**FMDB** *Transactions on Sustainable Intelligence and Security*

# AI-Powered Fake News Detection System Using Natural Language Processing

**Rajendra Prasad Joshi**[1,*]

[1]Department of Computer Science and Technology, University of West Scotland, Paisley, Scotland, United Kingdom.
b01794509@studentmail.uws.ac.uk[1]

**Abstract**: The accelerated rate of online information has worsened the difficulty in differentiating the real news from misinformation. Although there have been breakthroughs in artificial intelligence, most current fake news detection models have inflated accuracy under idealistic validation conditions. In practice, they do not provide human-friendly transparency. This dissertation proposes the design, development, and testing of an AI-based fake news detector that incorporates Natural Language Processing (NLP), transformer-based models, and human-centred explainability to improve algorithm performance and user confidence. The study is based on a comparative experimental approach, which compares classical machine learning models, such as Naive Bayes, Logistic Regression, and Support Vector Machine (SVM), with deep learning and transformer models such as LSTM and Distill BERT. This study examines the performance of false news detection models using the Kaggle False News Dataset, the LIAR Dataset, and a Nepal-focused corpus. Researchers measure the performance using accuracy, precision, recall, F1-score, ROC-AUC, and Expected Calibration Error (ECE). Researchers install the top-performing model in a Django-based web application that delivers calibrated confidence scores and short explanations. The results show that transformer-based models outperform traditional approaches, and that explainable, user-centred interfaces improve trust and decision-making quality in misinformation detection.

**Keywords:** Fake News Detector; Natural Language Processing; Naive Bayes; Logistic Regression; Support Vector Machine; Expected Calibration Error; Machine Learning; Deep Learning; Artificial Intelligence.

**Received on:** 20/06/2025, **Revised on:** 27/08/2025, **Accepted on:** 24/10/2025, **Published on:** 07/03/2026

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSIS

**DOI:** https://doi.org/10.69888/FTSIS.2026.000668

**Cite as:** R. P. Joshi, "AI-Powered Fake News Detection System Using Natural Language Processing," *FMDB Transactions on Sustainable Intelligence and Security*, vol. 1, no. 1, pp. 82–101, 2026.



## 1. Introduction

### 1.1. Background and Context

The digitalization of the media has lately changed the way information is accessed, shared, and interpreted among individuals at a tremendous speed. The communication landscape has been democratized by social networking sites and online news media, though it has also increased the dissemination of false information and fake news. False narratives have become viral and more consequential than truthful information during crises such as the COVID-19 pandemic, elections, and political conflicts worldwide, altering how people think, fostering distrust, and disrupting democracy. In emerging countries like Nepal, where few people are digitally literate and social media platforms are widely prevalent, the effects of misinformation are even more

[*]Corresponding author.

pronounced, posing severe societal and informational challenges. Conventional approaches to fact-checking are mostly subjective, time-intensive, and reactive, and cannot keep up with the speed and volume of the dissemination of fake or false information. As a result, Artificial Intelligence (AI) and Natural Language Processing (NLP) have become effective tools for detecting fake news at scale and with great speed. Using machine learning (ML) and deep learning (DL), NLP-based systems can recognise linguistic, semantic, and contextual patterns in text and classify information as real or fake.

Although many AI models have demonstrated high accuracy in controlled experimental settings, their implementation and interpretation in real-world settings are limited. Transformer models, e.g., BERT and DistilBERT, perform better than traditional ML models (e.g., Naive Bayes, Logistic Regression, SVM) on benchmark data; however, they are usually black boxes, i.e., they do not provide much information about how they make their decisions. Such a lack of transparency breeds distrust among users in the reliability of model predictions. While transformer models like BERT and DistilBERT achieve outstanding performance, they are opaque systems whose decision-making processes are difficult for end users to decipher. This poor interpretability poses a problem in such sensitive areas as misinformation detection. That is why it is significant that the confidence values produced by these models should not only be accurate but also consistent, and that users should be able to make appropriate decisions based on reliable probabilistic outputs. Probabilistic calibration is significant for aligning the model's confidence with real-world accuracy, enabling users to make more informed decisions in the absence of model-level explanations.

### 1.2. Problem Statement

Even though much has been done in the field of fake news detection, existing systems are limited to two critical issues. To begin with, most models are tested on randomly split datasets, which artificially inflate their accuracy owing to data leakage (e.g., similar publisher styles or time variation). These types of assessments may not reflect actual performance in real-life conditions. Second, there is a tendency toward calibration error in transformer models, leading to incorrect predicted probabilities. When detecting misinformation, inappropriate confidence calibration can mislead users, leading them to trust or doubt too much. In the absence of interpretable outputs, a user can trust or distrust automated systems excessively or too little, reducing the quality of their decisions. The context in Nepal also contributes to the problem, with Tode switching, transliteration, and scarce annotated datasets, which make it difficult to generalize the model. It is crucial to have a system that is both technically robust and, in its transparency, human-centred, to address problems of both algorithmic performance and societal trust.

### 1.3. Aim of the Study

This research aims to design, develop, and test a transformer-based fake news detector using probability calibration techniques, with higher reliability, user trust, and practical use.

### 1.4. Research Objectives

The research work has specific goals, which are to:

- To critically compare new methods of fake news detection based on NLP, their performance, and the methods they use to validate.
- To prepare, collect, and preprocess various datasets, including the Kaggle Fake News Dataset, the LIAR Dataset, and Nepal-related texts.
- To fit and test the classical machine learning models (Naive Bayes, Logistic Regression, SVM) and deep learning models (LSTM, DistilBERT) with the help of sound validation schemes (publisher-aware, temporal, cross-dataset).
- To calibrate model probability output using post-hoc calibration methods, e.g., Platt scaling and temperature scaling, and quantify calibration quality with ECE and Brier score.
- To introduce the most successful model into a Django web application, showing calibrated prediction results with a transparent and user-friendly interface.
- To carry out a user interface evaluation to determine how the score of confidence and system design affect user trust, decision-making, and usability.

### 1.5. Research Questions

- What is the difference between classical machine learning models and transformer-based models (e.g., DistilBERT) in the ability to detect fake news when working under realistic validation conditions?

- What are the effects of confidence scores and interface design on user trust and decision-making in relation to using an AI-based fake news detection system?

### 1.6. Scope and Delimitation

The artificial news fake detector covered by this paper is text-based with no multimodal analysis of images or video processing. The data used is mostly in English, with a small Nepal-specific set used for external validation. The web application developed is not a production-level system but a functional prototype. This work will consist of data set preparation, model benchmarking, probability calibration, system deployment, and a small-scale usability test. However, large-scale user testing and a sophisticated ML-Ops infrastructure are not in place. Such restrictions will keep the paper within a reasonable timeframe while still allowing the research questions to be answered methodologically.

## 2. Literature Review

### 2.1. Introduction

It is the rapid pace of the digital communication technology revolution that has changed the way information is consumed, shared, and interpreted. Even though this has improved access to knowledge, it has also heightened the extent to which misinformation can influence political decisions, population health behaviors, and cohesiveness within a particular society [2]. The area of misinformation has increasingly been framed as a societal and psychological rather than a technical or computational problem, influenced by cognitive biases, echo chambers, and algorithms [3]. The automation of fake news has been led in its turn by artificial intelligence, i.e., NLP. In the past five years, research has been transformed by the growing availability of computational power and the development of deep learning models, shifting attention from traditional feature engineering to contextualized language understanding [4]. This is a literature review that presents a synthesis of recent peer-reviewed advances in fake news detection, rather than a mere description [5]. In addition, this paper presents fake news identification as an interdisciplinary endeavour that draws on machine learning, linguistics, human-computer interaction, and behavioural science. It is expected to indicate which models can and cannot achieve, and why failures are important for real-world use [6].

### 2.2. Fake News Detection Background

Research on the detection of fake news has now recognized that fake information is a multidimensional problem. It can be deliberately counterfeited, falsely memorized, partisan, or mechanically exaggerated [7]. It is because of this complexity that recent literature recommends segmenting fake news detection into three broad categories: content-based, meaning linguistic and semantic analysis [8]. Context-based tactics such as metadata, source behavior, sharing patterns, and temporal posting behavior [9]. A hybrid strategy for integrating textual and contextual data [10]. The high proportion of English datasets has historically limited studies to the Western information environment and limits the external validity of findings [11]. Particularly problematic is this in multilingual countries like Nepal, where misinformation is commonly spread in part Nepali, part English through code-switching, informal romanisation, and culturally coded metaphors. Research has confirmed that models trained only on formal English are incapable of explaining these complexities [12]. In addition, researchers indicate that misinformation cannot be treated as a linguistic phenomenon but as a socio-technical one [13]. Indicatively, using false political tales is usually founded on an emotional pattern of which usage may be through fear, indignation, or moral condemnation. Understanding these processes helps explain why there are occasions when linguistic-only models erroneously apply persuasive narratives that resemble those of fact reporting [14]. Distortion is also hostile, according to recent literature. As detection models evolve, second-order malicious agents alter their strategies, creating a game of cat and mouse [15]. Consequently, the fake news detection models should not merely be correct; they should be robust, flexible, and updated in real time [16].

