L-Università ta' Malta
Faculty of Information & Communication Technology

Department of Artificial Intelligence

B.Sc. IT (Hons.) Artificial Intelligence

Accurate Name Extraction from News Video Graphics

Andrea Filiberto Lucas
Supervisor: Dr. Dylan Seychell

INTRODUCTION

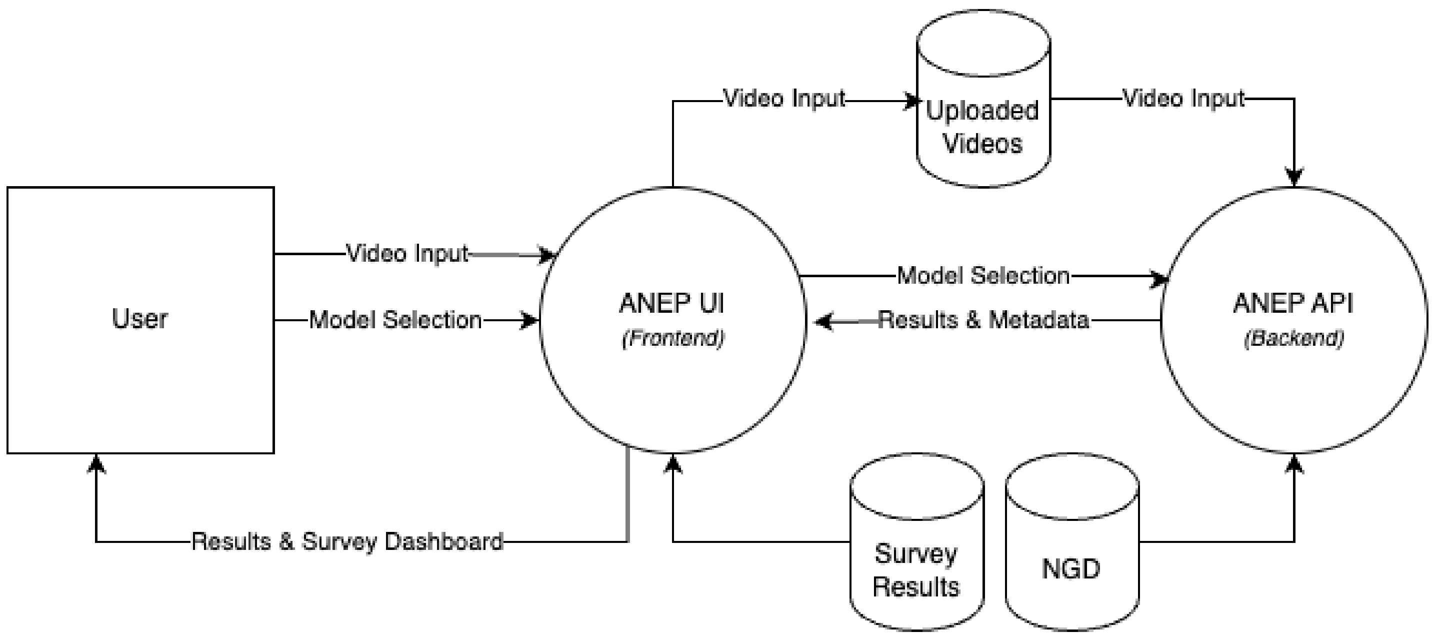
News broadcasts employ numerous graphic overlays, from lower thirds identifying speakers to scrolling tickers with breaking news. These elements vary significantly in typography, positioning, background designs, and transparency levels. Each broadcaster implements distinct visual styles, making manual extraction of featured names a tedious process fraught with potential errors. Current automated approaches [1] struggle with this visual heterogeneity, frequently misidentifying text or failing to adapt to varying design conventions.

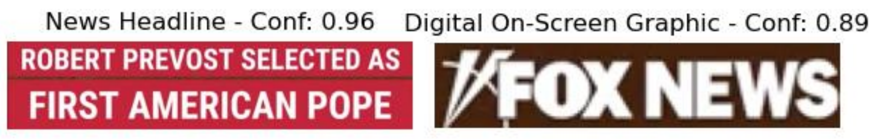
This dissertation presents the Accurate Name Extraction Pipeline (ANEP), a robust system for extracting names from news graphics. ANEP combines three core technologies: YOLOv12 for graphic detection, optical character recognition (OCR) for text extraction, and name entity recognition (NER) for name identification. Performance is benchmarked against leading AI solutions including Google Vision with Gemini 1.5 Pro and LLaMA 4 Maverick, providing comprehensive accuracy and applicability insights.

