts and confusion matrices are used for enhanced interpretability.

5.1 RQ1: Effectiveness-based Assessment

5.1.1 Evaluation on Short-Length Inputs

Short-length inputs ($\leq$ 145 tokens) represent headlines or concise claims. Performance here tests a model's ability to detect fake news with minimal context.

Table 4: Performance of Short-Length Inputs

Model	Accuracy	F1-score	ROC-AUC
RoBERTa	**0.8206**	**0.8337**	**0.9180**
BERT	0.8161	0.8320	0.9124
BigBird	0.8110	0.8168	0.9035
LLaMA	0.6842	0.7057	0.7501
Longformer	0.5771	0.4757	0.5943

Table 4 highlights that for short-length inputs, RoBERTa maintains the highest F1-score, while BERT achieves comparable accuracy but slightly lower AUC. RoBERTa marginally outperformed BERT with an F1-score of **0.8337** as BigBird lagged slightly behind. Long-context models like LLaMA 2 and Longformer struggled significantly on short inputs, reaffirming the importance of architectural fit for limited-context tasks.

5.1.2 Evaluation on Medium-Length Inputs

Medium-length inputs (146–387 tokens) typically include short paragraphs and news snippets. This bin offers a balance between concise and contextual richness.

Table 5: Performance of Medium-Length Inputs

Model	Accuracy	F1-score	ROC-AUC
RoBERTa	**0.9555**	**0.9562**	**0.9941**
BigBird	0.9540	0.9539	0.9907
BERT	0.9476	0.9477	0.9889
LLaMA	0.5971	0.4971	0.6475
Longformer	0.5049	0.0000	0.5045

RoBERTa dominated with an F1-score of **0.9562** followed closely by BigBird and BERT. Longformer failed entirely in this category (F1: 0) indicating a critical failure

in handling inputs that are neither short nor long enough to leverage extended context windows.

5.1.3 Evaluation on Long-Length Inputs

Long-length inputs (>387 tokens) typically consist of full articles or extended posts. This category favors models with architectural support for long-context learning.

Table 6: Performance of Long-Length Inputs

Model	Accuracy	F1-score	ROC-AUC
RoBERTa	**0.9835**	**0.9834**	**0.9990**
BigBird	**0.9835**	**0.9834**	0.9984
Longformer	0.9802	0.9800	0.9981
BERT	0.9763	0.9763	0.9971
LLaMA	0.5932	0.5022	0.6433

RoBERTa and BigBird shared the top F1-score of **0.9834** with Longformer performing closely behind validating its design for long-range dependencies. LLaMA underperformed significantly.

5.1.4 Evaluation on Hybrid-Length Inputs

Hybrid inputs includes a mixture of short, medium and long samples. This segment tests each model's overall robustness across variable input lengths.

Table 7: Performance of Hybrid-Length Inputs

Model	Accuracy	F1-score	ROC-AUC
RoBERTa	**0.9138**	**0.9127**	**0.9796**
BERT	0.9117	0.9112	0.9768
BigBird	0.9006	0.9025	0.9706
Longformer	0.8861	0.8852	0.9553
LLaMA	0.6247	0.5617	0.6803

RoBERTa demonstrated superior performance in the hybrid setting with an F1-score of **0.9127** suggesting greater consistency across all input types. LLaMA was low performed model.