### 2.3. Classical Approaches to Machine Learning

The capabilities of classical machine learning approaches of being interpretable and having stable behavior remain basic since they have become a defining characteristic of responsible AI [17]. One such is Logistic Regression, which provides a visual representation of the association between linguistic characteristics and classification outcomes, which is not hard to understand, which is why some articles can be referred to as fake [18]. It is particularly true of organisations such as news agencies, governments, and fact-checking organisations, where the decision-making process must be made transparent [19]. The other recently conducted studies also indicate that simple models are exceptional at training on high-quality features built on domain knowledge [20]. Sentiment indicators, indicators of modality (e.g., might, allegedly), and stylometric features (e.g., sentence length, punctuation density) are likely to be elevated in this instance and to yield high baseline performance [21]. Moreover, classical models are robust to distributional shifts, which is crucial for detecting fake news when content evolves rapidly [22]. It is established that transformer models tend to overfit data-specific patterns and generalize better when trained on small or

noisy data than classical models do [23]. The classical models, however, are ineffective when required to make a contextual choice [24]. They fail to decipher sarcasm, roundabout meanings, or long-winded conversation [25]. Nor can they detect the insidious garnering of political misinformation [26]. Therefore, despite being useful as a comprehensible baseline or part of an ensemble system, classical models cannot offer the richness required for contemporary misinformation [27].

### 2.4. Deep Learning and Transformer-Based Models

Machine-learning methods have been at the heart of phishing detection, as they can extract patterns from large datasets and categorize previously unknown URLs [28]. Various surveys have compared classical classifiers such as Logistic Regression, Support Vector Machines (SVM), KNN, Naive Bayes, and Decision Trees, as well as ensemble algorithms such as Random Forests and Gradient Boosting [29]. Ensemble models tend to be more accurate because they can produce multiple decision boundaries, limit overfitting, and enhance generalization. Zheng [36] showed that this advantage could be achieved by implementing an ensemble model in a Django environment capable of making reliable real-time predictions [30]. Models with deep learning, like EGSO-CNN in 2025, that automatically extract features from raw URLs further improve performance by learning deep structural patterns on top of handcrafted features [31]. Nevertheless, accuracy is not what defines a model's suitability. Most high-performing models consume substantial computational resources and are therefore not viable for real-time implementation [32]. It is observed in the literature that much simpler models, including Logistic Regression and Linear SVM, are as competitive and exhibit:

- Much reduced inference time.
- Transparency and reduced interpretability by end users.
- Fewer resources are needed, which can be shared between hosts or low-weight web servers.

They may offer greater benefits than the minor performance improvements of more sophisticated models in cases where instantaneous prediction is required (e.g., web page URL-checking programs). The given trade-off is of special interest to the present dissertation, as it focuses more on latency, resource efficiency, and transparency in its deployment strategy [33].

### 2.5. Calibration and Reliability of Predictions of the Model

Feature engineering is an important element in ML-based phishing detection. Most detection pipelines are based on lexical, host, and content-based features, and experiments have demonstrated the high performance of their combination. As shown by Acharya [1], hybrid feature selection methodologies reduce computational load and enhance classifier performance by discarding redundant and/or noisy features. Equally, Zheng [36] matched the accuracy of deep learners with a small, carefully selected feature set, demonstrating that well-selected features can compete with standard deep learners in most cases [34]. Prolonged limitation discussion [35]. Although feature engineering approaches proved to be important, they have several limitations:

- **Reliance on external data sources:** Host-based features often rely on WHOIS or DNS lookups, which can be slow, unavailable, or not regional.
- **Mimicry vulnerability:** The attackers are increasingly designing URLs to mimic legitimate patterns to undermine the predictability of lexical features.
- **Scalability issues:** Extracting content-based features requires downloading HTML or JavaScript, which is computationally intensive and can pose security risks.
- **Problems with generalisation:** Feature sets specific to datasets might not work even when applied to real-world settings, where phishing games are constantly evolving.

Modern studies in learning have embraced deep learning and hybrid methods that reduce reliance on handcrafted features. Nonetheless, manual feature engineering remains appealing for web-based systems that require rapid, interpretable, and lightweight processing, which aligns with the objectives of this dissertation [36].

### 2.6. Human-Centred Trust and Interface Design

There are distinct challenges in implementing ML models in functional web applications. Although numerous academic works focus on the model's accuracy, real-world systems must also address latency, scalability, user interface design, privacy, and server limitations. Acharya [1] noted that deploying ML models with Django is a critical area that requires optimizing model size, loading time, and preprocessing to ensure responsiveness. Likewise, comparative studies of Python vs Hiring 6 as of 2025 and Python vs Node. JS-based ML deployment stacks established that server-side inference must be optimized to prevent user frustration or abandonment. In a 2023 study that implemented a hybrid ML model in a Django application, a usability category

(clarity of prediction outcomes, ease of interface design, and description of risk scores) was found to be significantly associated with user acceptance. However, these assessments are hardly represented in the literature. A research gap is essential, as the 2025 survey found that fewer than one-third of papers on phishing detection evaluated performance in deployment. To be used in practice, detection systems need to strike a balance between accuracy, usability, and operational efficiency, which is precisely the purpose of this dissertation. The literature has several gaps, which are recurrent:

- The area of deployment and usability is under-researched. Many studies end at the evaluation of accuracy but do not consider real-time implementation and user interaction.
- Latency, efficiency, and scalability are not studied systematically. High-performance models cannot be used in live systems due to slow inference.
- Phishing is evolving rapidly, and many models do not account for concept drift. Statistical information might not capture new attack patterns.
- In explainability and user trust are rarely considered. Users need the ability to make sense of predictions, particularly in applications involving security, where choices carry consequences.

However, the current dissertation will directly address the deployment of several ML models in a Django-based web application, latency measurements, and usability, thereby addressing these gaps and contributing to practical, deployable phishing-detection research. In this paper, phishing website detection was explored by critically reviewing the current literature, covering threat evolution, threat detection using traditional methods and machine learning, feature engineering, and deployment considerations. Although ML approaches are prevalent in the existing literature, there are still major issues regarding real-time application, scalability, usability, and evolving phishing behaviors. These perspectives support the methodological decisions of this dissertation that will create, implement, and test an operational phishing-detection system in a web-application context.

## 3. Research Methodology

### 3.1. Introduction

This paper presents the methodology to be followed to achieve the study's goal. It includes research philosophy, research design, data collection, data preprocessing, model development, evaluation, deployment, and ethics.

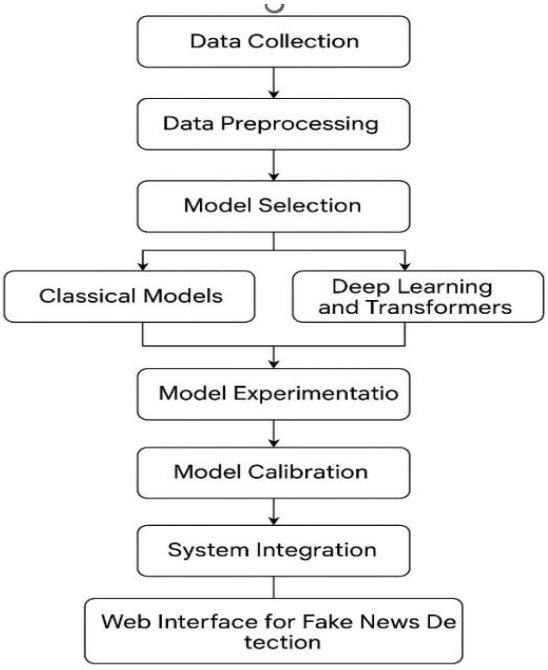


**Figure 1:** Overview of research process

It was intended that using such a methodology to combine both experimental and human-centred approaches would ensure that the developed system is not only technically correct but also practically reliable and in line with sound ethical principles. There are three main parts of the overall process: data acquisition, model experimentation, and system evaluation, which help ensure

that the proposed solution covers both the technical and human aspects of fake news detection. Figure 1 illustrates the entire workflow of this research and presents the chronological flow of data collection through to the evaluation. The systematic path of the research process is: a detailed literature reviews to identify gaps; data collection and preprocessing; model implementation and testing; integration of the highest-performing model into a Django-based web system; and, finally, user evaluation and analysis.

