Solent University
COM726 – Dissertation: Research Project
Pilot Study Report (Feasibility Report)
Fake News Detection Using Machine Learning Approaches

Student Name: [Your Name]
Submission Date: 09 June 2025

Table of Contents

---

1. Introduction and Problem Definition

The rise of digital platforms and social media has revolutionized information dissemination, but it has also facilitated the rapid spread of misinformation and fabricated content, commonly termed "fake news." Fake news poses risks to democratic processes, public health, and societal cohesion by distorting facts and eroding trust in legitimate media sources (Allcott & Gentzkow, 2017). Traditional manual fact-checking efforts struggle to keep pace with the volume and velocity of online content. Hence, there is an urgent need for automated, scalable detection systems that can flag dubious content in real time and provide transparent rationale for their decisions.

Fake news detection is challenging due to linguistic nuance, multilingual contexts, and adversarial behaviors where malicious actors deliberately manipulate language or metadata to evade detection (Zhou & Zafarani, 2020). Mechanical application of static machine learning models often leads to performance degradation as new misinformation patterns emerge. Moreover, black-box classifiers fail to provide interpretable insights, undermining user trust and limiting adoption among journalists and fact-checkers.

This pilot study report examines the feasibility of developing an adaptive, explainable machine learning pipeline for automated fake news detection. It defines clear research questions and hypotheses, reviews existing research, proposes a technical solution, and outlines the resources and timeline necessary to complete the proof-of-concept. By integrating user feedback into a lightweight web interface, the study aims to demonstrate both technical viability and potential for real-world integration into newsroom workflows.

2. Background and Justification

The prevalence and impact of fake news necessitate a thorough understanding of its origins, detection challenges, and the limitations of existing approaches. This section is organized into four subsections to provide a comprehensive backdrop for the proposed pilot study.

2.1 The Fake News Phenomenon

Fake news refers to intentionally false or misleading content crafted to appear authentic. While misinformation can emerge accidentally, fake news implies deliberate deception, often for political or financial gain. During the 2016 U.S. election, disinformation campaigns propagated false narratives that reached tens of millions of users, demonstrably swaying public opinion and voter behavior (Vosoughi, Roy, & Aral, 2018). Similarly, the COVID-19 pandemic saw an explosion of health-related rumors—ranging from fabricated cures to conspiracy theories—exacerbating public anxiety and endangering lives through harmful self-medicating practices.

2.2 Early Detection Approaches

Initial automated efforts focused on rule-based systems analyzing surface-level linguistic cues such as sentiment polarity, punctuation irregularities, and simple keyword patterns. These systems achieved only modest detection rates, plagued by high false positive counts and an inability to adapt to evolving writing styles. Hybrid approaches that incorporated metadata—user credibility scores, source reputation, and sharing networks—yielded improved performance but introduced privacy concerns and dependency on proprietary platform data, limiting generalizability and real-time applicability.

2.3 Advancements with Deep Learning

The advent of deep learning, particularly transformer-based architectures like BERT and RoBERTa, catalyzed significant gains in fake news detection. Fine-tuning pre-trained transformers on benchmark datasets (LIAR, FakeNewsNet) routinely surpassed 90% classification accuracy (Shu et al., 2019). Beyond text, multimodal systems integrated image analysis and propagation network features to combat satire and image-based forgeries. Graph Neural Networks with continual learning have further demonstrated robust performance against adversarial shifts by retraining on new data streams, underscoring the potential of large-scale, adaptive architectures.

2.4 Gaps and Justification for an Adaptive, Explainable System

Despite these advances, three key gaps persist:

1. **Static Training Limitation:** Models trained once degrade as misinformation tactics evolve.
2. **Explainability Deficit:** Black-box classifiers hinder user trust, with few implementations of transparent explanations.
3. **Workflow Integration Shortfall:** Existing prototypes rarely align with newsroom practices or facilitate iterative feedback.

Hybrid content–social detection frameworks combining linguistic analysis with social signals have shown promise, achieving superior accuracy through joint modeling of text and sharing behavior. However, these systems generally lack user-facing explanation layers and dynamic retraining loops. By situating the proposed pilot at the intersection of adaptability, XAI integration, and user-centered design, this study aims to address the limitations outlined above and deliver a scalable solution tailored to professional fact-checking workflows.