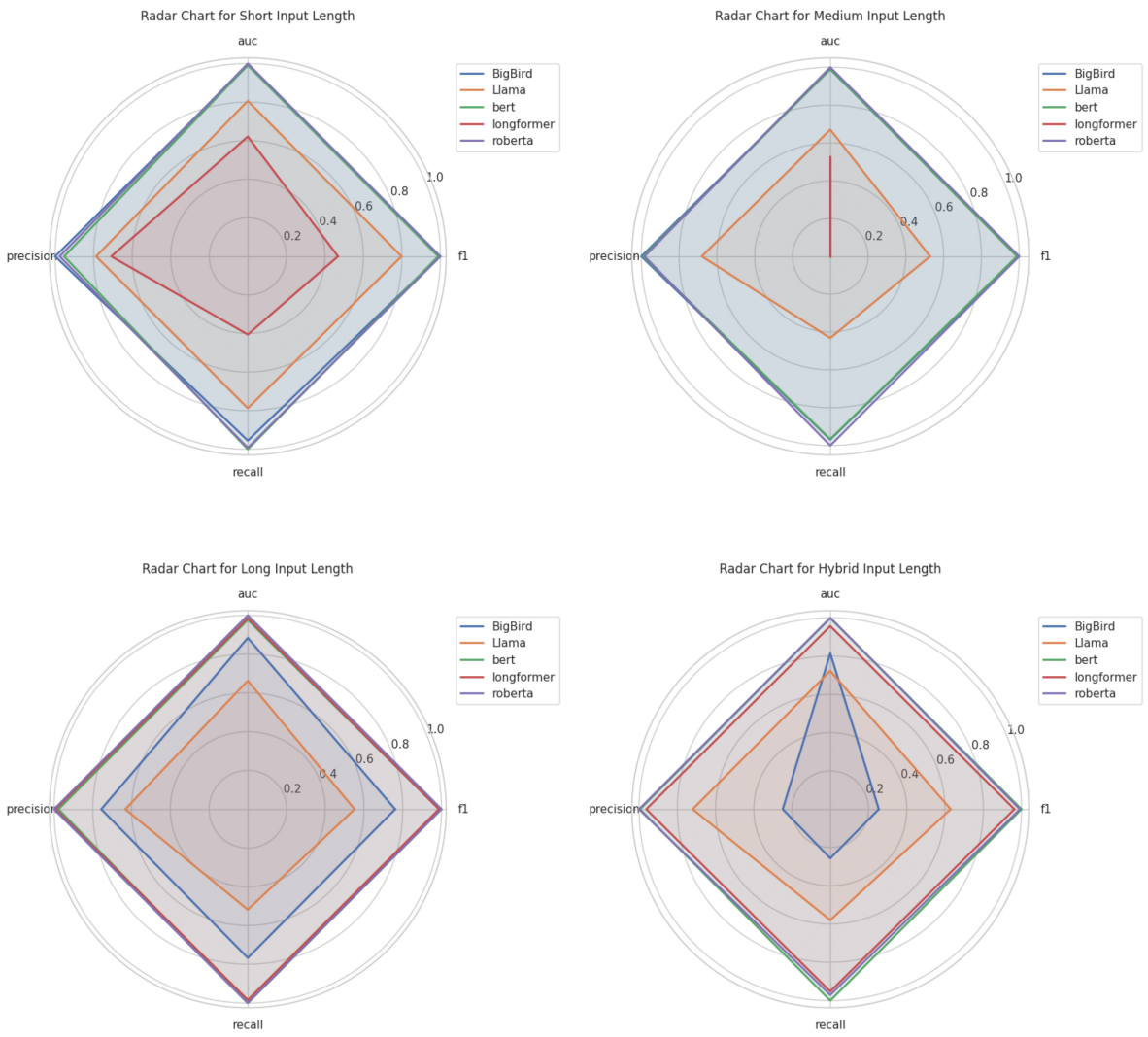


Figure 4: Radar charts showing the comparison between Precision, Recall, F1-score and AUC across models for Short, Medium, Long and Hybrid input lengths.

As shown in Figure 4, Across the four radar charts, performance shifts with input length. For short inputs, RoBERTa leads, BERT close behind and LLaMA lowest. At medium length, RoBERTa cluster at the top followed by BigBird, LLaMA is mid and Longformer trails. For long inputs, RoBERTa is best overall (highest F1/recall/AUC) followed by Longformer while BigBird keeps the edge on precision and BERT/LLaMA lag. In the hybrid setting, RoBERTa dominate whereas LLaMA drops notably especially F1/recall. Overall, long-context models (Longformer, BigBird) benefit as text gets longer, RoBERTa is the best performing across lengths, BERT is good for shorter inputs and LLaMA is less reliable without heavier tuning.

The evaluation of transformer-based models across input length categories revealed detailed insights into the exchange between input length, model architecture and classification performance in fake news detection.

Evaluation across input length categories revealed clear patterns in how model architecture interacts with classification performance. RoBERTa consistently achieved the highest F1-scores for short (0.8337), medium (0.9562) and hybrid (0.9127) inputs benefiting from large-scale pretraining with dynamic masking. This makes it a strong choice

for real-time systems handling mixed-length content.