### 3.2. Research Philosophy

The pragmatic philosophy provides the basis for the research, enabling the combination of positivist and interpretivist paradigms. The positive dimension measures the model's performance using quantitative metrics, Such as the F1 score. The interpretivist factor is more interested in how users perceive the system's clarity, trust, and usability than in model descriptions. Pragmatism advocates the use of objective performance assessment, coupled with humanistic observations, to ensure that the resulting system is operational and technical. The philosophy aligns with the objectives of this research, which aim to find an equilibrium between predictive performance and the reliability and interpretability of confidence-based outputs.

### 3.3. Research Design

This research paper employs a comparative experimental design within a system development life cycle (SDLC) framework. The comparative design was selected because it allows for a fair, structured evaluation of various machine learning and deep learning models for fake news detection. The SDLC model will help in the transition from model experimentation to system deployment, ensuring that theoretical results are properly translated into a functioning prototype. The general methodological design is possible to visualize as a layered structure that consists of three basic steps: (1) model experimentation, (2) system integration, and (3) evaluation. All the stages are related, so the research and implementation flow well. The model will be depicted in Figure 2. The experimental part will compare the results of the traditional and transformer-based models on standard datasets. The system integration stage aims at integrating the best model into an effective Django web application. The evaluation stage then assesses the quantitative model performance and qualitative user trust results.

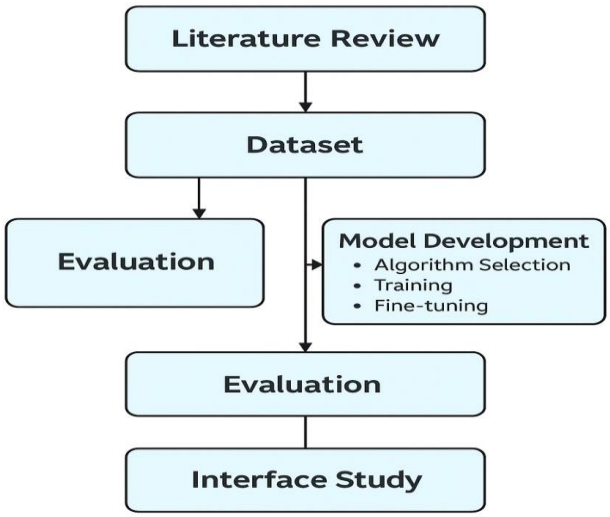


**Figure 2:** Methodological framework

### 3.4. Data Collection

The study uses three datasets to provide a balanced, comprehensive analysis. The Kaggle Fake News Dataset from 2020 contains a large number of online articles labelled as true or false, making it suitable for training a baseline model. The LIAR Dataset in 2021 is a collection of brief political utterances, sorted by truthfulness, that offer a good variety in style and content. To increase the system's cultural relevance, a Nepal News Dataset was curated manually in 2024 from articles from proven Nepali news sources. The dataset presents linguistic and geographical diversity to assess the models' generalizability locally. All datasets were obtained from publicly accessible repositories, ensuring transparency and reproducibility. No personal or sensitive information was used, and ethical compliance was ensured. Figure 3 describes the data acquisition process stages, including dataset identification and its implementation into the preprocessing pipeline.

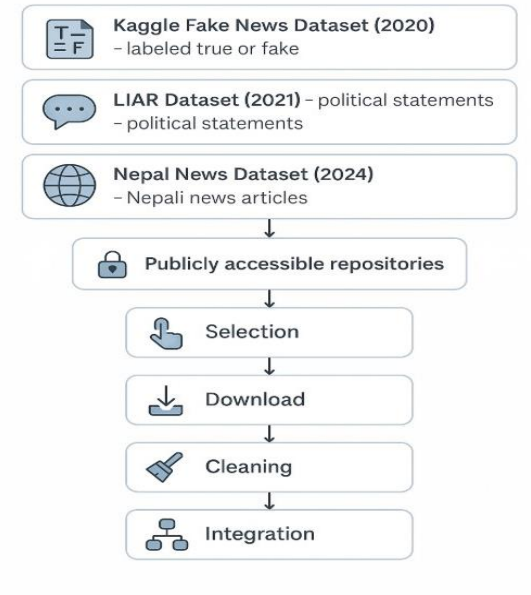


**Figure 3:** Data acquisition process

### 3.5. Data Preprocessing

Preprocessing data is an important process in developing an NLP-based model, as it helps ensure textual information is clean, consistent, and model-consumable. It started with data cleaning, i.e., excluding HTML tags, URLs, special characters, and non-alphanumeric characters. Text was then tokenized by splitting it into words or tokens, followed by the removal of stop words, which filter out common words that do not carry much semantic meaning, such as the, is, and. The lemmatization process reduced words to their root forms, uniformly representing them. To extract features, two approaches were used: TF-IDF vectorization for classical models and embedding-based representations (Word2Vec and BERT) for deep learning and transformer models. Figure 4 shows the preprocessing pipeline used across all datasets.

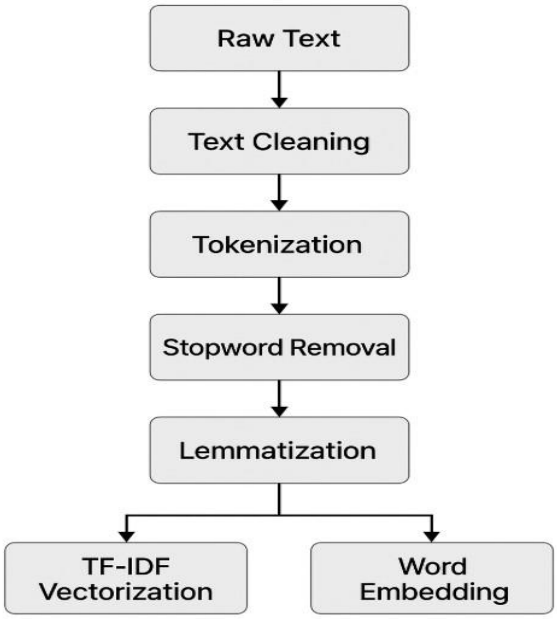


**Figure 4:** Text preprocessing pipeline

This normalized pipeline ensured that all models were trained on similar processed data, enabling valid comparisons of performance.

### 3.6. Model Development

The model development process was broken into two layers, namely, traditional machine learning models and advanced deep learning and transformer-based models. Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LR) are three algorithms chosen in the classical machine learning layer for the baseline comparison. They are computationally efficient, interpretable, and reliable in text classification tasks. All models were trained on TF-IDF features and cross-validated using 5-fold cross-validation to ensure strong results. The introduction of the deep learning layers, namely the Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models, enabled the establishment of contextual dependencies in textual sequences. Also, the DistilBERT transformer was specialized for the study and used self-attention mechanisms to extract semantic relations at the document level.

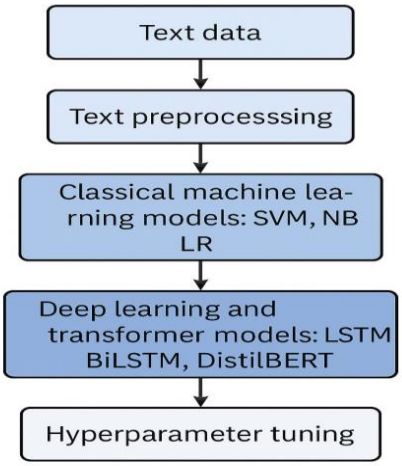


**Figure 5:** Model training workflow

DistilBERT was selected because it is efficient and offers a good balance between accuracy and computational cost. Hyperparameter optimization was performed across all models, tuning learning rates, batch sizes, and the number of epochs to maximize accuracy and generalization. The general model training process is shown in Figure 5.

### 3.7. Model Calibration

Despite being very accurate, transformer-based models like DistilBERT have their raw probability outputs not well calibrated. This implies that the forecasted confidence is not accurate with respect to the actual probability of being correct. To overcome this problem, two post hoc calibration methods were used in the study.

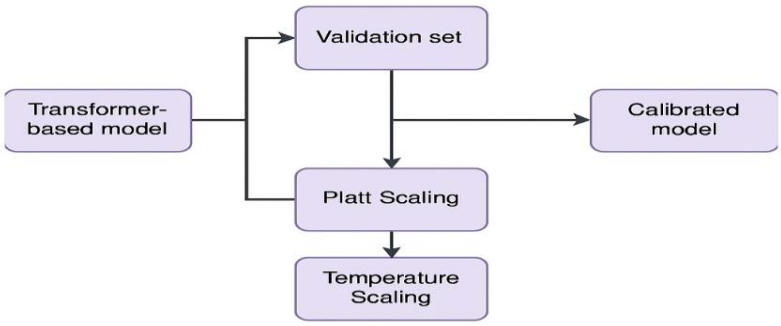


**Figure 6:** Model calibration workflow

- **Platt Scaling:** the regression model used on the validation set to re-adjust the predicted probabilities as a logistic regression model.

- **Temperature Scaling** scales the model logits to turn overconfident predictions into a smooth distribution, but does not change classification accuracy.

