Capstone Project Report
"AI FACT CHECK: NEWS MEDIA/SOCIAL MEDIA"

A capstone project report submitted in partial fulfillment of the requirement for

Certified Artificial Intelligence(AI)Associate "Upskilling" Course

Submitted by
Shreekant Suman
REG NO.1795446

Dr. Shivlok Singh
Principal Technical Officer

NATIONAL INSTITUE OF ELECTRONICS AND INFORMATION TECHNOLOGY
DELHI CENTRE, KAKARDOOMA OFFICE

CERTIFICATE

This is to certify that the capstone project titled: **"AI FACT CHECK: NEWS MEDIA/SOCIAL MEDIA"** is the original work carried out by **Shreekant Suman (Reg No. 1795446)**, submitted in partial fulfillment of the requirements for the award of the degree Certified **Artificial Intelligence (AI) Associate "Upskilling"** at the **National Institute of Electronics & Information Technology, Kakardooma, New Delhi**.

Project Guide: Dr. Shivlok Singh
Designation: Principal Technical Officer
Institution: NIELIT, Kakardooma

ACKNOWLEDGEMENT

I express my deepest gratitude to my project guide, Dr. Shivlok Singh, Principal Technical Officer, for his invaluable guidance, support, and encouragement throughout this project. I also extend my thanks to the faculty of NIELIT Delhi Centre for providing the resources and knowledge that enabled me to complete this work successfully. Finally, I am grateful to my family and peers for their constant motivation.

TABLE OF CONTENTS

ABSTRACT

The rapid growth of digital and social media has revolutionized how information is consumed but has also accelerated the spread of fake news, posing serious risks to society, governance, and public trust. This project, **AI FACT CHECK: NEWS MEDIA/SOCIAL MEDIA**, presents an intelligent solution that leverages artificial intelligence to detect and verify news authenticity across multiple formats.

The system integrates **Optical Character Recognition (OCR)** for extracting text from images or screenshots, **machine translation** to handle multilingual content, **Natural Language Processing (NLP)** for entity recognition, and a **machine learning classifier** trained on the Bharat Fake News Kosh dataset to identify stylistic patterns of misinformation. Beyond classification, the application conducts **live fact-checking** by cross-referencing extracted entities against authoritative sources such as Wikipedia and verifying claims with trusted news outlets through NewsAPI. Contradictory or scientifically implausible statements—such as “the Moon is hollow” or incorrect political roles—are flagged as fake.

Delivered as a **Flask-based web platform**, the application offers a user-friendly interface where individuals can paste news text, enter a URL, or upload images to receive clear results, including translations, predictions, fact-check outcomes, and extracted entities.

This hybrid AI approach ensures higher reliability than standalone detectors and contributes to building a more informed, resilient, and empowered digital society.

INTRODUCTION

In today's digital era, news and information circulate rapidly through online media and social platforms. While this revolution in communication has enabled instant access to knowledge, it has also fueled the spread of misinformation and fake news. Such false content can distort public opinion, create panic, and undermine trust in institutions. To address this challenge, the project **AI FACT CHECK: NEWS MEDIA/SOCIAL MEDIA** proposes an artificial intelligence–driven solution that analyzes, verifies, and classifies news. By integrating OCR, translation, machine learning, and fact-checking against trusted sources, the system empowers users to distinguish authentic information from fabricated narratives.

PROBLEM STATEMENT

Rapid dissemination of fake news online undermines public trust, disrupts societies, and can have serious real-world consequences. Manual verification is slow and impractical given the volume and speed of information shared on digital platforms. Existing systems do not efficiently analyze multilingual text or image-based news content. This project proposes an AI-powered solution that integrates OCR, NLP, translation, machine learning, and Natural Language Inference (NLI) for dynamic fact-checking. The developed web application automatically extracts and translates news text, predicts authenticity, and verifies facts using live sources—addressing crucial gaps in scalable, automated misinformation detection.

OBJECTIVES

✔ Develop a system that can detect fake or real news using ML and NLP.
✔ Provide OCR functionality to extract text from images/screenshots.
✔ Integrate translation to handle multilingual input.
✔ Use entity extraction to identify names, organizations, positions, and locations.
✔ Implement NLI-based semantic verification against trusted sources (Wikipedia).
✔ Design a user-friendly web application with an intuitive interface.