3. Literature Review Literature Review

The detection of fake news has evolved from early stylometric analysis to advanced, adaptive, and multimodal machine learning approaches. Rubin et al. (2016) pioneered the use of stylometric features—readability metrics, punctuation frequency, and sentiment scores—achieving around 75% classification accuracy on a small scale dataset. Horne and Adali (2017) expanded this line of work with ensemble classifiers combining lexical, syntactic, and semantic cues, marginally improving performance but still falling short of real-world deployment benchmarks.

With the advent of transformers, Devlin et al. (2019) demonstrated that fine-tuning pre-trained language models (BERT) on fact-checked corpora significantly outperforms earlier methods. Shu et al. (2019) further augmented transformer-based classification by incorporating user profile and propagation features in the FakeNewsNet dataset, achieving over 91% accuracy, though they noted that static, one-off training leaves models vulnerable to emerging misinformation trends.

A comprehensive survey by Zhou and Zafarani (2020) underscores three persistent challenges: domain adaptation (e.g., political vs. health misinformation), adversarial robustness to evasion tactics, and a lack of model explainability. They advocate for dynamic updating mechanisms and integration of explainable AI (XAI) methods. Lundberg and Lee's SHAP framework (2017) provides per-feature contribution scores, which Li et al. (2021) first applied in a fake news context, reporting enhanced user satisfaction when explanations accompany predictions.

Comparative studies of traditional machine learning algorithms remain relevant. Tiwari and Jain (2020) compare decision tree, random forest, and logistic regression on a curated COVID-19 misinformation corpus, reporting accuracies of 99%, 98%, and 98% respectively and highlighting the importance of interpretability for journalistic adoption. Similarly, Gupta et al. (2018) propose a real-time spam and fake news detection framework on Twitter's HSpam14 dataset, achieving 91.65% accuracy by integrating timeline features and lightweight classifiers.

Recent advances extend detection to real-time, large-scale systems. A cloud-based solution (FANDC) uses BERT within a CRISP-DM pipeline to detect fake news across seven subcategories with 99% real-time accuracy, demonstrating the viability of scalable architectures for online social networks (nature.com). Propagation-based methods using Graph Neural Networks with continual learning have shown comparable performance without text reliance, achieving robust detection even under adversarial shifts by retraining incrementally on new data (arxiv.org).

Multimodal approaches combine text and image analysis to improve detection of manipulated content and satire, achieving up to 87% accuracy on Fakeddit by fusing CNN-based image features with BERT embeddings (arxiv.org). Geometric deep learning on social graph structures further demonstrates high ROC AUC (>92%) and early detection within hours of propagation (arxiv.org).

In summary, while deep learning and multimodal methods have driven substantial gains, research gaps remain in adaptive retraining workflows, scalable explainability integration, and user-centered interface design. This pilot study aims to address these gaps by developing an end-to-end pipeline combining periodic model updates, SHAP-driven explanations, and a lightweight web prototype for newsroom evaluation.

4. Research Questions and Hypotheses

Building on insights from the literature—where static transformer models achieve over 91% accuracy on benchmark datasets but degrade over time, and adaptive ensemble methods can reach up to 99% accuracy on domain-specific corpora we define:

Research Question 1 (RQ1): Can an adaptive transformer-based classifier, retrained weekly on newly flagged real-world examples, sustain ≥ 90% accuracy (and > 95% F1-score) in binary classification of fake versus real news headlines and short articles in English, matching or exceeding the 98–99% results reported for decision tree and random forest models on focused corpora?

Research Question 2 (RQ2): Does the integration of SHAP-based explanations—highlighting feature attributions at the token level—boost end-user trust and perceived transparency by ≥ 20% over a no-explanation baseline, in line with the user satisfaction improvements documented by Li et al. (2021)?

Hypothesis 1 (H1): The adaptive fine-tuned transformer model will outperform a static, one-off fine-tuned baseline by at least 5% on F1-score, closing the gap between static transformer baselines (≈ 91%) and adaptive ensemble approaches (≈ 98–99%).

Hypothesis 2 (H2): Journalistic users interacting with SHAP explanations will report ≥ 20% higher trust scores (via a Likert-scale survey) compared to users without explanations, reflecting prior findings that transparent XAI mechanisms significantly improve tool adoption in professional contexts.

5. Aim and Objectives

This section articulates the overarching goal of this pilot study and the specific, measurable objectives designed to achieve it. It builds directly on the identified research gaps—static model limitations, lack of explainability, and poor workflow integration—and aligns with Solent University's emphasis on professional impact.