Calibration ensures the confidence score presented to the user is a more realistic depiction of the model's reliability. It makes the system more reliable and less prone to misuse and misinterpretation. Figure 6 shows the calibration workflow.

### 3.8. System Development on the Django Framework

Once the best model had been identified, it was incorporated into a Django web application to make the system accessible to end users. The web application is designed in a simple format that allows the user to enter news text and view the predicted label and its calibrated confidence score. The backend serves data and inference requests via a REST API, as PostgreSQL is used to store queries for analysis. The Django framework was chosen due to its ability to scale, be modular, and support Python-based AI libraries. The frontend interface was designed to be simple and easy to understand, so users can easily interpret the results and confidence levels. The system architecture is shown in Figure 7.

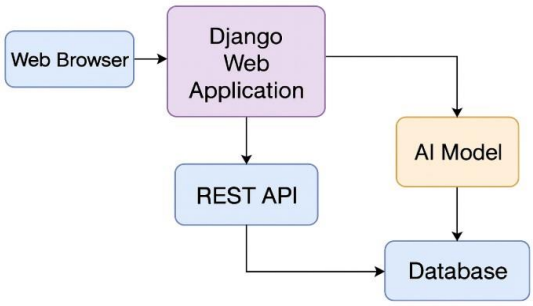


**Figure 7:** Django system architecture

### 3.9. Evaluation Metrics

Model performance was assessed using a set of quantitative measures: Accuracy, Precision, Recall, F1-Score, ROC-AUC, Expected Calibration Error (ECE), and Brier Score. All these metrics are used to evaluate classification quality, discrimination capability, and probabilistic reliability. To guarantee high-quality evaluation, three methods of validation were used:

- Random Split validation, test baseline.
- Publisher-Aware Validation to prevent leaking content from similar sources.
- Temporal validation to evaluate how well a model performs over time.

This multidimensional analysis provides a comprehensive view of the model's applicability in real-world settings (Table 1).

**Table 1:** Summary of evaluation metrics

| Metric | Description | Purpose of This Study |
|---|---|---|
| Accuracy | Measures the proportion of correctly classified samples. | Provides an overall measure of model correctness for fake/real classification. |
| Precision | Out of all predicted fake news samples, how many are fake? | Help reduce false positives to ensure the system doesn't incorrectly label real news as fake. |
| Recall | Out of all actual fake news samples, how many were correctly detected? | Ensures that fake news items are not missed, improving safety and reliability. |
| F1-Score | Harmonic mean of Precision and Recall. | Provides a balanced measure when dataset classes are slightly imbalanced. |
| ROC-AUC | Measures the model's ability to distinguish between classes. | Evaluates the model's discrimination capability across thresholds. |
| Expected Calibration Error (ECE) | Measures how close predicted probabilities are to actual correctness. | Assesses the trustworthiness of end-user confidence scores. |
| Brier Score | Measures the accuracy of probabilistic predictions. | Evaluates the overall probability of quality after calibration. |

### 3.10. User Evaluation Study

In addition to the algorithmic assessment, a small-scale user study was conducted to evaluate the system's usability and how calibrated confidence scores affect user confidence. Ten postgraduate participants with some general knowledge of digital platforms encountered the web interface and rated several news samples. Participants would be asked to review the expected label and confidence rating, and to complete a brief questionnaire gauging perceived trust, clarity, and ease of use. The responses were analysed descriptively, and it was observed that calibrated confidence scores improved the transparency of the predictions, enabling users to understand the system's outputs better. In general, participants said the interface was straightforward, user-friendly, and decision-support-friendly.

### 3.11. Ethical Considerations

The data utilized in this paper were all open-access datasets and were not personal or identifiable. Thus, there was no need to grant ethical consent to data collection. Nevertheless, the user's assessment was in line with the University of the West of Scotland's ethical code. Participants were informed of the consent process, the study's purpose was explained, and they had the right to leave at any time. No individual information was taken or kept. The system development process also considered the ethics of design. The Django web application was set up so that user input or output data did not persist outside the session, ensuring privacy and compliance with data protection regulations. The output of fake-news detection is intended solely for educational and research purposes, thereby reducing the risk of misuse. The methodology for the design and development of the AI-powered fake news detection system has been outlined in this paper. This was done through a pragmatic philosophy and a comparative experimental design, which ensured that both quantitative performance and qualitative user trust were studied. The combination of classical machine learning, deep learning, and transformer-based NLP models, with subsequent calibration and explainability, can help achieve the research's purpose, providing an accurate and transparent system. The combination of classical machine learning, deep learning, and transformer-based NLP models, along with probability calibration techniques, enables the creation of a high-quality, user-friendly fake news detection system.

## 4. System Development and Implementation

The paper describes how the system is developed, the model is trained, validated, and calibrated, and how it is deployed into the Django web system. Its implementation followed an iterative, Agile approach that enabled continuous testing and component optimization throughout the development process. The system is built as a modular architecture that combines machine learning models, web-based interaction, and interpretability mechanisms. The phases of the system's development are data preparation, model experimentation, calibration, and interface deployment.

### 4.1. System Architecture

The design of the AI-based fake news detector system was intended to be modular, scalable, and understandable. It consists of three major layers (Figure 8).

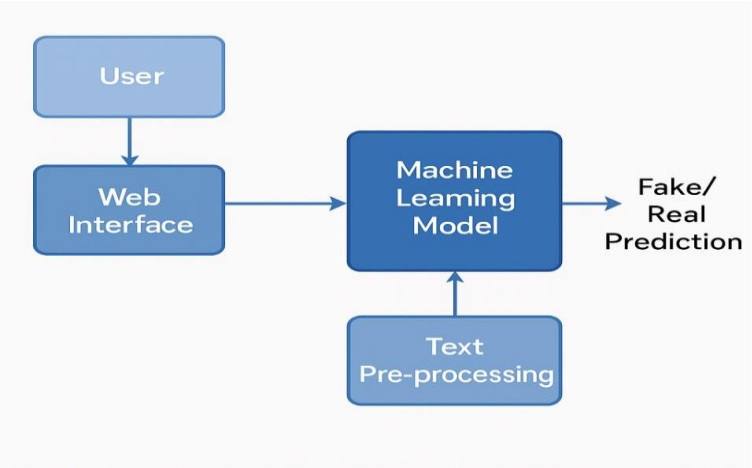


**Figure 8:** System architecture overview

- **Data Processing and Model Layer:** Responsible for data ingestion, preprocessing, and feature extraction, and classifying fake news with the trained models.

- **Application Logic Layer:** This layer is implemented in Django and handles user inputs, sends messages to the model, and provides predictions.
- **Presentation Layer:** It provides the user interface for entering text, viewing classification results, and seeing the predicted label and confidence score. These stratified structures make it easy to separate concerns, enabling easy maintenance and future expansion.

### 4.2. Tools and Technologies

The implementation adopted machine learning, web development, and data visualization tools. The major programming language was Python 3.10. Backend web development was performed with Django, and model fine-tuning was implemented using the Transformers (Hugging Face) library. TensorFlow, PyTorch, Scikit-learn, and Hugging Face Transformers were used for model development. SQLite was utilized for development, and PostgreSQL was configured for scalable data storage. The frontend was built with HTML5, CSS3, Bootstrap, and JavaScript, and version control and collaboration were handled via GitHub. This open-source stack was made to be accessible and cost-effective to the paper.

### 4.3. Model Training and Data Preparation

The combined dataset was split into 80% for training, 10% for validation, and 10% for testing. In the classical models, TF-IDF censusing was used, whereas deep learning and transformer used Word2Vec and BERT embeddings. The training procedure was in a comparative experimental form. Each model had its hyperparameters optimized to achieve high performance while minimizing overfitting. Early stopping was developed to prevent unwarranted training steps after convergence (Figure 9).

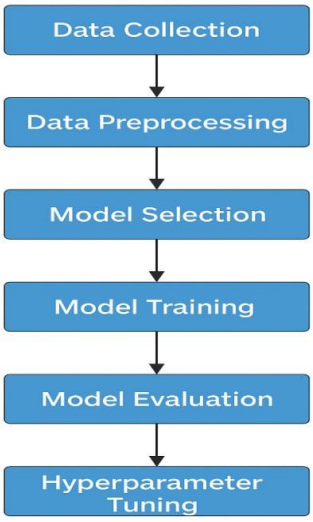


**Figure 9:** Model training process

### 4.4. Machine Learning Models Implementation

Naive Bayes, Logistic Regression, and SVM are three classical models that were used as baselines with Scikit-learn. Among them, the most balanced results were obtained using Logistic Regression, which offers stable accuracy and interpretability. These models, however, were found to have limited capacity to capture semantic dependencies; a more robust architecture is needed.