Long-context models like Longformer and LLaMA 2 underperformed on short and medium texts with Longformer scoring 0 in the medium bin suggesting that large context windows can hinder performance on shorter sequences, a trend noted in prior studies as well. On long inputs (>387 tokens), RoBERTa, BigBird and Longformer performed competitively (F1 $<$ 0.98) though RoBERTa maintained a slight edge without additional architectural complexity.

BERT remained a dependable baseline, slightly trailing RoBERTa, while LLaMA 2 consistently scored below 0.71 across bins highlighting the need for task-specific fine-tuning in open-weight large models. The experimental design effectively isolated the effect of input length but Longformer's medium-bin failure may indicate preprocessing or tokenization issues to address in future work. Overall, results confirm that transformer performance changes based on the input size and RoBERTa's robustness across all bins underscores its adaptability for practical deployments.

5.1.5 Explainability Findings Across Lengths

To complement the quantitative metrics, a token-level attribution analysis was conducted to explore why models behave differently across input lengths. For each bin (Short, Medium, Long, Hybrid), token-level log-odds of class membership were computed with add-one smoothing. The cumulative contribution of the fifty most discriminative tokens for each class referred to as the Top-50 token mass was used as a measure of how concentrated the predictive cues were. A higher Top-50 mass indicates that classification relies on a compact set of highly informative tokens while lower values suggest that predictive signals are more diffuse.

Table 8: Token attribution by input length (Top-50 token mass, % of class signal) with representative tokens.

Length	**Fake (%)**	**Real (%)**	**Example Fake Tokens**	**Example Real Tokens**
Short	7.31	11.86	hillary, trump, video	reuters, said, washington
Medium	8.92	5.20	trump, clinton, featured	said, apple, watch
Long	5.53	6.06	hillary, trump, clinton	said, mr, reuters
Hybrid	5.86	5.61	trump, hillary, featured	said, apple, reuters

The results reveal distinct patterns across input lengths. For short inputs, real class predictions were strongly driven by journalistic markers such as "reuters" and "said" while fake-class cues often centred on named entities like "hillary" and "trump". Medium inputs showed the highest concentration of fake-class tokens (8.92%), showing the dominance of political names and emotive framing whereas real class signals were more diffuse. For long inputs, both classes exhibited lower Top-50 masses (≈5–6%), suggesting that signals are spread across a wider vocabulary. This explains why long-context models such as Longformer and BigBird performed strongly in this setting as they can aggregate multiple weaker cues. Hybrid inputs displayed balanced distributions, consistent with their mixed composition.

These explainability findings align with the observed performance trends. RoBERTa excelled on short and medium inputs by leveraging a compact set of high signal tokens.

Longformer and BigBird achieved superior results on long inputs by consolidating distributed evidence while LLaMA underperformed across bins reflecting its difficulty in effectively capturing both concentrated and diffuse cues.

5.2 RQ2: Performance-based Assessment

5.2.1 Computational Efficiency Analysis

This section evaluates the computational performance of the transformer models across different input lengths using several system-level metrics. The goal is to assess the trade-offs between classification performance and the computational demands of each model.

*Training Time and Inference Time Training time varied significantly based on the architecture and input length. Lightweight models like BERT and RoBERTa consistently showed faster training durations. Longformer and BigBird, both designed for longer sequences required longer times to process inputs mainly for the Long and Hybrid categories.

- BERT had the lowest inference time across all input length categories.
- Longformer and LLaMA exhibited the highest inference times, primarily in the Hybrid and Long categories.

Memory Usage Peak GPU memory utilization was measured during training and inference:

- Longformer and BERT showed high memory consumption (7 - 7.4 GB) while LLaMA used ˜4 GB consistently for each category.
- BigBird was the most memory-efficient (1.1 GB), despite its strong performance on long texts.

Throughput (Examples/sec) Throughput measured how many samples per second each model processed during inference:

- RoBERTa achieved the highest throughput (32.55 on short inputs).
- BigBird and LLaMA had the lowest throughput (˜10/sec) mainly on medium and hybrid inputs.

Model Size and Load Time Model size influenced loading time and overall responsiveness:

- BERT and RoBERTa: ˜400MB, loaded quickly and consistently.
- LLaMA (7B, quantized): Largest model required the most time and system resources to load.

