# Multimodal Fact-Checking: Dataset and Baselines

## 1 Overview

This project aims to develop a multimodal fact-checking dataset and establish baseline verification methods. The dataset comprises unlabeled tweets collected systematically by journalists, focusing on potentially hateful or deceptive content. The study specifically targets misinformation related to immigrants, analyzing both the tweets and their accompanying images.

## 2 Research Goals

The primary objectives include dataset curation and classification model development for verifying multimodal content. The final dataset will undergo human verification to ensure reliability.

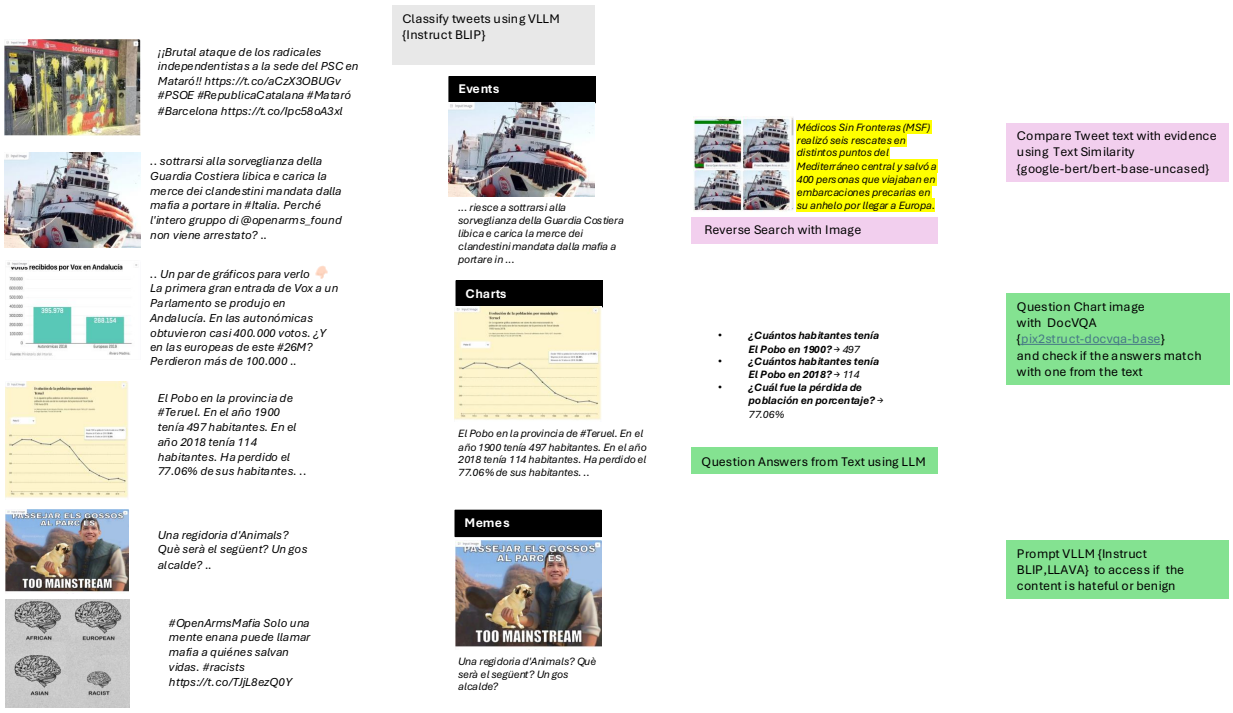


Figure 1: Pipeline for Proposed Work. Modules in Purple have starter code available. Modules in Grey are in development. Modules in Green will need to be implemented from scratch

## 3 Tasks and Methodology

The global project integrates several tasks as illustrated in Figure 1 and described below. The master project can focus on one or several of theses tasks, as it will be defined at the beginning of the project.

### 3.1 Filtering Task (Primary Objective 1)

We categorize tweet-image pairs into three main categories: **Memes, Charts,** and **Events** (including protests, conflicts, gatherings, and natural disasters). This classification is performed using InstructBLIP.

### 3.2 Classification Tasks

#### 3.2.1 Event Classification (Primary Objective 2)

We use reverse image search to find supporting evidence and validate it against tweet text using Hugging Face Sentence Transformers [1]. Based on the verification results, each event is labeled as **True, Fake,** or **Not Enough Data**.

#### 3.2.2 Chart Classification (Secondary Objective 1)

To verify chart accuracy, we generate structured questions from tweet text with a Large Language Model (LLM) and analyze the chart using DocVQA models like Pix2Struct-DocVQA [2].

#### 3.2.3 Meme Classification (Primary Objective 3)

We detect hate speech in text-image pairs using CLIP-based embeddings and fine-tuned multimodal large language models (MLLMs) such as InstructBLIP [3, 4] . Memes are labeled as **Hateful, Benign,** or **Unsure**.

### 3.3 Ground Truth Generation (Primary Objective 4)

All labeled samples (**True/Fake** for events and **Hateful/Benign** for memes) undergo manual verification. The finalized dataset, consisting of **Image, Text, Evidence,** and **Label**, will be publicly released for research purposes.

## 4 What You Will Learn

- **Multimodal Learning:** Understanding how to process and analyze text-image pairs using state-of-the-art models like InstructBLIP and CLIP-based embeddings.
- **Natural Language Processing (NLP):** Applying sentence transformers and large language models (LLMs) to verify textual claims in multimodal fact-checking.
- **Computer Vision for Misinformation Detection:** Utilizing reverse image search, document visual question answering (DocVQA), and chart analysis techniques.
- **Dataset Curation and Annotation:** Learning best practices for collecting, filtering, and manually verifying multimodal datasets for fact-checking tasks.
- **Fine-Tuning and Model Evaluation:** Finetuning MLLMs for tasks such as hate speech detection and misinformation classification.
- **AI for Social Good:** Understanding real-world challenges in misinformation detection, ethical AI, and the societal impact of multimodal verification systems.

By the end of the project, participants will have developed a solid foundation in multimodal AI techniques and gained practical experience in building and evaluating models for fact-checking applications.

## References


[1] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[2] T. Gao, Z. Wang, A. Bhaskar, and D. Chen, "Improving language understanding from screenshots," *arXiv preprint arXiv:2402.14073*, 2024.

[3] N. Rizwan, P. Bhaskar, M. Das, S. S. Majhi, P. Saha, and A. Mukherjee, "Zero shot vlms for hate meme detection: Are we there yet?" *arXiv preprint arXiv:2402.12198*, 2024.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.