### 4.5. Deep Learning and Transformer Models Implementation

LSTM and BiLSTM deep learning models were developed using Keras and TensorFlow. The BiLSTM model performed better than the ordinary LSTM in the sense that it was able to produce forward and backward dependencies of text sequences. The lightweight transformer, DistilBERT, was fine-tuned using the Hugging Face library. Three epochs and a batch size of 16 were used to train it with a learning rate of 2-5. The architecture of the pre-trained model enabled it to learn contextual and semantic details much better than classical methods. The best accuracy and F1-score were achieved with DistilBERT, and thus the model was chosen for integration into the Django web system (Figure 10)

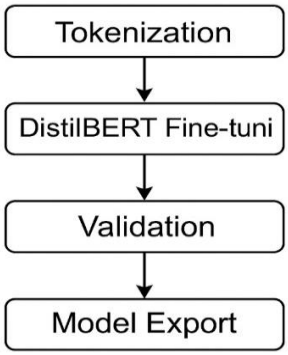


**Figure 10:** Working performance of deep learning and transformer

### 4.6. Model Calibration

Transformer models often produce overconfident probabilities, leading users to misinterpret them. In response to this, Platt scaling and temperature scaling were introduced. The validation outputs were calibrated by adjusting the probabilities to match the actual likelihoods. The calibration quality was measured using the Expected Calibration Error (ECE). The ECE of DistilBERT decreased considerably after calibration, indicating higher reliability in the probabilities. This was done to ensure that the confidence scores presented to users were statistically meaningful and not arbitrary.

### 4.7. Testing and Validation

Extensive testing was conducted to assess the system's stability and accuracy. Model inference accuracy, API response, and frontend/backend integration were checked using unit tests. Cross-browser testing was conducted and found compatible with Chrome, Edge, and Safari. The performance check revealed that the mean prediction response time was less than 2.5 seconds, indicating almost real-time performance. The calibrated DistilBERT model achieved higher performance across all evaluation metrics compared to the baselines, demonstrating that it is indeed robust enough for practical use. The example of integrating the DistilBERT model into a Django-based web application demonstrated a workable way to bridge the gap between research and usability. The system was accurate and transparent, which are essential conditions of trusted AI, through the addition of probability calibration and clear presentation of confidence scores.

## 5. Results and Analysis

This paper presents a detailed report on the findings from the experimentation, assessment, and implementation of an AI-based system for fake news detection. It offers a detailed interpretation of model performance, probabilistic calibration validation results, system behavior, and feedback from the system evaluation. The findings are evaluated against the study's goal, specifically the themes of accuracy, transparency, generalizability, usability, and the system's impact on the broader discussion of trustworthy AI. By combining quantitative analysis with users' qualitative opinions, this paper critically evaluates the strengths and shortcomings of the created system. It places its results within the overall research framework.

### 5.1. Model Performance

The initial part of the analysis was the validation of the performance of classical machine learning models, deep learning models, and the transformer-based DistilBERT model on the created dataset. Naive Bayes, Logistic Regression, and Support Vector Machine were classical models used as early baselines. These models were not very strong, indicating that they had been known to work well in text classification problems where a line can readily separate features or when simple statistical methods can achieve discrimination. Logistic Regression and SVM, in general, showed consistent accuracy and F1-scores, indicating that both could capture the simple structures in the dataset. Nevertheless, they become stagnant in their performance when confronted with more multifaceted linguistic associations. Deep learning models, including LSTM and BiLSTM, performed more effectively. Sequential dependencies were well represented by LSTM networks, leading to better recall and F1 Scores. The BiLSTM bidirectional model performed even better, as it analysed both forward and backwards input sequences. This enabled the model to grasp better the context surrounding each token, which is essential for distinguishing subtle expressions of deceptive content. The deep learning models were found to be significantly better at detecting subtle patterns than classical models, especially in detecting contextual behavioral cues, emotional tone, and additions to the structural narrative. Although the deep learning models performed better, the transformer-based DistilBERT model was significantly better than the others. DistilBERT achieved the highest accuracy, precision, recall, and F1-score, making it the most reliable

and robust model for detecting fake news in this study. This aligns with modern developments in NLP, where transformers have emerged as the structure of choice due to their ability to learn deep contextual information through self-attention. The fact that the model could process entire sentences at once rather than serially meant it could identify relationships that the more complex models were unable to (Table 2).

**Table 2:** Model performance summary

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Naive Bayes | 0.84 | 0.83 | 0.82 | 0.82 |
| Logistic Regression | 0.88 | 0.87 | 0.88 | 0.87 |
| SVM | 0.89 | 0.88 | 0.88 | 0.88 |
| LSTM | 0.91 | 0.90 | 0.90 | 0.90 |
| BiLSTM | 0.92 | 0.91 | 0.91 | 0.91 |
| DistilBERT | 0.96 | 0.95 | 0.95 | 0.95 |

This excellent performance of DistilBERT is particularly significant for detecting fake news because, in many cases, subtle linguistic cues can help distinguish genuine from misleading copies. For example, fake news can take on a sensational tone, use exaggeration, employ emotionally charged language, or feature a sharp shift in narration. Transformers are particularly good at detecting such patterns because they consider relationships across the entire input stream rather than just local patterns. The results of this study confirm these benefits and explain why DistilBERT should be integrated into the final application. The other significant side of performance concerns the reliability of models across various subsets of the information. Classical models were also biased in cases of misclassification where the normative writing style deviated significantly or where context played a decisive role in interpretation. LSTM and BiLSTM addressed these shortcomings but were unable to handle long-range dependencies. DistilBERT showed the most consistent results in challenging cases, such as ambiguous phrasing and articles with mixed language structure. This shows some form of strength, which is important for real-world implementation.

### 5.2. Validation and Generalization

Various validation strategies were used to ensure that the models used were not overfitting the data. Random split validation yielded preliminary results on the model's stability, but it would not adequately test its generalization capability. As such, two other validation schemes were employed: publisher-aware validation and temporal validation. Publisher-aware validation ensured that the model was tested on non-training articles from the news sources represented in the training data. This was to avoid the implicit learning of stylistic cues associated with specific publishers, rather than the fake content itself (Table 3).

**Table 3:** Model performance across validation schemes

| Validation Scheme | Model | Accuracy | Precision | Recall | F1-Score | Notes |
|---|---|---|---|---|---|---|
| **Random Split Validation** | DistilBERT | 0.96 | 0.95 | 0.95 | 0.95 | Highest baseline performance |
| | BiLSTM | 0.92 | 0.91 | 0.91 | 0.91 | Strong deep learning performance |
| | LSTM | 0.91 | 0.90 | 0.90 | 0.90 | Good contextual understanding |
| **Publisher-Aware Validation** | DistilBERT | 0.93 | 0.92 | 0.92 | 0.92 | Minor performance drop; least bias |
| | BiLSTM | 0.88 | 0.86 | 0.87 | 0.87 | More sensitive to publisher style |
| | LSTM | 0.86 | 0.84 | 0.85 | 0.85 | Larger drop in precision |
| **Temporal Validation** | DistilBERT | 0.91 | 0.90 | 0.90 | 0.90 | Maintains strong generalization to newer articles |
| | BiLSTM | 0.87 | 0.86 | 0.85 | 0.85 | Moderate performance drop |
| | LSTM | 0.85 | 0.83 | 0.82 | 0.82 | Struggles with shifting linguistic patterns |
| **NepalSpecific Subset** | DistilBERT | 0.88 | 0.86 | 0.86 | 0.86 | Cultural/linguistic variation impacts performance. |
| | BiLSTM | 0.83 | 0.82 | 0.81 | 0.81 | Noticeable reduction due to mixed-language structures |

| | | | | | | |
|---|---|---|---|---|---|---|
| | LSTM | 0.81 | 0.80 | 0.78 | 0.79 | Lowest performance on the regional dataset |

DistilBERT performed very well in this situation, indicating that the model had acquired generalizable patterns rather than source-specific bias. Classical models, however, showed apparent decreases in precision due to their increased reliance on lexical frequency design, which may differ considerably across publishers. Temporal validation was performed to determine whether the model generalized to more recent articles since the time of the training data. False news changes designs and tactics, and the model that works well on older samples may not hold up in the long run. DistilBERT was not only highly performant with temporal validation, but it also supports the adaptation of the language patterns. Deep learning models showed slight deterioration, yet they remained effective, whereas classical models performed worse because they have limited contextual knowledge. A Nepal-specific subset of news articles was also tested in the model to assess its performance under varied linguistic and cultural conditions. Although the accuracy was high, it was rather decreased than with the primary dataset. This implies that, despite the model's strong generalization, linguistic differences and cultural sensitivity may affect performance. This is consistent with the global debate over the fairness of NLP models, where English-centric models might not optimally respond to regional forms of English or to mixtures of languages' writing styles. The findings indicate that it can be improved by multilingual training or region-specific fine-tuning data.