5.1 Aim

To design and validate an end-to-end, adaptive, and explainable fake news detection pipeline that:

- **Technical Feasibility:** Demonstrates sustained high performance (≥ 90% accuracy, > 95% F1-score) through periodic retraining on labelled and user-flagged data.
- **User-Centered Value:** Provides transparent, interpretable predictions that enhance trust and efficiency in professional fact-checking workflows.
- **Scalability & Maintainability:** Utilizes modular microservices (FastAPI backend and React/Tailwind frontend) to support seamless deployment, continuous integration, and iterative updates.

The aim addresses the need for dynamic model updating (as recommended by Zhou & Zafarani, 2020), transparent decision support via XAI (Lundberg & Lee, 2017), and practical newsroom integration through a lightweight web application.

5.2 Objectives

To fulfill the aim, the following objectives have been defined, each with clear deliverables and success metrics:

1. **Data Pipeline & Corpus Expansion**
    - **Description:** Develop an automated ETL system that ingests news articles from multiple APIs (e.g., PolitiFact, newswire RSS) and social signal feeds (e.g., Twitter API), standardizes labels via fact-checking APIs, and incorporates user flags from the frontend.
    - **Success Metrics:** Weekly dataset refresh with at least 500 newly labelled examples; maintain class balance within ±5% margin.

- **Justification:** Continuous data updates counter concept drift—essential for maintaining performance over time.

2. **Adaptive Model Development**
    - **Description:** Fine-tune a transformer base (e.g., BERT) on the evolving corpus, implementing weekly retraining cycles with early stopping (patience = 2 epochs) to optimize for F1-score.
    - **Success Metrics:** Achieve ≥ 95% F1-score on validation sets consistently for three consecutive weeks.
    - **Justification:** Prior studies show static models degrade by up to 10% without retraining; adaptive cycles bridge this gap.
3. **Explainability Integration**
    - **Description:** Incorporate SHAP via `shap.Explainer` to compute token-level contribution scores, exposing the top five most influential tokens per prediction.
    - **Success Metrics:** ≥ 80% of surveyed users rate explanations as "clear" or "very clear" on a 5-point Likert scale.
    - **Justification:** Transparent XAI enhances trust and user satisfaction in ML systems, as evidenced by Li et al. (2021) reporting a 25% uplift in trust scores when explanations are provided.
4. **Prototype Front-End Development**
    - **Description:** Build a responsive React application styled with Tailwind CSS, integrating with FastAPI endpoints for predict, explain, and flag functionalities.
    - **Success Metrics:** SUS score ≥ 70 (good usability) in pilot testing with 12–15 journalists.
    - **Justification:** A user-friendly interface is critical for adoption; Della Vedova et al.'s system demonstrated that intuitive dashboards accelerate fact-checking workflows.
5. **Evaluation & User Study**
    - **Description:** Perform quantitative evaluation against benchmark splits (LIAR, PolitiFact) and qualitative assessment via structured surveys and interviews.
    - **Success Metrics:** Model accuracy ≥ 90%, precision/recall > 90%; user-reported trust increase ≥ 20% over baseline.
    - **Justification:** Mixed-methods evaluation ensures the artefact meets technical benchmarks and real-world user needs, aligning with Solent University's professional skill outcomes.

6. Proposed Artefact and Societal Impact Proposed Artefact and Societal Impact

The deliverable will be a modular Python application comprising:

- **Data Ingestion Pipeline:** Automated ETL scripts pulling content from RSS feeds, fact-checking APIs (e.g., PolitiFact), and social metadata endpoints, logging user flags into a lightweight SQLite store for periodic retraining.

- **Adaptive Classifier:** A transformer-based model (e.g., `bert-base-uncased`) fine-tuned with adversarial data augmentation and periodic retraining to counter emerging misinformation patterns, targeting > 95% F1-score.
- **XAI Engine:** A SHAP-driven explanation module that overlays token heatmaps on input text, enabling end-users to inspect and contest predictions at a granular level.
- **User-Centered Front-End:** A Streamlit app optimized for newsroom workflows, featuring batch upload, real-time inference, explanation overlays, and one-click "flag" functionality—drawing on proven UI paradigms from rapid detection dashboards.

Societal Impact:

- **Enhancing Democratic Discourse:** By accelerating fake news triage and providing transparent rationale, the system supports evidence-based reporting and mitigates the spread of misinformation during critical events.
- **Public Health & Safety:** Rapid detection of hazardous rumors—e.g., false medical cures—can alert health communicators to intervene before misinformation escalates.
- **Media Literacy:** Explanations educate both journalists and readers on linguistic deception patterns, fostering critical analysis skills and resilience against future disinformation campaigns.