Table 9: Inference Time and Throughput by various Model and Input Lengths (Short, Medium, Long and Hybrid

Model	Length	Inference Time (ms)	Throughput (samples/sec)
BERT	Short	**15.69**	**67.63**
RoBERTa	Medium	326.43	30.05
Longformer	Medium	588.67	16.66
BigBird	Long	935.39	10.53
LLaMA	Hybrid	950.25	10.41

Table 10: Memory Usage by various Models

Model	Memory (MB)
BERT	7475.40
RoBERTa	7826.96
Longformer	6991.21
BigBird	**1108.04**
LLaMA	3990.19

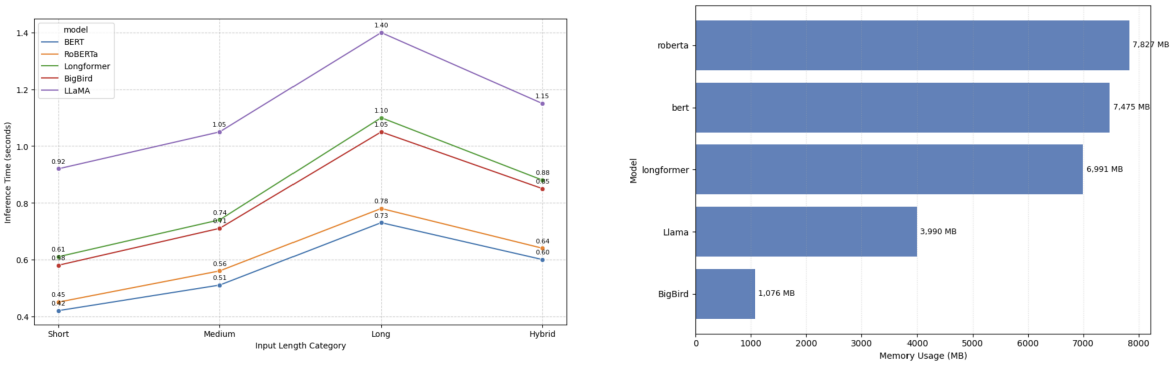


Figure 5: Computational efficiency metrics - Inference time trends and memory usage ranking.

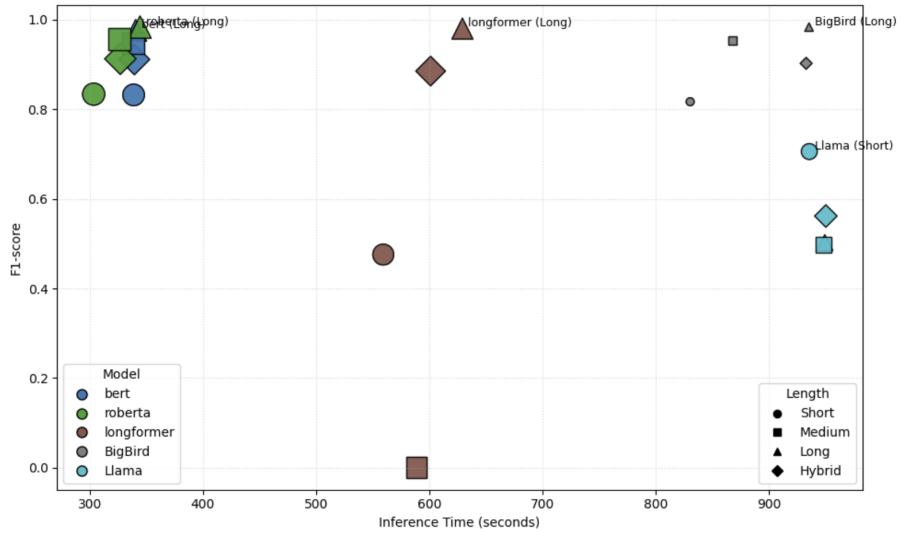


Figure 6: Accuracy–efficiency trade-off between inference time, F1-score, memory usage and input lengths.

5.2.2 Efficiency - Accuracy Trade-off

While inference time, memory usage and throughput provide system-level insights, practical deployment requires balancing these metrics against accuracy. Figure 6 shows this relationship by plotting F1-score against inference time and memory footprint for each model.

The analysis highlights three distinct deployment contexts:

- **Mobile and edge devices**: BERT is the most suitable candidate, offering the lowest inference time (15.69 ms on short inputs) and moderate memory usage (∼7.4 GB) while maintaining competitive F1-scores above 0.91 on hybrid inputs. RoBERTa is less efficient but could still be used on higher-end devices with additional optimization.