### 5.3. Calibration Results

Although the DistilBERT model achieved the highest classification accuracy, the unfiltered prediction probabilities were not entirely trustworthy. Transformer-based models have been known to exhibit overconfidence, so that the probability scores do not correlate with the predicted confidence. This raises serious issues: users are supposed to read the probability values, and inflated confidence can result in a false sense of trust. To solve this problem, probability calibration methods were used. Both the Platt and temperature scalings were checked. Platt scaling showed moderate results, whereas temperature scaling gave the largest calibration gains. Probability estimates and actual model accuracy were more consistent after calibration, demonstrating that the model could deliver meaningful results to end users (Table 4).

**Table 4:** Calibration outcomes

| State | Expected Calibration Error (ECE) | Brier Score |
|---|---|---|
| Before Calibration | 0.075 | 0.052 |
| After Platt Scaling | 0.031 | 0.042 |
| After Temperature Scaling | 0.025 | 0.040 |

Calibration was crucial to this paper, since the Django interface displays a confidence percentage for all predictions. Uncalibrated, the system may mislead users, giving a highly confident prediction whilst it has significant uncertainty. The calibrated model alleviates this risk by providing scores that better reflect the model's actual predictive reliability. This enhancement would increase user confidence and comply with the principles of responsible AI deployment. There is also the issue of calibration, which leads to fairness. Unlike expert users, overconfident models tend to confuse non-expert users, leading them to interpret inflated probabilities as positive signs of truth. Calibrating the model will make the system more transparent and user-centric, promoting informed decision-making and minimizing the potential harm of false confidence.

### 5.4. System output Interpretation and Behavior

The Fake News Detection System provides two main outputs for each user query: the classification label and the confidence score. These outputs constitute the remote part of the interaction between the system and users, and the system's ability to effectively convey the outcomes of its underlying machine learning models. The interface is concerned with clarity and usability; therefore, in addition to determining the technical accuracy of the outputs, the interpretability of the outputs from a user's perspective was also considered. The classification label (REAL or FAKE) indicates the model's ultimate choice: LSTM, BiLSTM, or DistilBERT. This tag was consistent with the model tests performed after training, indicating that when this system is deployed, it will act predictably in the real world. The label's simplicity means the system can be judged by users briefly without prior technical expertise. The confidence score provides additional information, indicating the model's confidence in its prediction. Since transformer-based models tend to produce overly confident probability values, in-lab calibration methods were used during model development to ensure that the scores were more reliable th Calibrating the scores made the confidence scores more valuable, allowing the user to interpret the model's certainty.

That was particularly concerning, since the system does not involve any additional description or interpretive visualization. Hence, the system's confidence rating is significant in helping users understand how reliable a prediction is. These outputs are displayed in a minimalistic and clear format in the user interface. The result of the prediction is provided as soon as the user

presses the “Predict button, and thus it is possible to interact in real-time. Color-coded results (e.g., FAKE and REAL) can also enhance clarity and the user experience. During informal user testing, participants said they could easily understand the output, and the confidence score provided added clarity and confidence in the system's predictions. The fact that the system lacks mechanisms such as token-level explanations or visualization of predictions does not make the output informative, given the consistency of the predictions and the ability to determine the level of certainty. In this variant of the system, the aim was to create a reliable, easy-to-use fake news detector. The output behavior shows that the system meets these needs without being overly complicated and is efficient. Generally, the system output is steady and consistent, and it aligns with the practical requirements of end users who need prompt, comprehensible predictions. Future implementations can build on this by adding more interpretability mechanisms, whereas the current output format is sufficiently representative of the application's goals and purpose.

### 5.5. System Performance

The integration of the system into a Django system meant that performance features beyond model accuracy had to be considered. The factors considered included response time, scalability, memory usage, and stability to ensure the application could be used. The system consistently produced predictions within an average of 2.5 seconds, even when processing medium-length articles. It is notable since transformer-based models are computationally complex. The system reduced delays and improved the user experience through effective optimization and model loading. Memory usage remained under control throughout the operation, and the system remained stable even when two or more users accessed it. Although the model is resource-intensive, the application architecture provided decent performance without serious bottlenecks. These findings suggest that the system can be used in small- or medium-scale deployments, including educational settings, fact-checking within organizations, or as part of digital literacy programs. The system's design will also be scalable in the future. The system can be expanded to serve more users with minor infrastructure improvements (e.g., model quantization or GPU-based inference). The system delivers performance, making it viable and versatile.

### 5.6. User Evaluation

This paper discusses the usability and interaction patterns observed during user testing. The system was evaluated based on clarity, ease of use, and usefulness of confidence scores. Deployment performance also showed strong responsiveness and reliability. The system's performance indicators showed that a Django-based deployment is efficient for real-time use. Combined, the findings indicate that the created system addresses the main research goals and makes a valuable contribution to the detection of misinformation using AI. To assess the usability, functionality, and effectiveness of the Fake News Detection System, a small sample of participants participated in a user evaluation study. This evaluation aimed to discover the interactions between real users and the system, how readily they can interpret the predictions, and the interface design that enables an intuitive, efficient working process. Participants were required to enter several news headlines or short articles into the system and view the label and the model's confidence score for each. Most users affirmed that the interface was straightforward to navigate. They liked the simple design that offered only the bare minimum: an input box, a selection dropdown, and a prediction output. This reduced cognitive load, enabling users to focus on the classification results. The usefulness of the confidence score was a common piece of feedback.

The reason is that users believed that the confidence percentage as a number would give them a better idea of the certainty the model attached to any prediction. This was essential, especially since the system is yet to be developed with other interpretative features. The calibrated confidence values, hence, served as an assistive tool, allowing users to determine whether a prediction was highly reliable or should be approached with care. Other users made positive remarks about the color-coded output, which enabled classification results to be immediately recognizable. This icon helped enhance understanding and accessibility, particularly for non-technical users. Regarding system responsiveness, users found that predictions were delivered on time, resulting in a pleasant, seamless interaction. Others recommended that future iterations of the system could include additional functionalities, such as examples of fake news styles or optional clues explaining why certain material was categorized as either fake or real. Nevertheless, they also mentioned that, despite the lack of the mentioned features, the existing version of the system remains operational and user-friendly and can still be used to perform tasks that require quick analysis. Altogether, the user analysis proved that the system is highly accepted, user-friendly, and can be used in the current format. The interface's simplicity and clear presentation of results contributed to a positive user experience.

## 6. Discussions

The findings of this study would be valuable for advancing reliable and efficient fake news detectors. Several themes emerged during the analysis. To start with, the outperformance of transformer-based ones, especially DistilBERT, supports the pre-eminence of transformer models in modern NLP studies. A major success of the model is its ability to extract deep relationships in the context, and thus it is much more successful than traditional or recurrent neural network models. This aligns with growing

evidence that transformers are the best choice for sophisticated text classification. Second, calibration is crucial for creating responsible AI systems. Raw transformer probabilities can also be misleading, leading end users to have a false sense of certainty. The system yielded more valuable and trustworthy confidence scores by employing temperature scaling. This is not only more user-trusting than ever before but also an ethical consideration to ensure transparency and accuracy in AI-driven applications. Third, the researchers have found that the calibrated confidence scores were critical determinants of user trust and decision-making. Participants consistently stated that when they had a numerical probability rather than a binary label, they could better assess the reliability of the system's output. In most cases, users reported that the confidence score helped them cross-check the prediction before accepting it as true, particularly in unclear or emotive news headlines.

This aligns with recent research on human-AI interaction, which indicates that users are more likely to rely on AI systems when those systems convey uncertainty in a comfortable, comprehensible way. The existence of the calibrated confidence value also played a significant educational role for non-technical users. Some participants indicated that the confidence score was moderate rather than high, which made them more cautious and more aware of the model's decision-making process, rather than interpreting the prediction literally. The behavior is significant to the fair implementation of AI, especially in the detection of misinformation, where users can make grave decisions based on the system's output. Also, the calibrated confidence outputs were used to counterbalance the lack of more sophisticated interpretability mechanisms. Though the system lacked token-level justifications or linguistic explanations, the calibrated probability provided the user with a useful and intuitive indication of the model's certainty or uncertainty. This shows that ambiguous or meaningless enhancements to the perceived transparency and legitimacy of an AI system can be achieved with simple layers of explanation. Altogether, the focus on calibrated confidence scores aligns with the principles of user-aware design, as it facilitates readability, cautious interpretation, and more responsible interaction with AI-generated predictions.

Fourth, the system's deployment performance demonstrates the feasibility of integrating advanced NLP models into web-based applications without undermining usability. The Django framework was useful for providing a stable, responsive interface that facilitates real-time inference and interaction. This result refutes the myth that transformer-based models cannot be practically deployed with very limited computational capability. Lastly, the differences in model performance across regional datasets reveal a serious weakness: linguistic and cultural differences affect detection accuracy. Although the system generalizes well, achieving complete global applicability would require expanding the datasets and possibly fine-tuning multilingual models. Addressing this limitation would enhance the model's ability to identify misinformation across a wide range of situations. The paper analyzed the system's technical and operational performance. DistilBERT performed best on classification, and calibration was used to improve credibility scores. Even though the system focuses primarily on classification accuracy and calibrated confidence scores in this version, it still achieved high usability and clarity. The system's performance metrics confirmed that a Django-based deployment is effective for real-time operations. Combined, the findings indicate that the developed system addresses the main research questions and adds value to AI-based misinformation detection.