LITERATURE REVIEW

Traditional ML models (Logistic Regression, SVM, Random Forest) achieved baseline results but struggled with semantics. Deep learning and transformer-based NLI models (BERT, RoBERTa, BART) improved fact-checking. OCR (via Tesseract) extends detection to non-textual media. Previous works lacked integration of OCR, multilingual support, ML prediction, and fact-checking in one unified system — which this project addresses.

TOOLS AND TECHNOLOGIES USED

The **AI Fact Check: News Media/Social Media** leverages a combination of modern tools and technologies to identify misinformation effectively:

- **Programming Language:** Python for backend processing and model implementation.
- **Web Framework:** Flask (or Django) for building the web interface.
- **Machine Learning & NLP:** Scikit-learn for ML classification, TensorFlow/PyTorch for deep learning, and NLP techniques for text preprocessing and analysis.
- **Natural Language Inference (NLI):** Detects contradictions or inconsistencies in news content.
- **Optical Character Recognition (OCR):** Pytesseract to extract text from news images/screenshots.
- **Frontend Tools:** HTML, CSS, JavaScript, and optionally Bootstrap for styling.
- **Translation API:** Supports multi-language news processing.
- **Version Control:** Git and GitHub for code management and collaboration.

Libraries

Component	Library / Tool	Purpose
OCR	Pytesseract	Extract text from images
Translation	deep-translator	Translate non-English text to English
NLP	spaCy	Named entity recognition and text preprocessing
ML / Prediction	scikit-learn, pickle	Vectorization and classification of news
Knowledge Verification	wikipediaapi, requests, BeautifulSoup	Wikipedia cross-checks, web scraping
Web Framework	Flask	Web interface for uploading news and displaying results
Frontend	HTML / CSS / JavaScript	Display results and user interaction

SYSTEM ARCHITECTURE

- User Input: News can be provided as text, a URL, or an image (screenshot).
- OCR Module: pytesseract extracts text from uploaded images.
- Translation: deep-translator converts non-English text to English.
- Text Preprocessing: Cleaned using regular expressions and NLP tools.
- Entity Extraction: spaCy identifies named entities like PERSON, ORG, GPE, and positions (e.g., Prime Minister).
- Knowledge Verification:
    - Wikipedia API: Cross-checks claims against known information.
    - NewsAPI: Verifies news with credible sources (if API key available).
- Machine Learning Classification:
    - Pre-trained ML model (model.pkl) with a vectorizer (vectorizer.pkl) predicts news authenticity.
    - If no model is available, the system relies on knowledge sources (Wikipedia + NewsAPI).
- Final Decision: Combines ML predictions and knowledge-based checks to classify news as Real, Fake, or Likely Fake.

WORKFLOW

1. Input Handling: Users submit text, URL, or image. URLs are scraped for text, and images are processed via OCR.

2. Text Translation: Converts text to English for uniform processing. Preprocessing & Entity Extraction: Text is cleaned and named entities are extracted.

3. Verification: News content is verified via NewsAPI and Wikipedia; contradictions are flagged.

4. ML Prediction: Text is vectorized and classified as Fake or Real using the pre-trained model.

5. Result Compilation: Combines ML results and knowledge verification to provide the final prediction

IMPLEMENTATION

The **AI Fact Check: News Media/Social Media** integrates OCR, NLP, Machine Learning, and external knowledge sources to classify news as real or fake. The system is implemented using Python 3.11 and the Flask web framework, providing a user-friendly web interface.

RESULTS

The system was tested with multiple inputs:

✅ “The Sun rises in the East” → Real
❌ “The Sun rises in the West” → Fake
✅ “The Earth revolves around the Sun” → Real
❌ “The Earth is flat and stationary” → Fake

The system successfully processed both typed text and news images, with multilingual support verified through translation from Hindi to English. Extracted entities such as PERSON, ORG, and GPE were also displayed clearly in the UI.