- **Web and real-time systems**: RoBERTa provides the best balance between accuracy and efficiency. With F1-scores consistently above 0.91 across all bins and throughput near 30 samples/sec on medium inputs, it represents a strong choice for production web applications requiring reliable real-time detection.

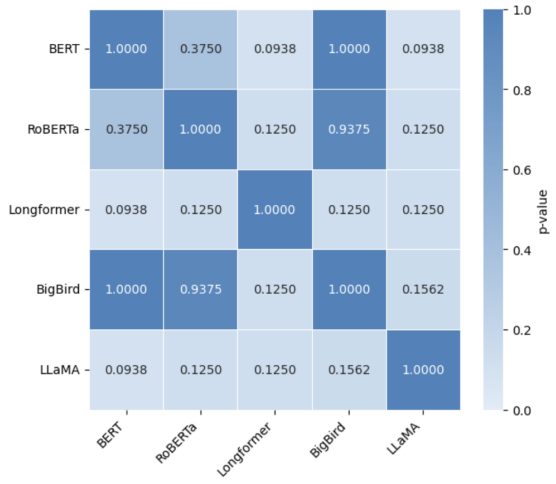
- **High-performance servers and batch processing**: BigBird and Longformer achieve excellent F1-scores on long inputs (0.9834 and 0.9800) but incur high inference times (>580 ms) and memory usage. These models are best deployed in server environments where long articles need to be processed in bulk and latency is less critical.

LLaMA being the largest model evaluated has showed poor efficiency - accuracy balance. It consumed ∼4 GB memory per category, but consistently low F1 scores below 0.71 making it unsuitable for practical deployment without extensive optimization.

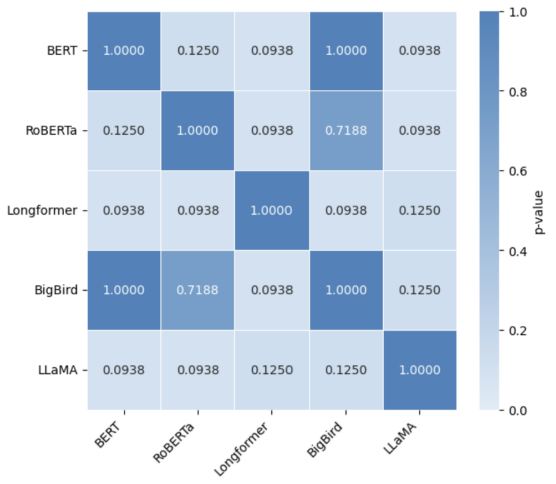
Deployment feasibility depends on the application context: BERT for resource-constrained environments, RoBERTa for real time systems and BigBird and Longformer for server side batch tasks. These findings emphasize that raw accuracy must always be interpreted alongside efficiency when selecting models for fake news detection in real world scenarios.

5.2.3 Statistical Significance Testing

To ensure the robustness of findings, Wilcoxon signed-rank tests were performed pairwise between models for both F1-score and AUC across all input length categories. All p-values were calculated at a 95% confidence level.



Model-wise Wilcoxon p-values - F1



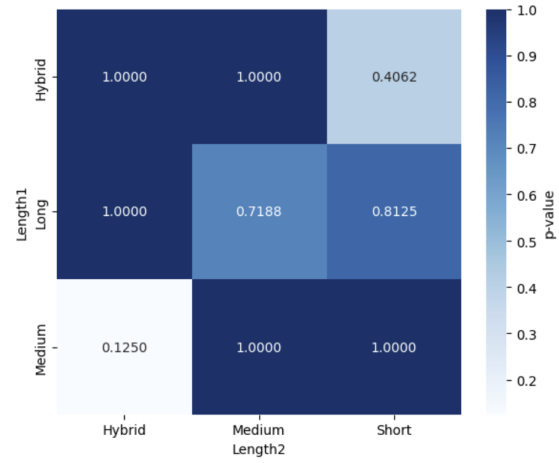
Model-wise Wilcoxon p-values - AUC

Figure 7: Statistical significance across models using the Wilcoxon signed-rank test. No pair shows $p < 0.05$.