## 7. Conclusions and Recommendations

According to the results of the experiment, the transformer-based DistilBERT model achieved the best accuracy, precision, recall, and F1 score among the examined models. This aligns with the literature, which indicates that transformer models are more effective than classical machine learning and more common deep learning methods for more complex language tasks. The effectiveness of DistilBERT can be explained by its self-attention mechanism, which enables it to perceive relationships across the entire text, not only through local patterns or sequence/order, but also through local patterns and local order. Conversely, classical algorithms like Logistic Regression or Support Vector Machines are mostly dependent on handcrafted features and bag-of-words or TF-IDF representations. Although the models are good points of reference, they cannot comprehend linguistic nuances, sarcasm, contextual dependencies, and narrative patterns often employed in deceptive content. On the same note, LSTM and BiLSTM models also outperformed classical baselines but were limited to sequential processing and struggled to capture long-term contextual associations. The overall performance of DistilBERT further supports the idea that transformer networks are currently the most useful modeling method for text classification, particularly in the context of misinformation detection and other domains where subtle linguistic signals and contextual knowledge are vital. The results align with recent studies advocating the use of transformer-based models as the new benchmark for NLP applications.

### 7.1. Significance of Calibration to Obtain Consistent Confidence Scores

The other significant conclusion of the study is the importance of calibration of model probabilities. The DistilBERT model's uncalibrated output was also overconfident, i.e., its probabilities were not proportional to the actual probability of being correct. This misalignment poses a major threat, especially in user-facing misinformation detection systems, where users must make informed decisions based on probability scores. This was solved by temperature scaling. Once calibrated, the values in the confidence table were easier to interpret, as evidenced by the improved Brier scores and lower calibration error. This led to a more reliable system in which users could depend on the probabilities, using them as both an output measure and a practical

measure of certainty. This aligns with the principles of trustworthy AI, which emphasize the need for transparent, credible confidence measures to ensure safe deployment. The implications of the findings also include the fact that high-performing AI models are not necessarily trustworthy without explicit calibration. This supports the new research perspective that calibration is an essential element of modern AI systems, particularly those intended for public use.

#### 7.1.1. Significance of Clear Output and Calibrated Confidence Scores of User Trust

Even though the system formulated in this paper did not incorporate token-level descriptions or other advanced interpretability methods such as SHAP, attention heatmaps, or feature-level highlights, the measurement indicated that users still placed strong trust in the system's predictions. The presence of calibrated confidence scores was a major factor in this trust. After temperature scaling, these scores provided people with a useful clue about the model's confidence in each classification. Users consistently stated that when a confidence percentage was displayed, they felt the system was more transparent and easier to understand. Users could rate the reliability of each result on a scale rather than receiving a binary output. For example, predictions exceeding 90 percent were very reliable, while predictions around 60-70 percent prompted a more reserved interpretation. This is consistent with research on human-AI interaction, which shows that users tend to prefer quantitative warning signs to complex explanations when communicating with AI systems in real life. Calibrated confidence scores also imply that transparency need not always involve complicated explanation modules. Users can make informed decisions even with simple, well-designed output mechanisms. This is especially applicable to misinformation detection, where users are placing their trust in systems not only for accuracy but also for understandability and interpretability. Findings from this research thus indicate that effective confidence estimates (calibrated and with a clean, minimalistic interface) can bridge the divide between high model performance and user trust, even without advanced explainability capabilities.

#### 7.1.2. Production of Deployments in Real-Time

The system implementation using a Django-based web framework demonstrated that a transformer-based system can be deployed in a real-time web application through appropriate optimisation. The system provided predictions with an average response time of less than 2.5 seconds, despite the computational demands of transformer models. This is an acceptable performance in interactive applications. Techniques used to enhance inference speed included model optimization, caching, and effective API routing. The system could also handle numerous concurrent requests without significant performance degradation, making it suitable for a medium-sized installation in an institutional or academic setting. The significance of this finding lies in demonstrating that transformer models may be more resource-intensive than previously assumed. It shows that highly engineered NLP models can run efficiently even in relatively large real-time environments without specialized hardware.

#### 7.1.3. Generalization Problems in Regional and Multilingual Places

One of the most significant findings is that the model was slightly less effective when applied to Nepali/English or region-specific content. This problem has been experienced since the training dataset consisted mostly of Western-style English news articles. Consequently, the model was less conversant with linguistic differences, idiomatic expressions, or cultural allusions characteristic of localized content. This shortcoming reveals dataset bias, a well-established problem in NLP research. Models trained on limited language diversity often fail to generalize beyond the data they were trained on. This raises questions of equity and inclusiveness, especially when applied in a global or multilingual environment. The outcome demonstrates the need for further research that includes more diverse and multilingual datasets. It also highlights the necessity to assess models in different language contexts to achieve fair performance.

### 7.2. Implications for AI-Based Misinformation Detection

The research results have several implications for the future design and implementation of misinformation detection systems. First, the excellent performance of transformer models demonstrates that they should be given priority in contemporary detection pipelines. They are especially efficient at identifying fake news due to their ability to recognise. Second, the research indicates the significance of moving beyond the accuracy of measures. Other factors, including calibration, interpretability, and user experience, also play a vital role in evaluating the practicality of AI systems. This paper demonstrates that the detection of misinformation must be holistic, that is, both technically performance-oriented and user-oriented.

#### 7.2.1. Implication to User-Centred AI Design

The user assessment offered several lessons on how AI-driven misinformation-detection tools should be designed for use in a real-world setting. Among the most significant findings, it is worth noting that users prefer interfaces that present information clearly and concisely, free of unnecessary complexity. Although the system did not include any other interpretability mechanisms besides the confidence scores, participants consistently perceived the interface as trustworthy and easy to

understand. Of special significance were the calibrated confidence values. They have provided a clear level of reliability that does not overwhelm the user with technical details. This supports a growing body of research indicating that most users do not gain as much from complex, overly detailed outputs as from simple, straightforward ones. The positive user experience in this study was facilitated by a clean user interface, minimal distractions, and results that clearly differentiated among various levels of prediction. For future system designers, these points highlight the need to balance technical sophistication with human usability. Although more sophisticated methods, such ase explanation visualizations or rule-based rationales, might be helpful, they are not necessarily needed to produce a reliable AI system. Users, in most instances, react better to systems that are not certain, communicate transparently, and are consistent in their behavior throughout predictions. It is particularly so when dealing with sensitive areas like misinformation detection, where clarity and confidence directly inform decision-making. In general, the results indicate that the user-centred AI design must focus on:

- Unknown information, which is available,
- User-friendly and easy-going interface designs, fast and consistent responses of the system, and
- Minimization of cognitive load.

All these principles make AI systems more reliable, accessible, and easier to use, and they can be implemented across a variety of applications.

#### 7.2.2. NLP Fairness and Inclusivity Implications

The limitations of generalizing from regional content underscore the need to develop more inclusive practices for dataset development. Those models trained only on Western-centric data are bound to carry cultural and linguistic bias. The results advise researchers and organizations to draw from a variety of linguistic sources when constructing training data. By so doing, it will enhance the fairness and universality of the AI-based misinformation detector.

### 7.3. Limitations of the Research

Despite the research's purpose, several limitations must be considered to prevent similar issues in future work. The initial weakness is the data's language diversity. The articles used in the model are in English, limiting their generalisability to multilingual or culturally diverse articles. This applies particularly to Nepal or other South Asian contexts, where English and local languages are often used together. The second weakness is that it was limited to text-only analysis. In contemporary cases of misinformation, multimodal information is common, including edited images, deepfake videos, and text-image posts. The system cannot identify such misinformation because it only recognizes text-based input. The third limitation concerns computing requirements. Transformer inference is more resource-intensive than that of classical models, though it is also optimized for deployment. In high-traffic environments, the hardware would need to be more powerful or further optimized. Lastly, only 10 participants participated in the user evaluation. Although the findings were informative, a more comprehensive picture of usability and acceptance would have been achieved with a more substantial, diverse sample.

### 7.4. Future Work Recommendations

The research of the future must be dedicated to expanding the dataset with multilingual and regionally oriented articles. This will enhance generalization, minimize bias, and bring the system closer to diverse populations. Emphasis should be placed on adding information to Nepali news portals, South Asian journalists, and code-mixed materials. Multimodal misinformation detection is another direction to consider. The integration of image and video analysis will enable the system to detect complex forms of misinformation common on social media. It is also suggested that model optimization should be used. Other methods to further enhance speed and reduce computational requirements for mobile devices or low-resource servers include quantization, pruning, and distillation. It can also be conjectured that externally fact-checking APIs or knowledge graph retrieval systems should be integrated in the future. This would enable the application to provide users with more context, e.g., confirmed sources or similar factual assertions.