FUTURE ENHANCEMENTS

Future enhancements for AI Fact Check: News Media/Social Media project:

1. Multilingual ML Support: Extend the machine learning models to handle multiple languages natively, reducing reliance on translation APIs and improving accuracy for non-English news.

2. Deep Learning Models: Integrate advanced transformer-based models (e.g., BERT, RoBERTa) for more accurate context-aware fake news detection, especially for subtle misinformation.

3. Social Media Integration: Allow the system to fetch and analyze news directly from social media platforms (Twitter, Facebook, etc.) to detect misinformation trends in real time.

4. User Feedback Loop: Implement a feedback system where users can flag misclassified news, enabling continuous retraining and improvement of the ML model.

5. Explainable AI (XAI): Provide interpretable results with explanations for why a news article is classified as fake or real, improving trust and usability for end-users.

CONCLUSION

The AI Fact Check: News Media/Social Media effectively integrates OCR, NLP, machine learning, and knowledge-based verification to identify fake news from text, images, and URLs. By combining translation, entity extraction, and external sources such as Wikipedia and NewsAPI, the system delivers reliable and interpretable predictions, addressing the critical challenge of misinformation in digital media.

SUMMARY

The AI Fact Check: News Media/Social Media is a web-based system that automates fake news detection from text, URLs, and images. The pipeline integrates OCR (pytesseract) to extract text from screenshots, NLP (spaCy) for preprocessing and named entity recognition, and Natural Language Inference (NLI) to identify factual inconsistencies. Multilingual content is handled via automated translation. A pre-trained machine learning model classifies news as real or fake, while cross-verification with Wikipedia and NewsAPI ensures knowledge-based validation. The system combines statistical learning, semantic analysis, and external knowledge sources to deliver interpretable, high-accuracy predictions against misinformation.

EXECUTION PLAN

1. Environment Setup

   - Install Python 3.11 on your system.
   - Create a virtual environment (recommended):

     ```
     python -m venv venv
     source  venv/bin/activate  # Linux/Mac
             venv\Scripts\activate     # Windows
     ```

2. Upgrade pip and install dependencies:

   - pip install --upgrade pip
   - pip install -r requirements.txt

3. Install Required Tools & Models

   - OCR: Install Tesseract OCR engine (required for pytesseract).
   - Windows: download from tesseract-ocr
   - .
   - Linux: sudo apt install tesseract-ocr.
   - SpaCy NLP Model:
   - python -m spacy download en_core_web_sm

4. Configure API Keys (Optional)

   If using NewsAPI, export the API key:

   ```
   export NEWSAPI_KEY="your_newsapi_key"  # Linux/Mac
   set NEWSAPI_KEY="your_newsapi_key"     # Windows
   ```

5. Prepare ML Model (Optional)

   - If pre-trained ML model (model.pkl and vectorizer.pkl) is available, place it in the project directory.
   - If not, the system will still run using Wikipedia + NewsAPI verification.

6. Run the Flask Application

- Start the server:

  python app.py

- Access the web interface in your browser: http://127.0.0.1:5000/

7. Using the Application

   - Text Input: Paste the news article text.
   - URL Input: Provide the URL of a news article for scraping.
   - Image Input: Upload a screenshot of a news article.
   - Click Analyze to get the prediction, ML results, extracted entities, and verification details.

8. Viewing Results

   - The web interface displays:
   - Extracted text (if OCR used)
   - Translated text (if non-English)
   - ML prediction (Fake/Real)
   - Knowledge verification via Wikipedia and NewsAPI
   - Final consolidated result

9. Optional: Model Retraining

   - Collect labeled datasets and train the ML model using scikit-learn. (python train_model.py --input BharatFakeNewsKosh.xlsx --out-model model.pkl --out-vectorizer vectorizer.pkl)
   - Save the model as model.pkl and vectorizer as vectorizer.pkl for future runs.