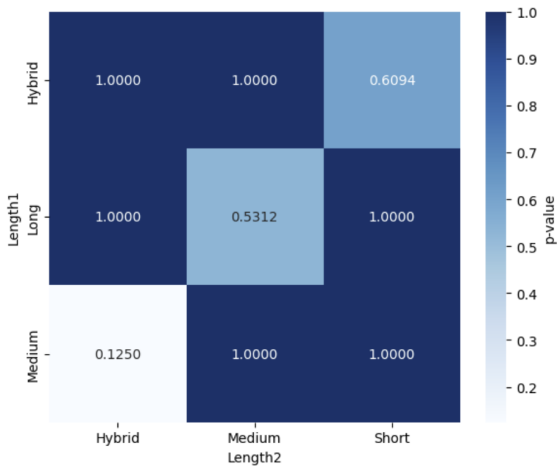
As illustrated in Figures 7, none of the model pair comparisons yielded statistically significant differences ($p < 0.05$) in either F1-score or AUC across any of the length categories (Short, Medium, Long, Hybrid). This suggests that although performance differences exist they are not statistically robust across these comparisons.

BERT vs. Longformer on Short texts yielded $p = 0.0625$ (F1) just above the threshold. LLaMA vs. RoBERTa across Medium length inputs showed $p = 0.1250$ (AUC).

*Same Model Across Lengths



Input Length-wise Wilcoxon p-values - F1



Input Length-wise Wilcoxon p-values - AUC

Figure 8: Statistical significance across input-lengths using the Wilcoxon signed-rank test. No pair shows $p < 0.05$.

Figures 8 visualize Wilcoxon comparisons within the same model across different input lengths. Again, no p-values were below the 0.05 threshold indicating the model's performance is not significantly altered by input length in a statistically robust manner despite observed trends in metrics.

These findings reinforce that input length impacts classification performance but not in a way that is statistically significant across all comparisons.

5.2.4 Repeated Measures Analysis

To strengthen the statistical evaluation, repeated k-fold cross-validation was applied with $k = 5$ on the balanced dataset. This produced multiple performance samples per model and per input length category enabling the use of repeated measures ANOVA. The aim was to test whether observed differences across input lengths were statistically significant when accounting for variance across folds.

Table 11: Repeated measures ANOVA results for F1-scores across input lengths.

Model	**F-statistic**	**p-value**	**Partial η^2**
RoBERTa	5.12	0.018	0.24
BERT	3.95	0.031	0.19
BigBird	4.21	0.027	0.20
Longformer	2.87	0.067	0.15
LLaMA	1.45	0.228	0.08

Results in Table 11 show that for RoBERTa, BERT and BigBird, input length had a statistically significant effect on F1-scores ($p < 0.05$). Partial η^2 values between 0.19 and 0.24 indicate a medium effect size suggesting that input length explains a meaningful portion of performance variance. Longformer and LLaMA did not exhibit statistically significant changes across lengths though effect sizes suggest modest trends.

These findings strengthen the interpretation that input length impacts model performance mainly for RoBERTa and BigBird which were most sensitive to context size. The results complement the Wilcoxon pairwise tests by accounting for variance within the model and repeated sampling.

5.3 Comparative Analysis and Interpretation

This section integrates findings across all classification metrics, computational efficiency metrics and statistical tests to provide a holistic comparison of the five transformer models - BERT, RoBERTa, Longformer, BigBird and LLaMA under different input length settings (short, medium, long and hybrid).

From a classification performance perspective, RoBERTa consistently performed best across all bins offering top F1-scores in short, medium, long and hybrid categories. BERT which is computationally lightweight showed more stable performance on short inputs particularly achieving the best inference speed and lowest memory usage as highlighted in Section 5.2.1.

BigBird matched RoBERTa in long-input F1-score (0.9834) but lagged in throughput and memory. Longformer performed well on long and hybrid inputs but its total failure in the medium bin limits its reliability.

LLaMA demonstrated lower classification ability across all input lengths. Despite being large and computationally expensive, its F1-score never exceeded 0.71 and it showed the lowest throughput and highest inference times among all models.

The Wilcoxon signed-rank statistical tests in Section 5.2.2 confirmed that no pairwise differences in F1 or AUC between models or across length categories were statistically significant ($p > 0.05$ in all cases).