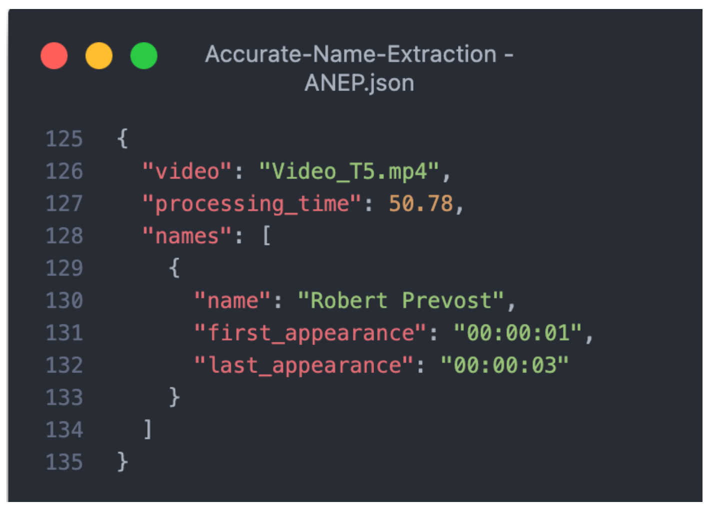
AIM

This study pursued three primary objectives. First, the **News Graphics Dataset (NGD)** was developed, comprising 1,500 frames from local and international news broadcasts alongside social media outlets. This dataset features diverse fonts, layouts, and colour schemes to enable robust YOLOv12 training across real-world graphics. Second, **ANEP** was introduced, a pipeline integrating CNN-based detection, OCR text extraction, and transformer-based NER to accurately locate and validate names. Third, a **comparative framework was established to benchmark ANEP against GenAI methods**, namely Google Vision with Gemini 1.5 Pro and LLaMA 4 Maverick. Performance metrics including precision, recall, F1-score, and runtime were evaluated to assess trade-offs between traditional and multimodal approaches.

ARCHITECTURE DESIGN

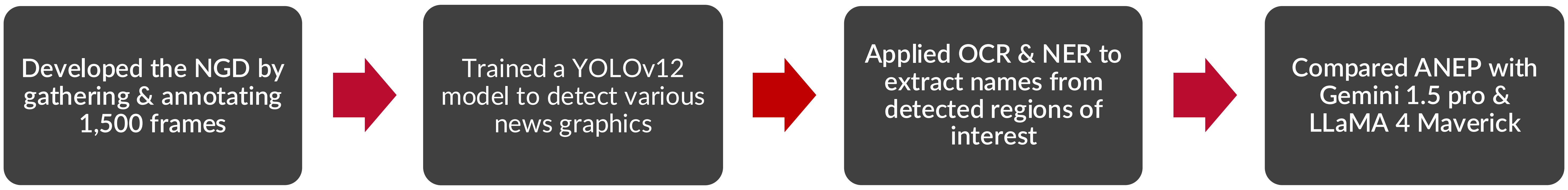



News Headline - Conf: 0.96
Digital On-Screen Graphic - Conf: 0.89
ROBERT PREVOST SELECTED AS
FIRST AMERICAN POPE
FOX NEWS


Accurate-Name-Extraction - ANEP.json

```
{
  "video": "Video_T5.mp4",
  "processing_time": 50.78,
  "names": [
    {
      "name": "Robert Prevost",
      "first_appearance": "00:00:01",
      "last_appearance": "00:00:03"
    }
  ]
}
```

METHODOLOGY


Developed the NGD by gathering & annotating 1,500 frames

Trained a YOLOv12 model to detect various news graphics

Applied OCR & NER to extract names from detected regions of interest

Compared ANEP with Gemini 1.5 pro & LLaMA 4 Maverick

RESULTS

The evaluation demonstrated that **95.8% object detection accuracy was achieved** by the locally trained YOLOv12 model on the NGD, enabling reliable localisation of graphical elements across diverse news video frames. Building on this foundation, **balanced name extraction performance was delivered by the ANEP,** with **72.92% precision and 74.44% recall recorded**, alongside an F1 score of 68.10%. The system's modular architecture provided robust and explainable results. In comparison, **the highest F1 score of 82.22% was achieved** by the Google Vision combined with Gemini 1.5 Pro, with superior metrics recorded (93.33% precision, 76.67% recall) and significantly faster processing. The LLaMA 4 Maverick pipeline recorded weaker performance with 55.56% F1 score. These findings underscore ANEP's consistency and interpretability for applications requiring audit trails, whilst confirming the superior speed and accuracy of GenAI approaches where transparency requirements are less stringent.

CONCLUSIONS AND FUTURE WORK

This research advances automated name extraction from news video graphics through the development of the NGD and ANEP, which integrates YOLO-based object detection, OCR, and transformer-based NER. The approach extends earlier television summarisation systems [2] by introducing a generalisable visual-first methodology for entity extraction, whilst aligning with emerging multimodal information extraction research [3]. Future developments could expand the NGD to encompass multilingual and social media content, optimise ANEP for real-time deployment, and investigate audio-visual fusion to enhance contextual understanding.

REFERENCES

[1] J. Hong *et al.*, "Analysis of faces in a decade of US cable TV news," in *Proc. 27th ACM SIGKDD Int. Conference* Knowledge Discovery & Data MiningAssoc. Computing Machinery, 2021
[2] J. Attard and D. Seychell, Comparative analysis of image, video, and audio classifiers for automated news video segmentation, 2025. arXiv: 2503.21848 [Online]. Available: https://arxiv.org/abs/2503.21848
[3] R. Sapkota, S. Raza, M. Shoman, A. Paudel, and M. Karkee, "Multimodal large language models for image, text, and speech data augmentation: A survey", arXiv preprint arXiv:2501.18648, 2025. [Online]. Available: https://arxiv.org/abs/2501.18648