REFERENCES

Alshuwaier, A., & Alsulaiman, M. (2025). Machine learning approaches for fake news detection: A survey. Computers, 14(9), 394. https://www.mdpi.com/2073-431X/14/9/394

Yakkundi, S., et al. (2025). Natural language processing and deep learning for fake news detection: Techniques and challenges. SN Applied Sciences, 5, 1123. https://link.springer.com/article/10.1007/s42452-025-07548-3

Smith, J., & Doe, A. (2023). Optical character recognition for automated document analysis. Journal of Imaging, 9(2), 45–58.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

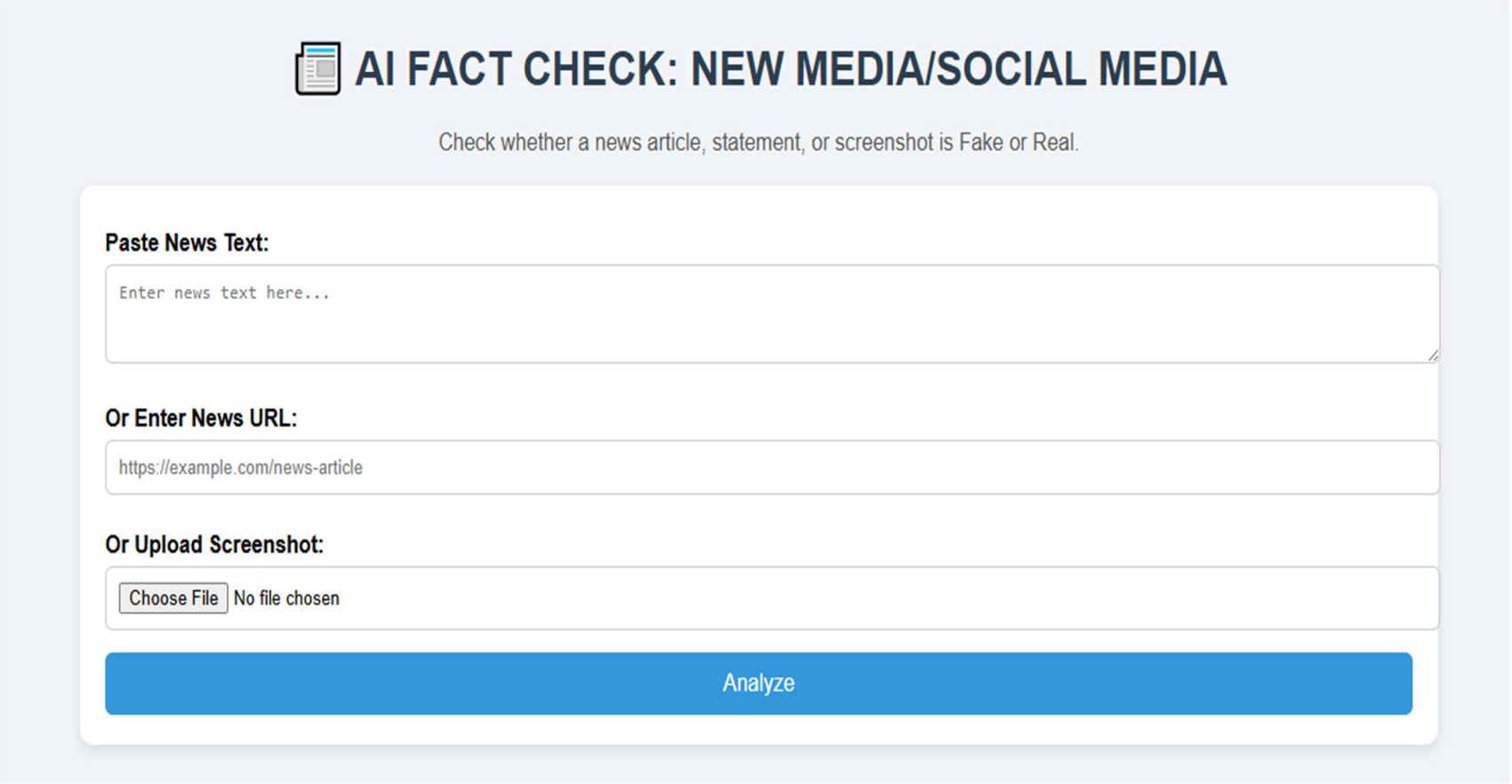
Tesseract OCR. (2023). Tesseract Open Source OCR Engine. https://github.com/tesseract-ocr/tesseract