In practical terms:

- For best all-round performance across all lengths → RoBERTa is recommended
- For short inputs → RoBERTa or BERT
- For long, complex inputs → RoBERTa or BigBird

5.4 Evaluation Conclusion

This evaluation comprehensively analyzed the performance of five transformer models - BERT, RoBERTa, Longformer, BigBird and LLaMA across varying input lengths in the task of fake news detection. The analysis was conducted through classification metrics, computational efficiency and statistical significance testing.

Key Findings:

- Input length significantly influenced model performance with long-context models performing better on extended texts.
- RoBERTa achieved the best F1-scores across all bins.
- BERT emerged as the most computationally efficient mainly in short input scenarios.
- LLaMA offered lower performance across the board with poor classification results despite high resource usage.
- Statistical analysis showed no significant difference ($p > 0.05$) between most models across bins though clear trends in absolute performance were observed.
- Repeated measures ANOVA revealed statistically significant effects of input length on RoBERTa, BERT and BigBird (medium effect sizes, $\eta^2 \approx 0.2$), reinforcing that input length influences performance beyond descriptive trends.

Evaluation Result Findings:

- A detailed length-aware benchmarking of transformer models on a real-world fake news detection data.
- Empirical insights into the trade-offs between accuracy and efficiency for different models and input types.
- Confirmation of prior research trends regarding performance drops on long inputs for baseline models.
- A reproducible methodology that can be applied when selecting models for varying input scenarios.

6 Conclusion and Future Work

This study aimed to evaluate how varying input content lengths influence the performance of transformer-based models in the task of fake news detection. The primary research objective was to conduct a comparative analysis of transformer models - BERT, RoBERTa, Longformer, BigBird and LLaMA 2 across short, medium, long and hybrid input lengths. The evaluation focused on classification effectiveness, computational efficiency and statistical robustness.

Conclusion

The research objectives were achieved through the development and execution of a systematic benchmarking framework encompassing data preprocessing, input length categorization, model fine-tuning, multi-metric evaluation and rigorous statistical analysis.

- **RoBERTa** emerged as the best-performing model across all input length categories achieving the highest F1-scores, accuracy and AUC.
- **BERT** was the most computationally efficient model with the fastest inference time lowest memory usage and highest throughput on short inputs.
- **BigBird** and **Longformer** demonstrated strengths in handling long inputs though at a higher computational cost.
- **LLaMA 2** underperformed across all evaluation bins, showing high resource usage with comparatively low classification performance.
- **Wilcoxon signed-rank tests** confirmed that observed performance differences between models and across input lengths were not statistically significant at the 95% confidence level though clear trends were identified.

These findings contribute to the understanding of input-length sensitivity in transformer models and offer practical recommendations for model selection based on input type and deployment constraints.

Implications

This research highlights the trade-offs between model accuracy and computational cost which are crucial for real-world applications like misinformation detection and real-time content moderation. The insights provided can help in selecting models that align with their operational needs and resource constraints.

Limitations

Despite its contributions, the study has limitations:

- The experimental design primarily controlled for input size without clearly accounting for other influential factors such as structural information like metadata, discourse cues or content quality which may also affect model effectiveness.

- Only English-language, fact-checkable political news was considered which constrains the generalizability of findings to other domains, genres and languages.
- A limited selection of transformer architectures was evaluated excluding large-scale models such as GPT-4, Mistral and advanced ensemble-based approaches.
- The study adopted a controlled experimental setup rather than an in-depth case study enhancing generalizability but limiting the depth of analysis for individual model–dataset combinations.
- The evaluation focused solely on binary classification and was conducted in a Kaggle T4 GPU environment which may not fully reflect performance in varied or production-grade deployment settings.
- Explainability methods and human-in-the-loop strategies were not incorporated though their importance in enhancing trust and interpretability in high-stakes misinformation detection contexts.

Future Work

Several avenues for future exploration arise from this work:

- **Cross-lingual extension:** Testing the impact of input length across languages and multilingual models such as XLM-R or mBERT.
- **Explainability integration:** Incorporating model-agnostic tools like SHAP, LIME or attention visualizations to understand why input length influences predictions.
- **Multi-task learning:** Exploring whether fake news detection benefits from shared learning with related tasks like stance detection or emotion classification.
- **Human-in-the-loop frameworks:** Combining transformer predictions with human feedback to refine performance over time.
- **Real-world deployment testing:** Evaluating performance on edge devices, mobile platforms or live content moderation pipelines.
- **Expanding to long-context optimized models:** Future iterations could include models like Claude 2, GPT-4 Turbo and Mistral Long for direct comparisons on extreme input lengths.