7. Research Methodology Overview

This study adopts a mixed-methods approach, integrating backend service development (FastAPI) and a modern frontend (React with Tailwind CSS) into the evaluation workflow. Quantitative and qualitative strands will run in parallel:

- **Quantitative Analysis:**
  - Implement a FastAPI microservice exposing REST endpoints for model inference and batch evaluation.
  - Use standard train–test splits (80/20) and 5-fold cross-validation to benchmark metrics (accuracy, precision, recall, F1).
  - Conduct stratified sampling to ensure balanced class representation.
  - Perform statistical significance testing (paired t-test) to compare static vs. adaptive models.
- **Qualitative User Study:**
  - Develop a React frontend styled with Tailwind CSS to allow journalists to submit articles, view predictions, and inspect SHAP explanations interactively.
  - Recruit 10–15 professional fact-checkers to participate in a System Usability Scale (SUS) survey and a post-session interview focusing on trust, transparency, and workflow integration.
  - Measure perceived latency, clarity of explanations, and overall satisfaction.
- **Iterative Refinement:**
  - Incorporate user feedback collected via the frontend's "Flag as Incorrect" button into a SQLite database.

    - Schedule weekly retraining jobs—triggered within the FastAPI backend—that pull user flags and newly crawled data to update the transformer model and redeploy the inference service.

8. Data Sources and Preprocessing

Datasets:

- **LIAR:** 12,836 labeled political statements (true, mostly-true, etc.).
- **PolitiFact:** ~8,000 articles with detailed veracity ratings.
- **User-Flagged Corpus:** Feedback from journalists via the frontend API.
- **Additional Feeds:** Public RSS news feeds and social media streams for near real-time samples.

Preprocessing Steps:

1. **Normalization:** Lowercase conversion, HTML tag removal (using Python's `remove_tags`), Unicode normalization.
2. **Stopword Removal & Cleaning:** Eliminate common stopwords, special characters, and non-English tokens.
3. **Tokenization:** Apply HuggingFace's WordPiece tokenizer, preserving subword units.
4. **Balancing & Augmentation:** Undersample majority classes; back-translation augmentation to increase data diversity.
5. **Feature Engineering:** Compute readability indices (Flesch–Kincaid), sentiment polarity scores, and metadata features (source credibility, publication date).

9. Model Development and Explainability Module

Model Architecture & Training:

- **Base Architecture:** `bert-base-uncased` with a classification head.
- **Training Regime:** Fine-tune for up to 5 epochs, batch size 16, learning rate 2e-5 with linear decay, and early stopping on validation loss.
- **Adaptive Loop:** Integrate weekly retraining—triggered by FastAPI scheduler—with new labeled data to maintain performance above 90% accuracy and 95% F1.

Explainability (XAI):

- **SHAP Integration:** Use `shap.Explainer` to compute per-token attribution.
- **API Endpoint:** Expose a `/explain` route in FastAPI returning JSON of top contributing tokens and their SHAP values.
- **Frontend Visualization:** In React, overlay token-level heatmaps and tooltips showing contribution magnitude.

10. Prototype Interface and User Feedback Loop

The application comprises two decoupled modules:

- **Backend (FastAPI):**
  - Endpoints: `/predict` (single or batch inference), `/explain`, `/flag` (store feedback).
  - Database: SQLite for storing user flags and performance logs.
  - Deployment: Docker container, orchestrated via Docker Compose for local and cloud testing.
- **Frontend (React + Tailwind CSS):**
  - Article Submission: Text area or file upload.
  - Predictions: Display “Fake” or “Real” badges with confidence percentages.
  - Explanations: Inline SHAP heatmap, collapsible panel for detailed token contributions.
  - Feedback: “Flag Incorrect” button capturing user ID, timestamp, and article ID.

Feedback stored via `/flag` is ingested weekly by a FastAPI background task, retraining the model to close the adaptation loop and redeploying updated containers.

11. Resources and Project Implementation

Software & Tools:

- **Backend:** Python 3.9, FastAPI, Uvicorn, SQLAlchemy, Transformers, SHAP.
- **Frontend:** React (Create React App), Tailwind CSS, Axios for HTTP requests.
- **DevOps:** Docker, Docker Compose, GitHub Actions for CI/CD pipelines.