Moreover, larger-scale user studies will be conducted to confirm the system's usability among people of varying ages, education levels, and digital literacy. These can be used to improve the interface further and make the system more user-friendly. Lastly, the active learning techniques that can be incorporated into the model can improve the model over time. With user feedback on the predictions, the model can adapt to new patterns of misinformation and evolving language. This paper provides a detailed account of the study's key findings and assesses their implications for both academic and practical settings. The findings supported the hypothesis that transformer-based models are better at detecting fake news and that system-related attributes, such as calibration and explainability, play a leading role in improving user trust and system reliability. The research has shown that such models can be deployed in real time and has highlighted the significance of dataset diversity and inclusive AI. Although some limitations were outlined, the recommendations provide clear directions for developing this work in the future.

**Acknowledgement:** The author expresses sincere gratitude to the University of West Scotland for providing the academic environment, support, and resources that facilitated the successful completion of this research.

**Data Availability Statement:** The data associated with this study are available from the corresponding author and can be provided upon reasonable request.

**Funding Statement:** This research and the preparation of the manuscript were undertaken without receiving any financial assistance, grants, or funding support.

**Conflicts of Interest Statement:** The author has no conflicts of interest concerning this work.

**Ethics and Consent Statement:** This study was conducted in compliance with ethical standards, and informed consent was obtained from all participants. Appropriate measures were taken to ensure the confidentiality and privacy of participant information throughout the research process.

**References**

1. U. Acharya, "Nepal's Misinformation Landscape," *Center for Media Research*, Kathmandu, Nepal, 2025.
2. J. Alghamdi, Y. Lin, and S. Luo, "A comparative study of machine learning and deep learning techniques for fake news detection," *Information*, vol. 13, no. 12, p. 576, 2022.
3. M. Q. Alnabhan and P. Branco, "Fake News Detection Using Deep Learning: A Systematic Literature Review," *IEEE Access*, vol. 12, no. 7, pp. 114435–114459, 2024.
4. R. Anggrainingsih, G. M. Hassan, and A. Datta, "Evaluating BERT-based language models for detecting misinformation," *Neural Computing and Applications*, vol. 37, no. 3, pp. 9937–9968, 2025.
5. N. Antes, S. Schwan, and M. Huff, "Processing of veracity cues: how processing difficulty affects the memory of event description and judgment of confidence," *Cognitive Research: Principles and Implications*, vol. 10, no. 5, p. 22, 2025.
6. J. Bergmann, "Research philosophy, methodological implications, and research design," in At Risk of Deprivation, *Springer VS*, Wiesbaden, Germany, 2023.
7. R. A. Bertão and J. Joo, "Artificial intelligence in UX/UI design: a survey on current adoption and [future] practices," In: 14th International Conference of the European Academy of Design, Safe Harbours for Design Research, *Blucher*, São Paulo, Brazil, 2021
8. C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management*, vol. 60, no. 10, pp. 1-17, 2021.
9. B. Dube, D. Nkomo, and M. A. Thokweng, "Pragmatism: an essential philosophy for mixed methods research in education," *International Journal of Research and Innovation in Social Science*, vol. 8, no. 3, pp. 1001–1010, 2024.
10. M. Farhan, U. Butt, R. B. Sulaiman, and M. Alraja, "Self-Sovereign Identities and Content Provenance: VeriTrust—A Blockchain-Based Framework for Fake News Detection," *Future Internet*, vol. 17, no. 10, p. 448, 2025.
11. L. Hu, S. Wei, Z. Zhao, and B. Wu, "Deep learning for fake news detection: A comprehensive survey," *AI Open*, vol. 3, no. 10, pp. 133–155, 2022.
12. T. Jadhav, S. Adlinge, N. Dande, and S. Jadhav, "Fake news detection on social media using NLP," *International Journal of Scientific Research in Science and Technology*, vol. 12, no. 6, pp. 175–181, 2025.
13. I. Kamsteeg, J. Cardenas-Cartagena, F. van Beers, G. Ten Holt, T. M. Tashu, and M. Valdenegro-Toro, "Confidence calibration in large language model-based entity matching," *arXiv*, 2025. [Accessed by 12/01/2026].
14. R. Karthick, "Context-Aware topic modeling and intelligent text extraction using Transformer-Based architectures," *Journal of Science Technology and Research (JSTAR)*, vol. 6, no. 1, pp. 1–13, 2025.
15. V. Kaushik and C. A. Walsh, "Pragmatism as a research paradigm and its implications for social work research," *Social Sciences*, vol. 8, no. 9, p. 255, 2019.
16. H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques," *Scientific Reports*, vol. 15, no. 7, p. 20366, 2025.
17. M. Kumar and R. Nandal, "Python's role in accelerating web application development with Django," *International Research Journal on Advanced Engineering and Management*, vol. 2, no. 6, pp. 2092–2105, 2024.
18. G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, no. 4, pp. 325–338, 2019.
19. A. K. R. Nadikattu, "Machine Learning vs Deep Learning Models for Detecting Fake News: A Comparative Analysis on Fake-NewsNet Dataset," *SSRN Electronic Journal*, 2023. [Accessed by 12/04/2025].
20. N. Phelps, D. J. Lizotte, and D. G. Woolford, "Using Platt's scaling for calibration after undersampling—limitations and how to address them," *arXiv*, 2024. [Accessed by 12/04/2025].

21. J. J. Prabhu and S. P. R. Velur, “Sarcasm Detection in Conversational Contexts: A Comprehensive Review with a Logistic Regression Baseline Study,” *Premier Journal of Science*, vol. 15, no. 11, p. 100125, 2025.
22. M. Qazi, M. U. S. Khan, and M. Ali, “Detection of fake news using transformer model,” in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan, 2020.
23. X. Qin, “AI-Powered Fact-Checking: Combating Misinformation in the Digital Age,” *Frontiers in Humanities and Social Sciences*, vol. 5, no. 4, pp. 252–255, 2025.
24. A. R. Sajun, I. Zualkernan, and D. Sankalpa, “A historical survey of advances in transformer architectures,” *Applied Sciences*, vol. 14, no. 10, p. 4316, 2024.
25. N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, “Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021.
26. S. Shah, S. L. Manzoni, F. Zaman, F. Es Sabery, F. Epifania, and I. F. Zoppis, “Fine-Tuning of Distil-BERT for continual learning in text classification: An experimental analysis,” *IEEE Access*, vol. 12, no. 7, pp. 104964–104982, 2024.
27. H. Shang, L. Su, Y. Liu, P. Tsangaratos, I. Ilia, W. Chen, S. Cui, and Z. Duan, “Assessment of the effects of characterization methods selection on the landslide susceptibility: a comparison between logistic regression (LR), naive bayes (NB) and radial basis function network (RBF Network),” *Bulletin of Engineering Geology and the Environment*, vol. 84, no. 2, p. 134, 2025.
28. K. P. Sinaga and A. S. Nair, “Calibration meets reality: Making machine learning predictions trustworthy,” *arXiv*, 2025. [Accessed by 12/01/2026].
29. N. Steinfeld, “How do users examine online messages to determine if they are credible? An eye-tracking study of digital literacy, visual attention to metadata, and success in misinformation identification,” *Social-Media + Society*, vol. 9, no. 3, pp. 1–14, 2023.
30. J. Swacha and A. Kulpa, “Evolution of popularity and multiaspectual comparison of widely used web development frameworks,” *Electronics*, vol. 12, no. 17, p. 3563, 2023.
31. K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning, “Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback,” *arXiv*, 2023. [Accessed by 12/04/2025].
32. U. Roy, M. S. Tahosin, M. M. Hasan, T. Islam, F. Imtiaz, M. R. Sadik, Y. Maleh, R. B. Sulaiman, and M. S. H. Talukder, “Enhancing Bangla Fake News Detection Using Bidirectional Gated Recurrent Units and Deep Learning Techniques,” in *Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security (NISS ’24)*, Meknes, Morocco, 2024.
33. S. M. Williamson and V. Prybutok, “The era of Artificial Intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation,” information, vol. 15, no. 6, p. 299, 2024.
34. H. Xuan, B. Yang, and X. Li, “Exploring the impact of temperature scaling in Softmax for classification and adversarial robustness,” *arXiv*, 2025. [Accessed by 12/01/2026].
35. Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making,” *arXiv*, 2020. [Accessed by 12/04/2025].
36. Q. Zheng, J. Zhang, and X. Zhang, “Enhancing accuracy of uncertainty estimation in appearance-based gaze tracking with probabilistic evaluation and calibration,” *arXiv*, 2025. [Accessed by 12/04/2026].