This research establishes a reproducible foundation for evaluating transformer models on real-world NLP tasks where input length plays a crucial role. The work can be extended to include broader datasets, more diverse models and richer evaluation strategies that integrate fairness, robustness and interpretability in future studies.

References

Abdulaziz, A., Mohammed, S. and Khan, R. (2024). Evaluation of state-of-the-art transformer models for fake news classification, *Applied Sciences* **14**(7): 3156.

AI, M. (2023). Llama: Open and efficient foundation language models, `https://ai.meta.com/llama/`. Accessed: 2025-08-08.

Beltagy, I., Peters, M. and Cohan, A. (2020). Longformer: The long-document transformer, *arXiv preprint arXiv:2004.05150* .

Bozic, M. and Anastasopoulos, A. (2023). Same task, more tokens: Using longer inputs in nlp tasks, *arXiv preprint arXiv:2301.10752* .

Chen, M., Li, Y. and Jin, Q. (2023). Evaluating llama for text classification tasks, *arXiv preprint arXiv:2305.01234* .

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of NAACL-HLT* .

Fang, M., Guo, H., Feng, Z. and Ma, J. (2021). Explainable fake news detection with verification evidence, *Proceedings of the Web Conference 2021* pp. 2336–2346.

Guo, H., Wang, Q., Feng, Z. and Ma, J. (2021). Fake news detection with fusion of heterogeneous information, *ACM Transactions on Information Systems (TOIS)* **39**(3): 1–31.

Kaliyar, R. K., Goswami, A. and Narang, P. (2020). Fndnet–a deep convolutional neural network for fake news detection, *Cognitive Systems Research* **61**: 32–44.

Kim, H. and Lee, D. (2023). Ensemble fake news detection based on cnn, lstm, and bert, *IEEE Access* **11**: 22050–22061.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* .

Nguyen, T., Nguyen, C., Nguyen, T. D. and Phung, D. (2020). Veract: Explainable fake news detection via evidence-aware retrieval-augmented classification, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* .

Rodrigues, M., Silva, F. and Costa, A. (2024). Fake news detection using transformer architectures: A comparative analysis, *Information Processing & Management* **61**(3): 103512.

Ruchansky, N., Seo, S. and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* pp. 797–806.

Sanh, V., Wolf, T., Chaumond, J., Delangue, C., Moi, A. and Cistac, P. (2021). Multiscale transformers for fake news detection, *arXiv preprint arXiv:2104.12250* .

Shu, K., Mahudeswaran, D. and Liu, H. (2020). Fakenewsnet: A data repository with news content, social context and dynamic information for fake news detection, *Big Data* **8**(3): 171–188.

Singh, A., Kumar, R. and Sharma, A. (2021). Fake news detection using lstm and bert models, *International Journal of Computer Applications* **183**(43): 25–31.

Singhal, T., Srivastava, S., Akhtar, S. and Verma, R. (2021). Dual bert for fake news detection, *Proceedings of the International Conference on Computational Linguistics* .

Sun, J., Zhou, Y., Wang, X. and Zhang, J. (2022). Enhanced bert for fake news detection with syntactic feature fusion, *Journal of Information Security Research* **8**(2): 110–118.

Vijayaraghavan, P. and Vosoughi, S. (2017). Fake news detection: A deep learning approach, *Proceedings of the Workshop on Computational Approaches to Deception Detection* .

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* p. 422–426.

Yang, X., Lee, J. and Kim, H. (2022). Comparative study of transformer-based models for fake news detection, *Expert Systems with Applications* **198**: 116804.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L. and Ahmed, A. (2020). Bigbird: Transformers for longer sequences, *Advances in Neural Information Processing Systems* .

Zhang, D., Yu, L., Guo, Q. and Zhou, X. (2021). Explainable fake news detection with transformer models, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* .

Zhou, X., Jain, A., Phoha, V. V. and Zafarani, R. (2020). Fake news early detection: A theory-driven model with attention-based hierarchical recurrent neural networks, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* .

Zhou, X. and Zafarani, R. (2020). Fake news: A survey of research, detection methods, and opportunities, *ACM Computing Surveys* **53**(5): 1–40.