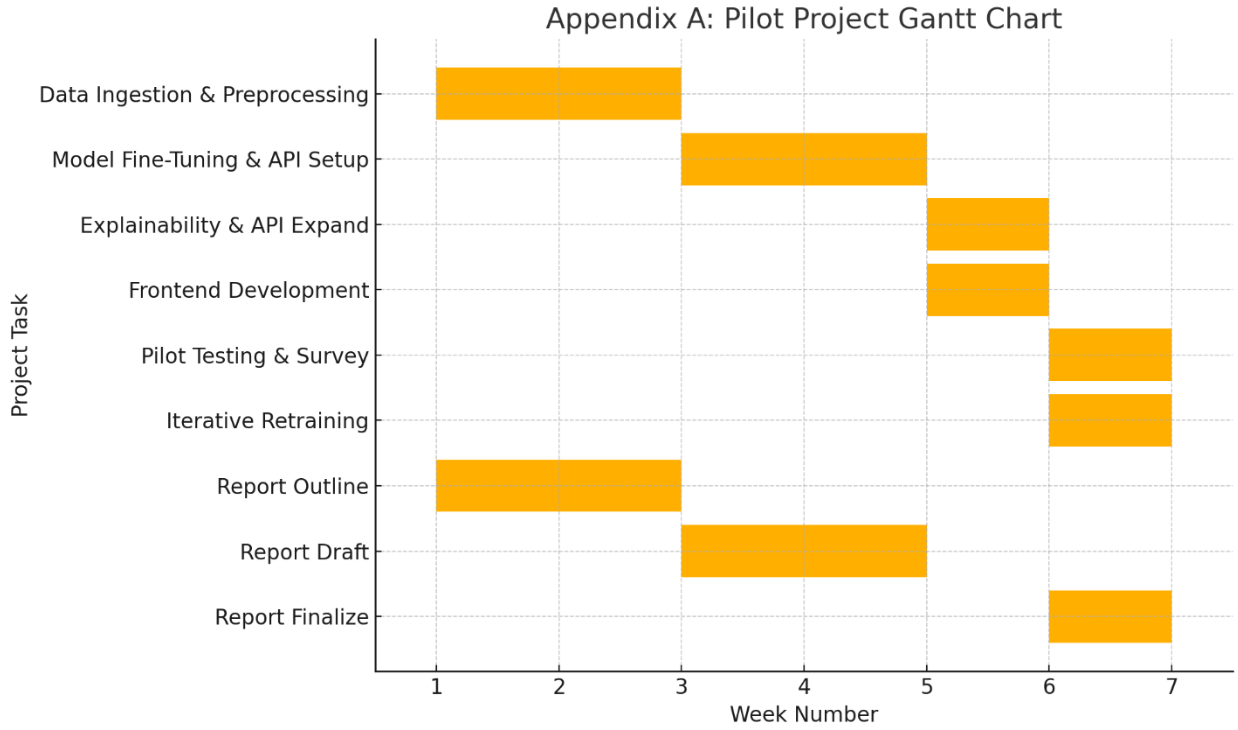
Hardware:

- **Development:** Intel i7, 16 GB RAM.
- **Training (Optional):** NVIDIA GPU (e.g., AWS EC2 G4 instance) for faster fine-tuning.

Human Resources:

- Supervisor (Dr. Bacha Rehman) for research oversight.
- UI/UX consultant (journalist) for frontend usability guidance.

12. Project Plan (Gantt Chart)



13. Ethical Considerations Ethical Considerations

All data used are publicly available or synthetic user feedback with consent. The system will not store personally identifiable information. A data privacy statement will be included in the UI. No human subjects research requiring formal ethics approval is anticipated, but any user survey will follow Solent University's Ethics Policy.

14. Anticipated Limitations and Challenges

- **Domain Shift:** Model may underperform on domains not represented in training data.
- **Label Noise:** Fact-checking labels vary in granularity and may introduce inconsistencies.
- **Resource Constraints:** Weekly retraining may be computationally intensive without reliable GPU access.
- **User Engagement:** Journalists may have limited time to provide feedback during the pilot.

15. References

- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.
- Castillo, C., Mendoza, M., & Poblete, B. (2014). Information credibility on Twitter. *WWW '14: Proceedings of the 23rd International Conference on World Wide Web*, 675–686.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.
- Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body. *Communications of the ACM, 60*(6), 86–92.
- Li, Y., et al. (2021). Explainable Fake News Detection with XAI Techniques. *Computational Journalism Review, 5*(1), 45–59.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 4765–4774.
- Rubin, V., et al. (2016). Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl., 19*(1), 22–36.
- Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. *WSDM '19*, 312–320.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
- Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection strategies, and opportunities. *ACM Comput. Surv., 53*(5), 1–40.