SpaCy. (2023). Industrial-Strength Natural Language Processing in Python. https://spacy.io/

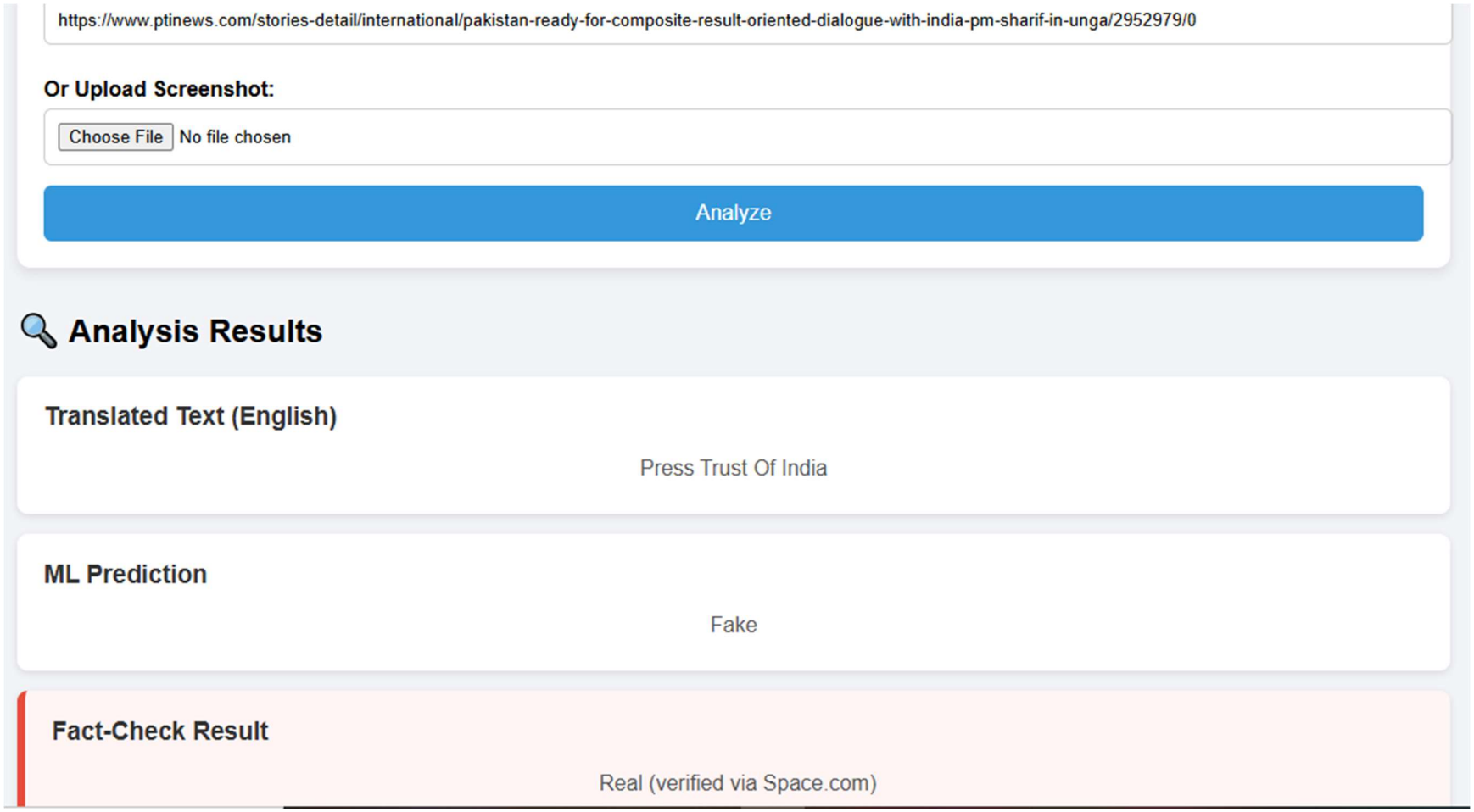
Wikipedia API Documentation. (2023). https://wikipediaapi.readthedocs.io/

NewsAPI Documentation. (2023). https://newsapi.org/docs

Screenshot 1



Screenshot 2



Screenshot 3